

# Solutions

10-315 Machine Learning  
Spring 2023  
Exam 2 Practice Problems  
April 25, 2023  
Time Limit: N/A

Name:  
Andrew Email:  
Room:  
Seat:  
Exam Number:

---

## Instructions:

- Fill in your name and Andrew ID above. Be sure to write neatly, or you may not receive credit for your exam.
  - Clearly mark your answers in the allocated space **on the front of each page**. If needed, use the back of a page for scratch space, but you will not get credit for anything written on the back of a page. If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
  - No electronic devices may be used during the exam.
  - Please write all answers in pen.
  - You have N/A to complete the exam. Good luck!
-

## Instructions for Specific Problem Types

For “Select One” questions, please fill in the appropriate bubble completely:

**Select One:** Who taught this course?

- Pat Virtue
- Marie Curie
- Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

**Select One:** Who taught this course?

- Pat Virtue
- Marie Curie
- Noam Chomsky

For “Select all that apply” questions, please fill in all appropriate squares completely:

**Select all that apply:** Which are scientists?

- Stephen Hawking
- Albert Einstein
- Isaac Newton
- I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

**Select all that apply:** Which are scientists?

- Stephen Hawking
- Albert Einstein
- Isaac Newton
- I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

**Fill in the blank:** What is the course number?

10-315

10-~~7~~315

# 1 Optimization

1. **Select all that apply:** Which of the following are correct regarding Gradient Descent (GD) and stochastic gradient descent (SGD)

- Each update step in SGD pushes the parameter vector closer to the parameter vector that minimizes the objective function.
- The gradient computed in SGD is, in expectation, equal to the gradient computed in GD.
- The gradient computed in GD has a higher variance than that computed in SGD, which is why in practice SGD converges faster in time than GD.

B.

A is incorrect, SGD updates are high in variance and may not go in the direction of the true gradient. C is incorrect, for the same reason. D is incorrect since they can converge if the function is convex, not just strongly convex.

2. (a) Determine if the following 1-D functions are convex. Assume that the domain of each function is  $\mathbb{R}$ . The definition of a convex function is as follows:

$$f(x) \text{ is convex} \iff f(\alpha x + (1 - \alpha)z) \leq \alpha f(x) + (1 - \alpha)f(z), \forall \alpha \in [0, 1] \text{ and } \forall x, z.$$

Select all convex functions:

- $f(x) = x + b$  for any  $b \in \mathbb{R}$
- $f(x) = c^2x$  for any  $c \in \mathbb{R}$
- $f(x) = ax^2 + b$  for any  $a \in \mathbb{R}$  and any  $b \in \mathbb{R}$
- $f(x) = 0$
- None of the Above

$f(x) = x + b$  for any  $b \in \mathbb{R}$ ,  $f(x) = c^2x$  for any  $c \in \mathbb{R}$ ,  $f(x) = 0$ .

- (b) Consider the convex function  $f(z) = z^2$ . Let  $\alpha$  be our learning rate in gradient descent.

For which values of  $\alpha$  will  $\lim_{t \rightarrow \infty} f(z^{(t)}) = 0$ , assuming the initial value of  $z$  is  $z^{(0)} = 1$  and  $z^{(t)}$  is the value of  $z$  after the  $t$ -th iteration of gradient descent.

Select all that apply:

- $\alpha = 0$
- $\alpha = \frac{1}{2}$
- $\alpha = 1$
- $\alpha = 2$

None of the Above

$$\alpha = \frac{1}{2}$$

Give the range of all values for  $\alpha \geq 0$  such that  $\lim_{t \rightarrow \infty} f(z^{(t)}) = 0$ , assuming the initial value of  $z$  is  $z^{(0)} = 1$ . Be specific.

$(0, 1)$ .

## 2 Logistic Regression

1. [2 pts] If today I want to predict the probability that a student sleep more than 8 hours on average (SA) given the Course loading (C), I will choose to use linear regression over logistic regression.

Circle one:      True      False

False.

2. Answer the following questions with brief explanations where necessary.

- a) [2 pts] A generalization of logistic regression to a multiclass settings involves expressing the per-class probabilities  $P(y = c|x)$  as the softmax function  $\frac{\exp(w_c^T x)}{\sum_{d \in C} \exp(w_d^T x)}$ , where  $c$  is some class from the set of all classes  $C$ .

Consider a 2-class problem (labels 0 or 1). Rewrite the above expression for this situation, to end up with expressions for  $P(Y = 1|x)$  and  $P(Y = 0|x)$  that we have already come across in class for binary logistic regression.

$$P(y = 1|x) = \frac{\exp(w_1^T x)}{\exp(w_0^T x) + \exp(w_1^T x)} = \frac{\exp((w_1 - w_0)^T x)}{1 + \exp((w_1 - w_0)^T x)} = \frac{\exp(w^T x)}{1 + \exp(w^T x)} = p$$

$$\text{Therefore, } 1 - p = \frac{1}{1 + \exp(w^T x)}$$

- b) [3 pts] Given 3 data points  $(1, 1)$ ,  $(1, 0)$ ,  $(0, 0)$  with labels 0, 1, 0 respectively. Consider 2 models, Model 1:  $\sigma(w_1 x_1 + w_2 x_2)$ , Model 2:  $\sigma(w_0 + w_1 x_1 + w_2 x_2)$  ( $\sigma(z)$  is the sigmoid function  $\frac{1}{1 + e^{-z}}$ ) that compute  $p(y = 1|\mathbf{x})$ . Using the given data, we can learn parameters  $\hat{w}$  by maximizing the conditional log-likelihood.

Suppose we switched  $(0, 0)$  to label 1 instead.

Do the parameters learnt for Model 1 change?

Circle one:      True      False

One-line explanation:

False. The parameters learnt for Model 1 don't change because  $w_1 x_1 + w_2 x_2 = 0$  for  $(0, 0)$ . Hence  $p = 0.5$  irrespective of the labels or the values of  $w$ .

What about Model 2?

Circle one:      True      False

One-line explanation:

True. This model has a bias term which remains non-zero for  $(0, 0)$ , and can thus change the model depending on the label assigned.

- c) [2 pts] For logistic regression, we need to resort to iterative methods such as gradient descent to compute the  $\hat{w}$  that maximizes the conditional log likelihood. Why?

There is no closed-form solution.

- d) [3 pts] Considering a Gaussian prior, write out the MAP objective function  $J(w)_{MAP}$  in terms of the MLE objective  $J(w)_{MLE}$ . Name the variant of logistic regression this results in.

$J_{MAP}(\mathbf{w}) = J_{MLE}(\mathbf{w}) - \lambda \|\mathbf{w}\|_2^2$ . This is L2 regularized logistic regression.

3. Given a training set  $\{(x_i, y_i), i = 1, \dots, n\}$  where  $x_i \in \mathbb{R}^d$  is a feature vector and  $y_i \in \{0, 1\}$  is a binary label, we want to find the parameters  $\hat{w}$  that maximize the likelihood for the training set, assuming a parametric model of the form

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w^T x)}.$$

The conditional log likelihood of the training set is

$$\ell(w) = \sum_{i=1}^n y_i \log p(y_i, |x_i; w) + (1 - y_i) \log(1 - p(y_i, |x_i; w)),$$

and the gradient is

$$\nabla \ell(w) = \sum_{i=1}^n (y_i - p(y_i|x_i; w))x_i.$$

- a) [5 pts.] Is it possible to get a closed form for the parameters  $\hat{w}$  that maximize the conditional log likelihood? How would you compute  $\hat{w}$  in practice?

There is no closed form expression for maximizing the conditional log likelihood. One has to consider iterative optimization methods, such as gradient descent, to compute  $\hat{w}$ .

- b) [5 pts.] For a binary logistic regression model, we predict  $y = 1$ , when  $p(y = 1|x) \geq 0.5$ . Show that this is a linear classifier.

Using the parametric form for  $p(y = 1|x)$ :

$$\begin{aligned} p(y = 1|x) \geq \frac{1}{2} &\implies \frac{1}{1 + \exp(-w^T x)} \geq \frac{1}{2} \\ &\implies 1 + \exp(-w^T x) \leq 2 \\ &\implies \exp(-w^T x) \leq 1 \\ &\implies -w^T x \leq 0 \\ &\implies w^T x \geq 0, \end{aligned}$$

so we predict  $\hat{y} = 1$  if  $w^T x \geq 0$ .

- c) Consider the case with binary features, i.e,  $x \in \{0, 1\}^d \subset \mathbb{R}^d$ , where feature  $x_1$  is rare and happens to appear in the training set with only label 1. What is  $\hat{w}_1$ ? Is the gradient ever zero for any finite  $w$ ? Why is it important to include a regularization term to control the norm of  $\hat{w}$ ?

If a binary feature fired for only label 1 in the training set then, by maximizing the conditional log likelihood, we will make the weight associated to that feature be infinite. This is because, when this feature is observed in the training set, we will want to predict predict 1 irrespective of everything else. This is an undesired behaviour from the point of view of generalization performance, as most likely we do not believe this rare feature to have that much information about class 1. Most likely, it is spurious co-occurrence. Controlling the norm of the weight vector will prevent these pathological cases.

4. Given the following dataset,  $\mathcal{D}$ , and a fixed parameter vector,  $\boldsymbol{\theta}$ , write an expression for the binary logistic regression conditional likelihood.

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)} = 0), (\mathbf{x}^{(2)}, y^{(2)} = 0), (\mathbf{x}^{(3)}, y^{(3)} = 1), (\mathbf{x}^{(4)}, y^{(4)} = 1)\}$$

- Write your answer in terms of  $\boldsymbol{\theta}$ ,  $\mathbf{x}^{(1)}$ ,  $\mathbf{x}^{(2)}$ ,  $\mathbf{x}^{(3)}$ , and  $\mathbf{x}^{(4)}$ .
- Do not include  $y^{(1)}$ ,  $y^{(2)}$ ,  $y^{(3)}$ , or  $y^{(4)}$  in your answer.
- Don't try to simplify your expression.

**Conditional likelihood:**

$$\left(1 - \frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}^{(1)}}}\right) \left(1 - \frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}^{(2)}}}\right) \frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}^{(3)}}} \frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}^{(4)}}}$$

5. Write an expression for the decision boundary of binary logistic regression with a bias term for two-dimensional input features  $x_1 \in \mathbf{R}$  and  $x_2 \in \mathbf{R}$  and parameters  $b$  (the intercept parameter),  $w_1$ , and  $w_2$ . Assume that the decision boundary occurs when  $P(Y = 1 | \mathbf{x}, b, w_1, w_2) = P(Y = 0 | \mathbf{x}, b, w_1, w_2)$ .

- (a) Write your answer in terms of  $x_1$ ,  $x_2$ ,  $b$ ,  $w_1$ , and  $w_2$ .

**Decision boundary equation:**

- (b) What is the geometric shape defined by this equation?

(a)  $0 = b + w_1 x_1 + w_2 x_2$  (b) A line.

6. We have now feature engineered the two-dimensional input,  $x_1 \in \mathbf{R}$  and  $x_2 \in \mathbf{R}$ , mapping

it to a new input vector:  $\mathbf{x} = \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \end{bmatrix}$

- (a) Write an expression for the decision boundary of binary logistic regression with this feature vector  $\mathbf{x}$  and the corresponding parameter vector  $\boldsymbol{\theta} = [b, w_1, w_2]^T$ . Assume that the decision boundary occurs when  $P(Y = 1 | x, \boldsymbol{\theta}) = P(Y = 0 | x, \boldsymbol{\theta})$ . Write your answer in terms of  $x_1, x_2, b, w_1,$  and  $w_2$ .

**Decision boundary expression:**

- (b) What is the geometric shape defined by this equation?

- (c) If we add an L2 regularization on  $[w_1, w_2]^T$ , what happens to **parameters** as we increase the  $\lambda$  that scales this regularization term?

- (d) If we add an L2 regularization on  $[w_1, w_2]^T$ , what happens to the **decision boundary shape** as we increase the  $\lambda$  that scales this regularization term?

(a)  $0 = b + w_1x_1^2 + w_2x_2^2$  (b) An ellipse. Probably decent partial credit for circle. (c) The magnitude of the parameters will decrease. (d) The parameters shrink, so the ellipse will get bigger.



### 3 Feature Engineering and Regularization

1. **Model Complexity:** In this question we will consider the effect of increasing the model complexity, while keeping the size of the training set fixed. To be concrete, consider a classification task on the real line  $\mathbb{R}$  with distribution  $D$  and target function  $c^* : \mathbb{R} \rightarrow \{\pm 1\}$  and suppose we have a random sample  $S$  of size  $n$  drawn iid from  $D$ . For each degree  $d$ , let  $\phi_d$  be the feature map given by  $\phi_d(x) = (1, x, x^2, \dots, x^d)$  that maps points on the real line to  $(d + 1)$ -dimensional space.

Now consider the learning algorithm that first applies the feature map  $\phi_d$  to all the training examples and then runs logistic regression as in the previous question. A new example is classified by first applying the feature map  $\phi_d$  and then using the learned classifier.

- a) [4 pts.] For a given dataset  $S$ , is it possible for the training error to increase when we increase the degree  $d$  of the feature map? **Please explain your answer in 1 to 2 sentences.** No. Every linear separator using the feature map  $\phi_d$  can also be expressed using the feature map  $\phi_{d+1}$ , since we are only adding new features. It follows that the training error must decrease for any given sample  $S$ .
- b) [4 pts.] Briefly **explain in 1 to 2 sentences** why the true error first drops and then increases as we increase the degree  $d$ . When the dimension  $d$  is small, the true error is high because it is not possible to the target function is not well approximated by any linear separator in the  $\phi_d$  feature space. As we increase  $d$ , our ability to approximate  $c^*$  improves, so the true error drops. But, as we continue to increase  $d$ , we begin to overfit the data and the true error increases again.

## 4 Neural Networks

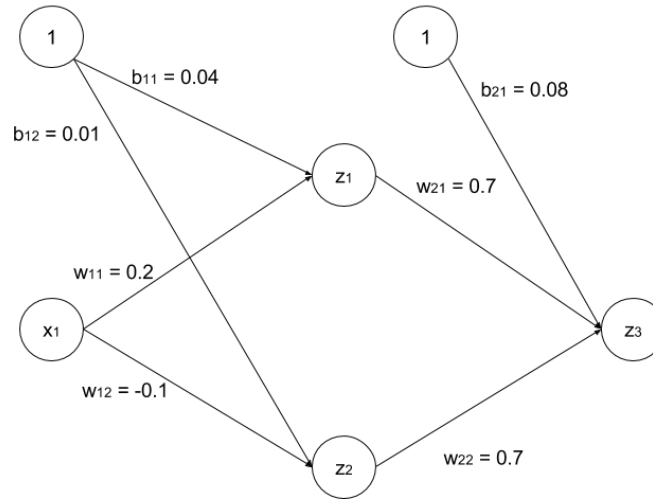


Figure 1: neural network

- Consider the neural network architecture shown above for a 2-class  $(0, 1)$  classification problem. The values for weights and biases are shown in the figure. We define:

$$a_1 = w_{11}x_1 + b_{11}$$

$$a_2 = w_{12}x_1 + b_{12}$$

$$a_3 = w_{21}z_1 + w_{22}z_2 + b_{21}$$

$$z_1 = \text{relu}(a_1)$$

$$z_2 = \text{relu}(a_2)$$

$$z_3 = \sigma(a_3), \sigma(x) = \frac{1}{1+e^{-x}}$$

Use this information to answer the questions that follow.

- [6 pts]** For  $x_1 = 0.3$ , compute  $z_3$ , in terms of  $e$ . **Show all work.**

$$z_3 =$$

$$z_3 = \frac{1}{1+e^{-0.15}}$$

- [2 pts]** To which class does the network predict the given data point ( $x_1 = 0.3$ ), i.e.,  $\hat{y} = ?$  Note that  $\hat{y} = 1$  if  $z_3 > \frac{1}{2}$ , else  $\hat{y} = 0$ .

**Circle one:**      0      1

$$\hat{y}(x_1 = 0.3) = 1$$

- (iii) [6 pts] Perform backpropagation on the bias  $b_{21}$  by deriving the expression for the gradient of the loss function  $L(y, z_3)$  with respect to the bias term  $b_{21}$ ,  $\frac{\partial L}{\partial b_{21}}$ , in terms of the partial derivatives  $\frac{\partial \alpha}{\partial \beta}$ , where  $\alpha$  and  $\beta$  can be any of  $L, z_i, a_i, b_{ij}, w_{ij}, x_1$  for all valid values of  $i, j$ . Your backpropagation algorithm should be as explicit as possible—that is, make sure each partial derivative  $\frac{\partial \alpha}{\partial \beta}$  cannot be decomposed further into simpler partial derivatives. Do *not* evaluate the partial derivatives.

$$\frac{\partial L}{\partial b_{21}} = \frac{\partial L}{\partial z_3} \frac{\partial z_3}{\partial a_3} \frac{\partial a_3}{\partial b_{21}}$$

- (iv) [6 pts] Perform backpropagation on the bias  $b_{12}$  by deriving the expression for the gradient of the loss function  $L(y, z_3)$  with respect to the bias term  $b_{12}$ ,  $\frac{\partial L}{\partial b_{12}}$ , in terms of the partial derivatives  $\frac{\partial \alpha}{\partial \beta}$ , where  $\alpha$  and  $\beta$  can be any of  $L, z_i, a_i, b_{ij}, w_{ij}, x_1$  for all valid values of  $i, j$ . Your backpropagation algorithm should be as explicit as possible—that is, make sure each partial derivative  $\frac{\partial \alpha}{\partial \beta}$  cannot be decomposed further into simpler partial derivatives. Do *not* evaluate the partial derivatives.

$$\frac{\partial L}{\partial b_{12}} = \frac{\partial L}{\partial z_3} \frac{\partial z_3}{\partial a_3} \frac{\partial a_3}{\partial z_2} \frac{\partial z_2}{\partial a_2} \frac{\partial a_2}{\partial b_{12}}$$

2. In this problem we will use a neural network to classify the crosses ( $\times$ ) from the circles ( $\circ$ ) in the simple dataset shown in Figure 2a. Even though the crosses and circles are not linearly separable, we can break the examples into three groups,  $S_1$ ,  $S_2$ , and  $S_3$  (shown in Figure 2a) so that  $S_1$  is linearly separable from  $S_2$  and  $S_2$  is linearly separable from  $S_3$ . We will exploit this fact to design weights for the neural network shown in Figure 2b in order to correctly classify this training set. For all nodes, we will use the threshold activation function

$$\phi(z) = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0. \end{cases}$$

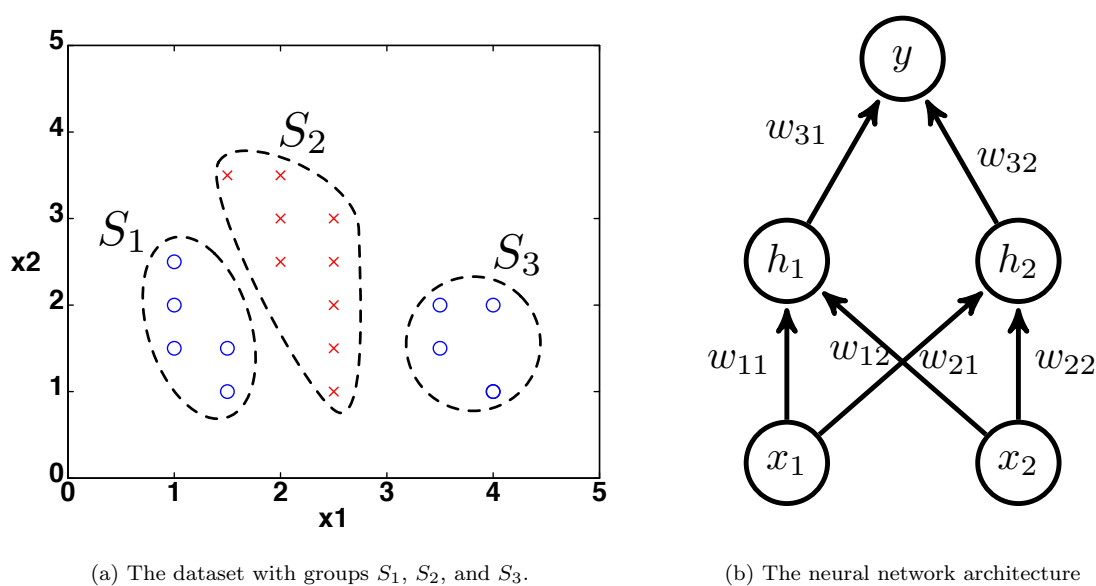


Figure 2

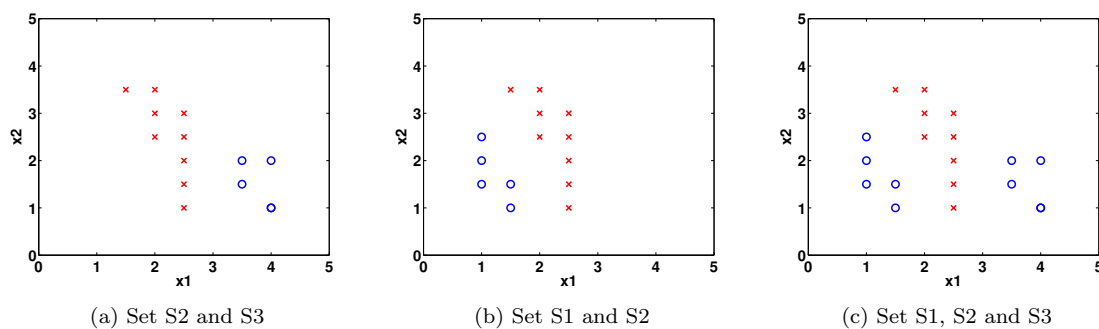


Figure 3: NN classification.

- (i) First we will set the parameters  $w_{11}$ ,  $w_{12}$  and  $b_1$  of the neuron labeled  $h_1$  so that its output  $h_1(x) = \phi(w_{11}x_1 + w_{12}x_2 + b_1)$  forms a linear separator between the sets  $S_2$  and  $S_3$ .
- (a) [1 pt.] On Fig 3a, draw a linear decision boundary that separates  $S_2$  and  $S_3$ .
- (b) [1 pt.] Write down the corresponding weights  $w_{11}$ ,  $w_{12}$ , and  $b_1$  so that  $h_1(x) = 0$  for all points in  $S_3$  and  $h_1(x) = 1$  for all points in  $S_2$ . One solution would suffice and the same applies to (ii) and (iii).
- $w_{11} = -1, w_{12} = 0, b_1 = 3$
- (ii) Next we set the parameters  $w_{21}$ ,  $w_{22}$  and  $b_2$  of the neuron labeled  $h_2$  so that its output  $h_2(x) = \phi(w_{21}x_1 + w_{22}x_2 + b_2)$  forms a linear separator between the sets  $S_1$  and  $S_2$ .
- (a) [1 pt.] On Fig 3b, draw a linear decision boundary that separates  $S_1$  and  $S_2$ .

- (b) [1 pt.] Write down the corresponding weights  $w_{21}$ ,  $w_{22}$ , and  $b_2$  so that  $h_2(x) = 0$  for all points in  $S_1$  and  $h_2(x) = 1$  for all points in  $S_2$ .

$$w_{21} = 3, w_{22} = 1, b_2 = -7$$

- (iii) Now we have two classifiers  $h_1$  (to classify  $S_2$  from  $S_3$ ) and  $h_2$  (to classify  $S_1$  from  $S_2$ ). We will set the weights of the final neuron of the neural network based on the results from  $h_1$  and  $h_2$  to classify the crosses from the circles. Let  $h_3(x) = \phi(w_{31}h_1(x) + w_{32}h_2(x) + b_3)$ .

- (a) [1 pt.] Compute  $w_{31}$ ,  $w_{32}$ ,  $b_3$  such that  $h_3(x)$  correctly classifies the entire dataset.  $w_{31} = 1, w_{32} = 1, b_3 = -1.5$

- (b) [1 pt.] Draw your decision boundary in Fig 3c.

(iv) **Back propagation**

In the above example, we need to learn the weights by according to the data. At first step, we need to get the gradients of the parameters of neural networks.

Suppose there  $m$  data points  $x_i$  with label  $y_i$ , where  $i \in [1, m]$ .  $x_i$  is a  $d \times 1$  vector and  $y_i \in \{0, 1\}$ . We use the data to train a neural network with one hidden layer:

$$\begin{aligned} h(x) &= \sigma(W_1x + b_1) \\ p(x) &= \sigma(W_2h(x) + b_2), \end{aligned}$$

where  $\sigma(x) = \frac{1}{1+\exp(-x)}$  is the sigmoid function,  $W_1$  is a  $n$  by  $d$  matrix and  $b_1$  is a  $n$  by 1 vector,  $W_2$  is a 1 by  $n$  matrix and  $b_2$  is a 1 by 1 vector.

We use cross entropy loss function and minimize the negative log likelihood to train the neural network:

$$l = \frac{1}{m} \sum_i l_i = \frac{1}{m} \sum_i -(y_i \log p_i + (1 - y_i) \log(1 - p_i)),$$

where  $p_i = p(x_i)$ ,  $h_i = h(x_i)$ .

- (a) Describe how you would drive the gradients w.r.t the parameters  $W_1$ ,  $W_2$  and  $b_1$ ,  $b_2$ . (No need to write out the detailed mathematical expression.) **Use chain rule.**
- (b) When  $m$  is large, we typically use a small sample of all the data set to estimate the gradient, this is call stochastic gradient descent (SGD). Explain why we use SGD instead of gradient descent.  
**SGD converges faster than gradient descent.**
- (c) Work out the following gradient:  $\frac{\partial l}{\partial p_i}, \frac{\partial l}{\partial W_2}, \frac{\partial l}{\partial b_2}, \frac{\partial l}{\partial h_i}, \frac{\partial l}{\partial W_1}, \frac{\partial l}{\partial b_1}$ . When deriving the gradient w.r.t. the parameters in lower layers, you can may assume the gradient in upper layers are available to you (i.e., you can use them in your equation). For example, when calculating  $\frac{\partial l}{\partial W_1}$ , you can assume  $\frac{\partial l}{\partial p_i}, \frac{\partial l}{\partial W_2}, \frac{\partial l}{\partial b_2}, \frac{\partial l}{\partial h_i}$  are known.

$$\begin{aligned} \frac{\partial l}{\partial p_i} &= \frac{1}{m} \left( -\frac{y_i}{p_i} + \frac{1-y_i}{1-p_i} \right) \\ \frac{\partial l}{\partial W_2} &= \frac{1}{m} \sum_i \frac{\partial l_i}{\partial p_i} \frac{\partial p_i}{\partial W_2} = \frac{1}{m} \sum_i \frac{\partial l_i}{\partial p_i} p_i (1-p_i) h_i^T \\ \frac{\partial l}{\partial b_2} &= \frac{1}{m} \sum_i \frac{\partial l_i}{\partial p_i} p_i (1-p_i) \\ \frac{\partial l}{\partial h_i} &= \frac{\partial l}{\partial p_i} \frac{\partial p_i}{\partial h_i} = \frac{\partial l}{\partial p_i} p_i (1-p_i) W_2^T \\ \frac{\partial l}{\partial W_1} &= \frac{1}{m} \sum_i \frac{\partial l_i}{\partial h_i} \frac{\partial h_i}{\partial W_1} = \frac{1}{m} \sum_i \left[ \frac{\partial l_i}{\partial h_i} \circ h_i \circ (1-h_i) \right] x_i^T \\ \frac{\partial l}{\partial b_1} &= \frac{1}{m} \sum_i \frac{\partial l_i}{\partial h_i} \frac{\partial h_i}{\partial b_1} = \frac{1}{m} \sum_i \frac{\partial l_i}{\partial h_i} \circ h_i \circ (1-h_i) \end{aligned}$$

3. Consider the following neural network for a 2-D input,  $x_1 \in \mathbb{R}$  and  $x_2 \in \mathbb{R}$  where:

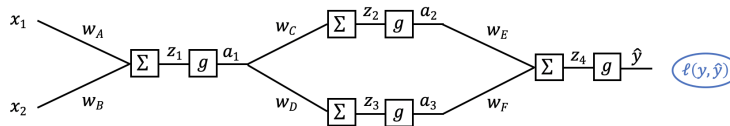


Figure 4: Neural Network

- All  $g$  functions are the same arbitrary non-linear activation function with no parameters
- $\ell(y, \hat{y})$  is an arbitrary loss function with no parameters, and:

$$z_1 = w_A x_1 + w_B x_2 \quad a_1 = g(z_1)$$

$$z_2 = w_C a_1 \quad a_2 = g(z_2)$$

$$z_3 = w_D a_1 \quad a_3 = g(z_3)$$

$$z_4 = w_E a_2 + w_F a_3 \quad \hat{y} = g(z_4)$$

**Note:** There are no bias terms in this network.

- (a) What is the chain of partial derivatives needed to calculate the derivative  $\frac{\partial \ell}{\partial w_E}$ ?

Your answer should be in the form:  $\frac{\partial \ell}{\partial w_E} = \frac{\partial?}{\partial?} \frac{\partial?}{\partial?} \dots$ . Make sure each partial derivative  $\frac{\partial?}{\partial?}$  in your answer cannot be decomposed further into simpler partial derivatives. **Do not evaluate the derivatives.** Be sure to specify the correct subscripts in your answer.

$$\frac{\partial \ell}{\partial w_E} =$$

$$\frac{\partial \ell}{\partial w_E} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_4} \frac{\partial z_4}{\partial w_E}$$

- (b) The network diagram from above is repeated here for convenience: What is the

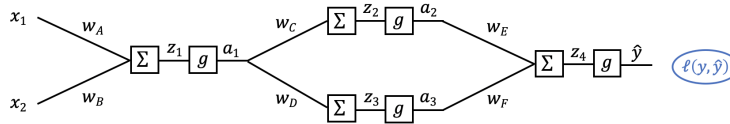


Figure 5: Neural Network

chain of partial derivatives needed to calculate the derivative  $\frac{\partial \ell}{\partial w_C}$ ?  
Your answer should be in the form:

$$\frac{\partial \ell}{\partial w_C} = \frac{\partial ?}{\partial ?} \frac{\partial ?}{\partial ?} \dots$$

Make sure each partial derivative  $\frac{\partial ?}{\partial ?}$  in your answer cannot be decomposed further into simpler partial derivatives. **Do not evaluate the derivatives.** Be sure to specify the correct superscripts in your answer.

$$\frac{\partial \ell}{\partial w_C} =$$

$$\frac{\partial \ell}{\partial w_C} = \frac{\partial \ell}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z_4} \frac{\partial z_4}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial w_C}$$

- (c) The gradient descent update step for weight  $w_c$  is:

$$w_c \leftarrow w_c - \alpha \frac{\partial Q}{\partial t} = \frac{\partial s}{\partial t}$$

where  $\alpha$  (alpha) is the learning rate (step size).

Now, we want to change our neural network objective function to add an L2 regularization term on the weights. The new objective is:

$$\ell(y, \hat{y}) + \lambda \frac{1}{2} \|w\|_2^2$$

where  $\lambda$  (lambda) is the regularization hyperparameter and  $\mathbf{w}$  is all of the weights in the neural network stacked into a single vector,  $\mathbf{x} = [w_A, w_B, w_C, w_D, w_E, w_F]^T$ . Write the right-hand side of the new gradient descent update step for weight  $w_C$  given this new objective function. You may use  $\frac{\partial \ell}{\partial w_C}$  in your answer.

**Update:**  $w_C \leftarrow \dots$

Update for  $w_C$ :  $w_C \leftarrow w_C - \alpha \left( \frac{\partial \ell}{\partial w_C} + \lambda w_C \right)$



## 5 MLE/MAP

1. Please circle **True** or **False** for the following questions, providing brief explanations to support your answer.

- (i) [2 pts] Consider the linear regression model  $y = w^T x + \epsilon$ . Assuming  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  and maximizing the conditional log-likelihood is equivalent to minimizing the sum of squared errors  $\|y - w^T x\|_2^2$ .

Circle one:      True      False

One line justification (only if False):

True. The squared error term comes from the squared term in the Gaussian distribution.

- (ii) [4 pts] Consider  $n$  data points, each with one feature  $x_i$  and an output  $y_i$ . In linear regression, we assume  $y_i \sim \mathcal{N}(wx_i, \sigma^2)$  and compute  $\hat{w}$  through MLE.

Suppose  $y_i \sim \mathcal{N}(\log(wx_i), 1)$  instead. Then the maximum likelihood estimate  $\hat{w}$  is the solution to the following equality:

$$\sum_{i=1}^n x_i y_i = \sum_{i=1}^n x_i \log(wx_i)$$

Circle one:      True      False

Brief explanation:

False. The likelihood function can be written as

$$\prod_{i=1}^n \frac{\exp(-(y_i - \log(wx_i))^2/2)}{\sqrt{2\pi}} = \frac{\exp(-\sum_{i=1}^n (y_i - \log(wx_i))^2/2)}{\sqrt{2\pi}}$$

Differentiating wrt  $w$  and setting to zero gives us

$$\sum_{i=1}^n 2(y_i - \log(wx_i)) \frac{x_i}{wx_i} = 0 \implies \sum_{i=1}^n y_i = \sum_{i=1}^n \log(wx_i)$$

2. **Select all that apply:** Which of the following are correct regarding Gradient Descent (GD). Assume data log-likelihood is  $L(\theta|X)$ , which is a function of the parameter  $\theta$ , and the objective function is negative log-likelihood .

- GD requires that  $L(\theta|X)$  is concave with respect to parameter  $\theta$  in order to converge

- GD requires that  $L(\theta|X)$  is convex with respect to parameter  $\theta$  in order to converge
- GD update rule is  $\theta \leftarrow \theta - \alpha \nabla_{\theta} L(\theta|X)$
- Given a fixed small learning rate (say  $\alpha = 10^{-10}$ ), GD will always reach the optimum after infinite iterations (assume that the objective function satisfies the convergence condition).

A

Analysis: C should replace minus with plus. D is wrong because it is possible that  $\theta$  will jump around the minimum and never reach the optimum even though  $\alpha$  is (finitely) small.

3. Let  $X_1, X_2, \dots, X_N$  be i.i.d. data from a uniform distribution over a diamond-shaped area with edge length  $\sqrt{2}\theta$  in  $\mathbb{R}^2$ , where  $\theta \in \mathbb{R}^+$  (see Figure 6). Thus,  $X_i \in \mathbb{R}^2$  and the distribution is

$$p(x|\theta) = \begin{cases} \frac{1}{2\theta^2} & \text{if } \|x\| \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where  $\|x\| = |x_1| + |x_2|$  is  $L1$  norm. Please find the maximum likelihood estimator of  $\theta$ .

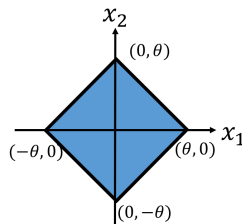


Figure 6: Area of  $\|x\| \leq \theta$

Analysis:

The likelihood function is

$$L(X_1, X_2, \dots, X_N; \theta) = \frac{1}{(2\theta^2)^N} \mathbb{1} \left\{ \max_{1 \leq i \leq N} \|X_i\| \leq \theta \right\}$$

To maximize likelihood, we want  $\theta$  to be as small as possible with the constraint of  $\max_{1 \leq i \leq N} \|X_i\| \leq \theta$ , otherwise the likelihood drops to 0. So the MLE of  $\theta$  is

$$\hat{\theta} = \max_{1 \leq i \leq N} \|X_i\|$$

4. **Short answer:** Suppose we want to model a 1-dimensional dataset of  $N$  real valued features  $(x^{(i)})$  and targets  $(y^{(i)})$  by:

$$y^{(i)} \sim \mathcal{N}(\exp(wx^{(i)}), 1)$$

Where  $w$  is our unknown (scalar) parameter and  $\mathcal{N}$  is the normal distribution with probability density function:

$$f(a)_{\mathcal{N}(\mu, \sigma^2)} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(a - \mu)^2}{2\sigma^2}\right)$$

Can the maximum conditional negative log likelihood estimator of  $w$  be solved analytically? If so, find the expression for  $w_{\text{MLE}}$ . If not, say so and write down the update rule for  $w$  in gradient descent.

Cannot be found analytically.

Taking the derivative of the negative log likelihood with respect to  $w$  yields:

$$\frac{\partial \text{NLL}}{\partial w} = \sum_i^N -x^{(i)} y^{(i)} \exp(wx^{(i)}) + x^{(i)} \exp(2wx^{(i)})$$

Update rule is thus

$$w \leftarrow w - \eta \frac{\partial \text{NLL}}{\partial w}$$

5. Assume we have  $n$  iid random variables  $x_i, i \in [1, n]$  such that each  $x_i$  belongs to a normal distribution with mean  $\mu$  and variance  $\sigma^2$ .

$$p(x_1, x_2, \dots, x_n | \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

- a) Write the log likelihood function  $l(x_1, x_2 \dots x_n | \mu, \sigma^2)$

$$\log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}\right) = \sum_{i=1}^n \left[\log\left(\frac{1}{\sqrt{2\pi\sigma}}\right) - \frac{(x_i - \mu)^2}{2\sigma^2}\right] \quad (1)$$

$$= -n \log(\sqrt{2\pi\sigma}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \quad (2)$$

- b) Derive an expression for the Maximum Likelihood Estimate for the variance ( $\sigma^2$ )

We can find the estimator by solving  $\nabla_{\sigma} l(x_1, x_2 \dots x_n | \mu, \sigma^2) = 0$ .

$$-n \frac{1}{\sqrt{2\pi\sigma}} \sqrt{2\pi} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0 \quad (3)$$

$$\frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = \frac{n}{\sigma} \quad (4)$$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = \sigma^2 \quad (5)$$

6. Assume we have a random variable that is Bernoulli distributed  $X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$ . We are going to derive its MLE. Recall that in a Bernoulli  $X = \{0, 1\}$  and the pdf of a Bernoulli is

$$p(X; \theta) = \theta^x(1 - \theta)^{1-x}$$

- a) Derive the likelihood,  $L(\theta; X_1, \dots, X_n)$

$$\begin{aligned} L(\theta; X_1, \dots, X_n) &= \prod_{i=1}^n p(X_i; \theta) \\ L(\theta; X_1, \dots, X_n) &= \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i} \\ L(\theta; X_1, \dots, X_n) &= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \end{aligned}$$

Either of the final two steps are acceptable.

- b) Derive the following formula for the log likelihood

$$l(\theta; X_1, \dots, X_n) = \left( \sum_{i=1}^n X_i \right) \log(\theta) + \left( n - \sum_{i=1}^n X_i \right) \log(1 - \theta)$$

$$\begin{aligned} l(\theta; X_1, \dots, X_n) &= \log L(\theta; X_1, \dots, X_n) \\ l(\theta; X_1, \dots, X_n) &= \log \left[ \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \right] \\ l(\theta; X_1, \dots, X_n) &= \left( \sum_{i=1}^n x_i \right) \log(\theta) + \left( n - \sum_{i=1}^n x_i \right) \log(1 - \theta) \end{aligned}$$

- c) Derive the MLE,  $\hat{\theta}$ , and show that  $\hat{\theta} = \frac{1}{n}(\sum_{i=1}^n X_i)$

Take the derivative of the log likelihood and set it to zero

$$\begin{aligned} \frac{dl}{d\theta} &= \frac{d}{d\theta} \left[ \left( \sum_{i=1}^n x_i \right) \log(\theta) + \left( n - \sum_{i=1}^n x_i \right) \log(1 - \theta) \right] = 0 \\ \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1 - \theta} &= 0 \\ \left( \sum_{i=1}^n x_i \right) (1 - \theta) - \left( n - \sum_{i=1}^n x_i \right) \theta &= 0 \\ \sum_{i=1}^n x_i - n\theta &= 0 \\ \hat{\theta} &= \frac{1}{n} \left( \sum_{i=1}^n X_i \right) \end{aligned}$$

7. Assume we have a random sample that is Exponential distributed  $X_1, \dots, X_n \sim \text{Exponential}(\theta)$ . We are going to derive the MLE for  $\theta$ . Recall that a exponential random variable  $X$  has p.d.f:

$$P(X; \theta) = \theta \exp(-\theta X).$$

- a) Derive the likelihood,  $L(\theta; X_1, \dots, X_n)$ .

$$\begin{aligned} L(\theta; X_1, \dots, X_n) &= \prod_{i=1}^n p(X_i; \theta) \\ L(\theta; X_1, \dots, X_n) &= \prod_{i=1}^n \theta \exp^{-\theta x_i} \end{aligned}$$

- b) Find  $\theta$  that maximizes  $L(\theta; X_1, \dots, X_n)$ .

$$\begin{aligned} l(\theta; X_1, \dots, X_n) &= \log L(\theta; X_1, \dots, X_n) \\ l(\theta; X_1, \dots, X_n) &= \log \left[ \prod_{i=1}^n \theta \exp^{-\theta x_i} \right] \end{aligned}$$

$$l(\theta; X_1, \dots, X_n) = \sum_{i=1}^n \log(\theta) + \sum_{i=1}^n -\theta x_i$$

$$\frac{dl}{d\theta} = \frac{d}{d\theta} \left[ \sum_{i=1}^n \log(\theta) + \sum_{i=1}^n -\theta x_i \right] = 0$$

$$\sum_{i=1}^n \frac{1}{\theta} - \sum_{i=1}^n x_i = 0$$

$$\theta = \frac{n}{\sum_{i=1}^n x_i}$$

8. For each question state **True** or **False** and give one line justifications.

- a) **T or F** The value of the Maximum Likelihood Estimate (MLE) is equal to the value of the Maximum A Posteriori (MAP) Estimate with a uniform prior.

True. Since we know posterior is proportional to the product of likelihood and prior, i.e.,

$$p(\theta|x) \propto p(x|\theta)p(\theta). \quad (6)$$

Since uniform prior gives us constant value on  $p(\theta)$ , after proper normalization, we know likelihood and posterior are the same, so do their estimator.

- b) **T or F** The bias of the Maximum Likelihood Estimate (MLE) is typically less than or equal to the bias of the Maximum A Posteriori (MAP) Estimate.

True. Since MAP estimate allows us to incorporate more prior knowledge, it is likely to be more biased.

- c) **T or F** The MAP estimate is always better than the MLE.

False. When the prior is chosen poorly, it will take the MAP estimate longer to converge to a good estimate because it needs to see enough examples to overcome the bad prior.

- d) **T or F** In the limit as  $n$  (the number of samples) increases, the MAP and MLE estimates become the same.

True. As the number of examples increases, the data likelihood goes to zero very quickly, while the magnitude of the prior stays the same. In the limit, the prior plays an insignificant role in the MAP estimate and the two estimates will converge.

- e) **T or F** Naive Bayes can only be used with MAP estimates, and not MLE estimates.

False. Naive Bayes can be used with any technique for estimating the parameters of a distribution. In homework 2, we used both the MAP and MLE estimates.

## 6 Probability, Naive Bayes and MLE

### 6.1 Probability

1. For each question, circle the correct option.

1. Which of the following expressions is equivalent to  $p(A|B, C, D)$ ?

(a)  $\frac{p(A,B,C,D)}{p(C|B,D)p(B|D)p(D)}$

(b)  $\frac{p(A,B,C,D)}{p(B,C)p(D)}$

(c)  $\frac{p(A,B,C,D)}{p(B,C|D)p(B)p(C)}$

Answer is (a).  $p(A|BCD) = \frac{p(A,B,C,D)}{p(B,C,D)} = \frac{p(A,B,C,D)}{p(C|B,D)p(B,D)} = \frac{p(A,B,C,D)}{p(C|B,D)p(B|D)p(D)}$

2. Let  $\mu$  be the mean of some probability distribution.  $p(\mu)$  is always non-zero.

(a) True

(b) False

No, taking the example of a distribution that put 0.5 probability on +1 and 0.5 probability on -1. Though their mean is 0,  $p(0)$  is zero.

2. Assume we have a sample space  $\Omega$ . Just state **T** or **F**, no justification needed.

1. If events  $A$ ,  $B$ , and  $C$  are disjoint then they are independent.

False. If they are disjoint, i.e. mutually exclusive, they are very dependent! (what does disjoint mean in terms of the probabilities of  $A$ ,  $B$ , and  $C$ ? What about independent?)

2.  $P(A|B) \propto \frac{P(A)P(B|A)}{P(A|B)}$ .

False  $P(A|B) \propto \frac{P(A)P(B|A)}{P(B)}$

3.  $P(A \cup B) \leq P(A)$ .

False  $P(A \cup B) \geq P(A)$

4.  $P(A \cap B) \geq P(A)$ .

False  $P(A \cap B) \leq P(A)$

### 6.2 Naive Bayes

1. Consider the following data. It has 4 features  $\mathbf{X} = (x_1, x_2, x_3, x_4)$  and 3 labels  $(+1, 0, -1)$ . Assume that the probabilities  $p(\mathbf{X}|y)$  and  $p(y)$  are both Bernoulli distributions. Answer the questions that follow under the Naive Bayes assumption.

$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	1	0	1	+1
0	1	1	0	+1
1	0	1	1	0
0	1	1	1	0
0	1	0	0	-1
1	0	0	1	-1
0	0	1	1	-1

1. Compute the Maximum Likelihood Estimate for  $p(x_i = 1|y), \forall i \in [1, 4], \forall y \in \{+1, 0, -1\}$ .

	$y = +1$	$y = 0$	$y = -1$
$x_1 = 1$	0.5	0.5	1/3
$x_2 = 1$	1	0.5	1/3
$x_3 = 1$	0.5	1	1/3
$x_4 = 1$	0.5	1	2/3

2. Compute the Maximum Likelihood Estimate for the prior probabilities  $p(y = +1), p(y = 0), p(y = -1)$

$$p(y = +1) = \frac{2}{7}, p(y = 0) = \frac{2}{7} \text{ and } p(y = -1) = \frac{3}{7}.$$

3. Use the values computed in the above two parts to classify the data point  $(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1)$  as either belonging to class +1, 0 or -1

According to naive bayes assumption, samples are independent given  $y$ , thus we can write the conditional joint probability as

$$p(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1) = p(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1|y)p(y) \quad (7)$$

$$= p(y) \prod_{i=1}^4 p(x_i = 1|y). \quad (8)$$

We calculate the probability given different value on  $y$  and pick the  $y$  that gives us largest probability.

$$p(y = +1) \prod_{i=1}^4 p(x_i = 1|y = +1) = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{7} = \frac{1}{28} \quad (9)$$

$$p(y = 0) \prod_{i=1}^4 p(x_i = 1|y = 0) = \frac{1}{2} \cdot \frac{1}{2} \cdot 1 \cdot 1 \cdot \frac{2}{7} = \frac{1}{14} \quad (10)$$

$$p(y = -1) \prod_{i=1}^4 p(x_i = 1|y = -1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{7} = \frac{2}{189} \quad (11)$$

Since  $y = 0$  yields the largest value, we classify the data as  $\hat{y} = 0$ .

2. You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a machine learning classifier to predict the sex of a student, so you randomly split the data into a training set and a testing set. Here are the specifications of the data set:

- sex  $\in$  {male, female}
- height  $\in$  [0,300] centimeters
- hair  $\in$  {brown, black, blond, red, green}
- 3240 men in the data set
- 6760 women in the data set

Under the assumptions necessary for Naive Bayes (not the distributional assumptions you might naturally or intuitively make about the dataset) answer each question with **T** or **F** and a one sentence explanation of your answer:

1. **T or F:** Height is a continuous valued variable. Therefore Naive Bayes is not appropriate since it cannot handle continuous valued variables.

False. Naive Bayes can handle both continuous and discrete values as long as the appropriate distributions are used for conditional probabilities. For example, Gaussian for continuous and Bernoulli for discrete

2. **T or F:** Since there is not a similar number of men and women in that dataset Naive Bayes will have high test error.

False. Since the data was randomly split, the same proportion of male and female will be in the training and testing sets. Thus this discrepancy will not affect testing error.

3. **T or F:**  $p(\text{height}|\text{sex, hair}) = p(\text{height}|\text{sex})$ .

True. This results from the conditional independence assumption required for Naive Bayes.

4. **T or F:**  $p(\text{height, hair}|\text{sex}) = p(\text{height}|\text{sex}) * p(\text{hair}|\text{sex})$ .

True. This results from the conditional independence assumption required for Naive Bayes.

### 6.3 Naive Bayes, Logistic Regression

1. Suppose you wish to learn  $P(Y|X_1, X_2, X_3)$ , where  $Y, X_1, X_2$  and  $X_3$  are all boolean-valued random variables. You consider both Naive Bayes and Logistic Regression as possible approaches.

For each of the following, answer True or False, and give a *one sentence* justification for your answer.



1. T or F: In this case, a good choice for Naive Bayes would be to implement a Gaussian Naive Bayes classifier.

Answer: False. Given the  $X_i$  are boolean, it is better to model  $P(X_i|Y)$  with a Bernoulli rather than Gaussian distribution

2. T or F: To learn  $P(Y|X_1, X_2, X_3)$  using Naive Bayes, you must make conditional independence assumptions, including the assumption that  $Y$  is conditionally independent of  $X_1$  given  $X_2$ .

Answer: False. Naive Bayes assume  $(\forall i \neq j) X_i$  is conditionally independent of  $X_j$  given  $Y$ .

3. T or F: Logistic regression is certain to be the better choice in this case.

Answer: False. It will depend on the number of training examples available, and whether the Naive Bayes assumptions are satisfied.

## 2. Parameter estimation

1. How many parameters must be estimated for your Gaussian Naive Bayes classifier, and what are they (i.e., please list them).

7 - We need  $P(Y = 1)$  and  $P(Y = 0) = 1 - P(Y = 1)$ . Then we need  $P(X_i = 0|Y = 0), P(X_i = 1|Y = 0), P(X_i = 0|Y = 1), P(X_i = 1|Y = 1)$  for  $i \in \{1, 2, 3\}$ .

But given the binary parameters,  $P(X_i = 1|Y = 0) + P(X_i = 0|Y = 0) = 1$ , so we only need to estimate one of these conditional probabilities ( $P(X_i = 0|Y = 0)$ , for example). So in total we need only  $2*3 + 1 = 7$

2. How many parameters must be estimated for your Logistic Regression classifier, and what are they (i.e., please list them).

4 - Weights for: Bias,  $X_1$ ,  $X_2$ , and  $X_3$

3. T or F: We can train Naive Bayes using maximum likelihood estimates for each parameter, but not MAP estimates. Justify your answer *in one sentence*.

False: MAP estimates are just MLE spiced up with priors on the parameters of  $P(X_i|Y_j)$  (prior knowledge that we can inject into the model), so there's no reason we can't add it in.

4. T or F: We can train Logistic Regression using maximum likelihood estimates for each parameter, but not MAP estimates. Justify your answer *in one sentence*.

False: We can assign priors on the regression model by assuming  $y = w^T x + \epsilon$  where  $\epsilon$  is "noise" from a distribution (e.g. Gaussian)

3. Mixing discrete and continuous variables. Suppose we add a numeric, real-valued variable  $X_4$  to our problem. Note we now have a mix of some discrete-valued  $X_i$  and one continuous  $X_i$ .

1. Explain in *two sentences* why we can no longer use Naive Bayes, or if we can, how we would modify our first solution.

We can't use our original model because a Bernoulli distribution can't model the new data.

We must modify our solution so that a different Naive Bayes model is trained on the continuous variables, using a different distribution than a Bernoulli one (i.e. a Gaussian), and then the result is a multiplication of the two.

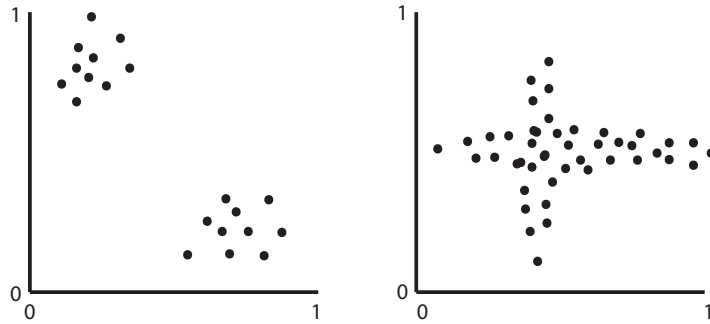
Since our assumption is that each parameter is conditionally independent anyhow, we can multiply the results of the two models together safely.

2. Explain in *two sentences* why we can no longer use Logistic Regression, or if we can, how we would modify our first solution.

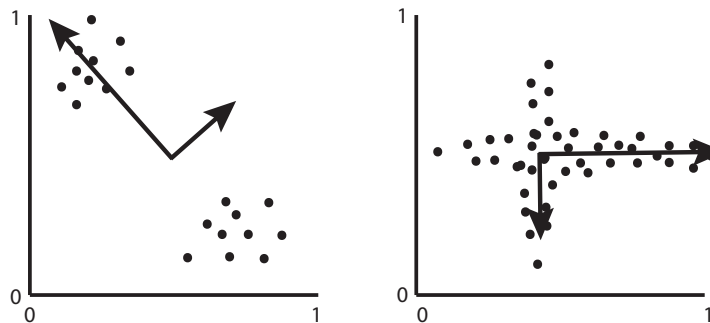
Assuming the discrete variables were already transformed into one-hot features, we can simply add one weight per continuous feature to our model.

## 7 Principal Component Analysis

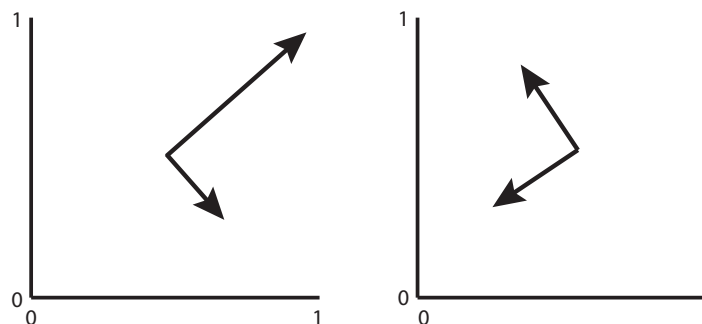
1. (i) [5 pts] Consider the following two plots of data. Draw arrows from the mean of the data to denote the direction and relative magnitudes of the principal components.



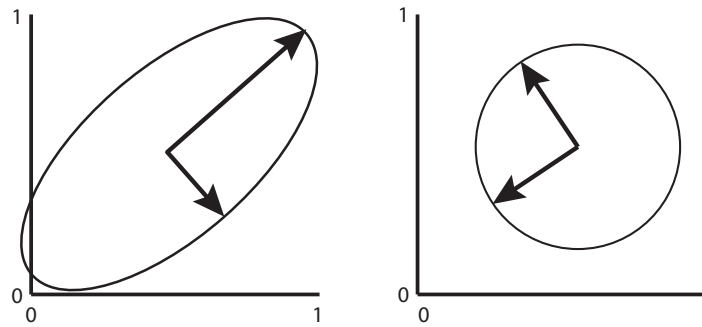
Solution:



- (ii) [5 pts] Now consider the following two plots, where we have drawn only the principal components. Draw the data ellipse or place data points that could yield the given principal components for each plot. Note that for the right hand plot, the principal components are of equal magnitude.



Solution:



2. Circle one answer and explain.

In the following two questions, assume that using PCA we factorize  $X \in \mathbb{R}^{n \times m}$  as  $Z^T U \approx X$ , for  $Z \in \mathbb{R}^{m \times n}$  and  $U \in \mathbb{R}^{m \times m}$ , where the rows of  $X$  contain the data points, the rows of  $U$  are the prototypes/principal components, and  $Z^T U = \hat{X}$ .

- (i) [2 pts] Removing the last row of  $U$  and  $Z$  will still result in an approximation of  $X$ , but this will never be a better approximation than  $\hat{X}$ .

Circle one:     True     False

True.

- (ii) [2 pts]  $\hat{X} \hat{X}^T = Z^T Z$ .

Circle one:     True     False

True.

- (iii) [2 pts] The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

Circle one:     True     False

False

- (iv) [2 pts] The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

Circle one:     True     False

False

## 8 K-Means

1. For **True or False** questions, circle your answer and justify it; for **QA** questions, write down your answer.

(i) For a particular dataset and a particular  $k$ ,  $k$ -means always produce the same result, if the initialized centers are the same. Assume there is no tie when assigning the clusters.

True

False

**Justify your answer:**

---

True. Every time you are computing the completely same distances, so the result is the same.

(ii)  $k$ -means can always converge to the global optimum.

True

False

**Justify your answer:**

---

False. It depends on the initialization. Random initialization could possibly lead to a local optimum.

(iii)  $k$ -means is not sensitive to outliers.

True

False

**Justify your answer:**

---

False.  $k$ -means is quite sensitive to outliers, since it computes the cluster center based on the mean value of all data points in this cluster.

(iv)  $k$  in  $k$ -nearest neighbors and  $k$ -means has the same meaning.

True

False

**Justify your answer:**

---

False. In knn, k is the number of data points we need to look at when classifying a data point. In k-means, k is the number of clusters.

- (v) What's the biggest difference between k-nearest neighbors and k-means?

**Write your answer in one sentence:**

---

knn is a supervised algorithm, while k-means is unsupervised.

2. In k-means, random initialization could possibly lead to a local optimum with very bad performance. To alleviate this issue, instead of initializing all of the centers completely randomly, we decide to use a smarter initialization method. This leads us to k-means++.

The only difference between k-means and k-means++ is the initialization strategy, and all of the other parts are the same. The basic idea of k-means++ is that instead of simply choosing the centers to be random points, we sample the initial centers iteratively, each time putting higher probability on points that are far from any existing center. Formally, the algorithm proceeds as follows.

**Given:** Data set  $x^{(i)}, i = 1, \dots, N$

**Initialize:**

$$\mu^{(1)} \sim \text{Uniform}(\{x^{(i)}\}_{i=1}^N)$$

For  $j = 2, \dots, k$

Computing probabilities of selecting each point

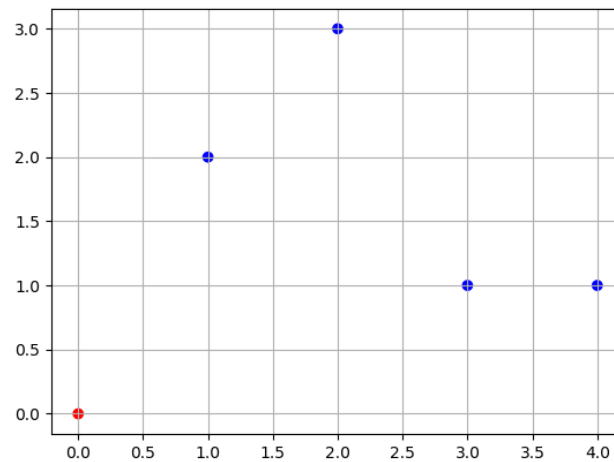
$$p_i = \frac{\min_{j' < j} \|\mu^{(j')} - x^{(i)}\|_2^2}{\sum_{i'=1}^N \min_{j' < j} \|\mu^{(j')} - x^{(i')}\|_2^2}$$

Select next center given the appropriate probabilities

$$\mu^{(j)} \sim \text{Categorical}(\{x^{(i)}\}_{i=1}^N, \mathbf{p}_{1:N})$$

Note: n is the number of data points, k is the number of clusters. For cluster 1's center, you just randomly choose one data point. For the following centers, every time you initialize a new center, you will first compute the distance between a data point and the center closest to this data point. After computing the distances for all data points, perform a normalization and you will get the probability. Use this probability to sample for a new center.

Now assume we have 5 data points (n=5): (0, 0), (1, 2), (2, 3), (3, 1), (4, 1). The number of clusters is 3 (k=3). The center of cluster 1 is randomly chosen as (0, 0). These data points are shown in the figure below.



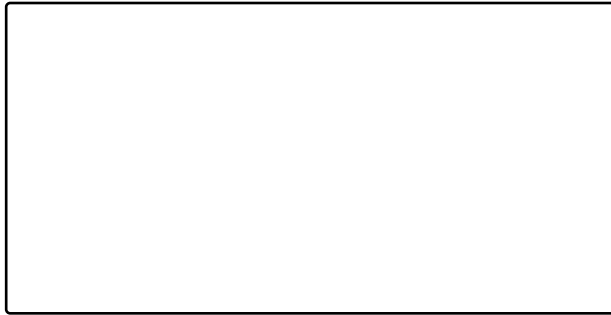
- (i) [5 pts] What is the probability of every data point being chosen as the center for cluster 2? (The answer should contain 5 probabilities, each for every data point)

(0, 0): 0  
 (1, 2): 0.111  
 (2, 3): 0.289  
 (3, 1): 0.222  
 (4, 1): 0.378

- (ii) [1 pts] Which data point is mostly likely chosen as the center for cluster 2?

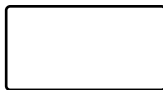
(4, 1) is mostly likely chosen.

- (iii) [5 pts] Assume the center for cluster 2 is chosen to be the most likely one as you computed in the previous question. Now what is the probability of every data point being chosen as the center for cluster 3? (The answer should contain 5 probabilities, each for every data point)



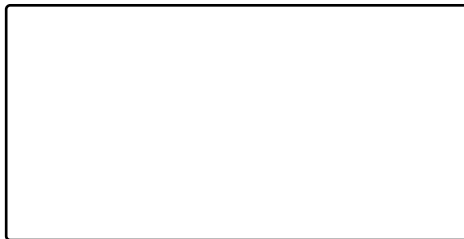
(0, 0): 0  
(1, 2): 0.357  
(2, 3): 0.571  
(3, 1): 0.071  
(4, 1): 0

- (iv) [1 pts] Which data point is mostly likely chosen as the center for cluster 3?



(2, 3) is mostly likely chosen.

- (v) [3 pts] Assume the center for cluster 3 is also chosen to be the most likely one as you computed in the previous question. Now we finish the initialization for all 3 centers. List the data points that are classified into cluster 1, 2, 3 respectively.



cluster 1: (0, 0)  
cluster 2: (1, 2), (2, 3)  
cluster 3: (3, 1), (4, 1)

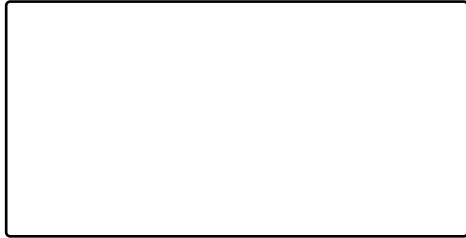
- (vi) [3 pts] Based on the above clustering result, what's the new center for every cluster?





center for cluster 1: (0, 0)  
center for cluster 2: (1.5, 2.5)  
center for cluster 3: (3.5, 1)

- (vii) [2 pts] According to the result of (ii) and (iv), explain how does k-means++ alleviate the local optimum issue due to initialization?



k-means++ tends to initialize new cluster centers with the data points that are far away from the existing centers, to make sure all of the initial cluster centers stay away from each other.

## 9 Kernel Methods

1. [3 pts.] Applying the kernel trick enables features to be mapped into a higher dimensional space, at a cost of higher computational complexity to operate in the higher dimensional space.

Circle one:      True      False

False. We didn't increase computational complexity, that's the whole point of kernel trick.

2. [3 pts.] Since the VC dimension for an SVM with a Radial Base Kernel is infinite, such an SVM must have a larger generalization error than an SVM without kernel which has a finite VC dimension.

Circle one:      True      False

False. The learning theory gives a trivial upper bound for generalization error when VC dimension is infinite. It doesn't imply that RBF kernel won't work in practice.

3. [3 pts.] Suppose  $\phi(x)$  is an arbitrary feature mapping from input  $x \in \mathcal{X}$  to  $\phi(x) \in \mathbb{R}^N$  and let  $K(x, z) = \phi(x) \cdot \phi(z)$ . Then  $K(x, z)$  will always be a valid kernel function.

Circle one:      True      False

True. This is the definition of a kernel function.

4. [3 pts.] Suppose  $\phi(x)$  is the feature map induced by a polynomial kernel  $K(x, z)$  of degree  $d$ , then  $\phi(x)$  should be a  $d$ -dimensional vector.

Circle one:      True      False

False. The dimension of  $\phi(x)$  is not a constant but increases with the dimension of  $x$ . For example if  $x \in \mathbb{R}^2$  and  $K(x, z) = (1 + x^T z)^2$  has a degree of  $d = 2$ , then  $\phi(x) = (1, x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)^T$ .

5. [3 pts.] The decision boundary that we get from a Gaussian Naive Bayes model with class-conditional variance is quadratic. Can we in principle reproduce this with an SVM and a polynomial kernel?

Circle one:      Yes      No

Yes. Quadratic decision boundaries can be reproduced with an SVM with polynomial kernel of degree two.

6. Let's go kernelized!

(a) **Perceptron review.** Assume we have a binary classification task with dataset  $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{\infty}$  where  $x^{(i)} \in \mathbb{R}^d$  and  $y^{(i)} \in \{-1, 1\}$ . Recall that the perceptron learns a linear classifier  $y = \text{sign}(w^T x)$  by applying the following algorithm.

**Algorithm 1:** Perceptron algorithm

---

```

Initialize the weights  $w = 0$ ;
for  $i = 1, 2, \dots$  do
    Predict  $\hat{y}^{(i)} = \text{sign}(w^T x^{(i)})$ ;
    if  $\hat{y}^{(i)} \neq y^{(i)}$  then
        Update  $w = w + y^{(i)}x^{(i)}$ ;
end
Final classifier:  $h(x) = \text{sign}(w^T x)$ 

```

---

Show that the final weight vector  $w$  is a linear combination of all the samples  $x^{(i)}$  ( $i = 1, 2, \dots, T$ ) it has been trained on, and hence for prediction we can write  $w^T x$  in the form of  $w^T x = \sum_{i=1}^T \alpha_i K(x^{(i)}, x)$  for some  $\alpha_i$  where  $K(x, z) = x^T z$ .

The final weight can be written as  $w = \alpha_1 x^{(1)} + \dots + \alpha_T x^{(T)}$  where  $\alpha_i = y^{(i)}$  if a mistake is made at iteration  $i$  and 0 otherwise. Hence  $w^T x = \sum_{i=1}^T \alpha_i (x^{(i)})^T x = \sum_{i=1}^T \alpha_i K(x^{(i)}, x)$ .

- (b) **Kernelized perceptron.** Now we are going to introduce a kernel function  $K(x, z)$  to kernelize the perceptron algorithm. Based on your findings in the previous question, fill in the blanks below to complete the kernelized perceptron algorithm using the kernel  $K(x, z)$ . Assume the training loop stops after it has seen  $T$  training samples.

**Algorithm 2:** Kernelized Perceptron

---

```

Initialize _____;
for  $i = 1, 2, \dots$  do
    Predict  $\hat{y}^{(i)} =$  _____;
    if  $\hat{y}^{(i)} \neq y^{(i)}$  then
        _____;
end
Final classifier:  $h(x) =$  _____

```

---

- (1) Initialize all  $\alpha_i$  to 0;
- (2)  $\hat{y}^{(i)} = \text{sign}(\sum_{j=1}^{i-1} \alpha_j K(x^{(j)}, x^{(i)}))$ ;
- (3) Update  $\alpha_i = \alpha_i + y^{(i)}$  (or Set  $\alpha_i = y^{(i)}$ );
- (4)  $h(x) = \text{sign}(\sum_{j=1}^T \alpha_j K(x^{(j)}, x))$ ;

- (c) **Short answer.** Describe one advantage and one disadvantage of using kernelized perceptron compared to using vanilla perceptron.

Advantage example: kernels can introduce complex features to perceptron, so that non-linear decision boundaries can be learned and the resulting model should be more powerful.

Disadvantage example: kernelized perceptron has to store all training samples it has seen. This requires a lot of storage and yields a higher time complexity for prediction when the number of training samples is large.

7. Suppose we have six training samples that lie in a two-dimensional space as is shown in Figure 7a. Four of them belong to the blue class:  $(0, 0.5)$ ,  $(0, 2)$ ,  $(0.5, 0)$ ,  $(2, 0)$ , and two of them belong to the red class:  $(\sqrt{2}/2, \sqrt{2}/2)$ ,  $(1.5, 1.5)$ . Unfortunately, this dataset is not linearly separable. You recall that kernel trick is one technique you can take advantage of to address this problem. The trick uses a kernel function  $K(x, z)$  which implicitly defines a feature map  $\phi(x)$  from the original space to the feature space. Consider the following normalized kernel:

$$K(x, z) = \frac{x^T z}{\|x\|_2 \|z\|_2}.$$

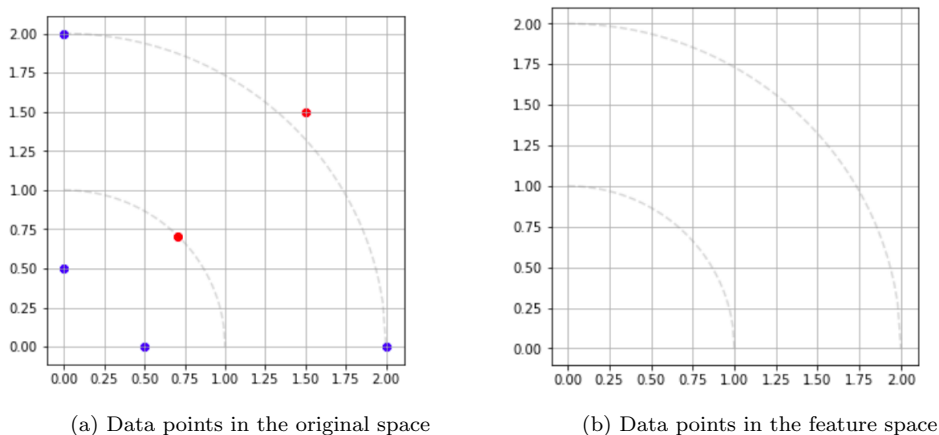


Figure 7: Data points in two spaces

- (a) What is the feature map  $\phi(x)$  that corresponds to this kernel? Draw  $\phi(x)$  for each training sample in Figure 7b.

$\phi(x) = x/\|x\|_2$ . Blue points are mapped to  $(0, 1)$  and  $(1, 0)$ . Red points are mapped to  $(\sqrt{2}/2, \sqrt{2}/2)$ .

- (b) The samples should now be linearly separable in the feature space. The classifier in the feature space that gives the maximum margin can be represented as a line  $w^T x + \alpha = 0$ . Draw the decision boundary of this classifier in Figure 7b. What are the coefficients in the weight vector  $w = (w_1, w_2)^T$ ? Hint: you don't need to compute them.

$w = (1, 1)^T$  due to symmetry. Observe that  $1 \leq -\alpha \leq \sqrt{2}$ , so the exact value of  $\alpha$  if anyone is interested should be  $-(1 + \sqrt{2})/2$ .

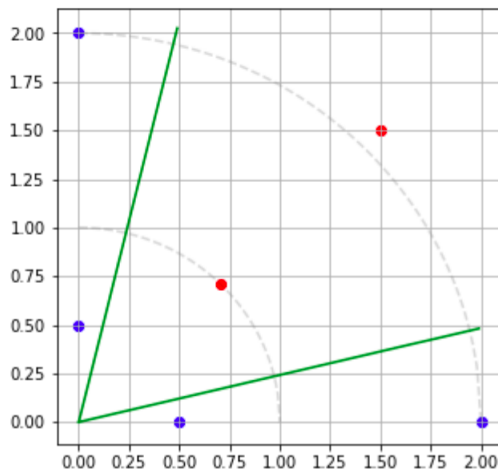
- (c) Now we map the decision boundary obtained in (b) back to the original space. Write down the corresponding boundary in the original space in the format of an explicit equation. You can keep  $\alpha$  in your equation. Try to plot its rough shape in Figure 7a.

The decision boundary in the original space is

$$w^T \phi(x) + \alpha = 0 \implies \frac{x_1 + x_2}{\sqrt{x_1^2 + x_2^2}} + \alpha = 0 \implies (x_1 + x_2)^2 = \alpha^2(x_1^2 + x_2^2)$$

$$\implies x_1^2 + x_2^2 - \frac{2}{\alpha^2 - 1}x_1x_2 = 0 \implies x_2 = \frac{\eta \pm \sqrt{\eta^2 - 4}}{2}x_1 \quad (\eta = \frac{2}{\alpha^2 - 1} > 2)$$

The final step is obtained by solving the quadratic equation. It's okay if you didn't work to the simplest form. So the decision boundary would be two straight lines. A possible rough shape is shown below.



## 10 Recommender Systems

1. [5pts] Applied to the Netflix Prize problem, which of the following methods does NOT always require side information about the users and the movies?

Select all that apply:

- Neighborhood methods
- Content filtering
- Latent factor methods
- Collaborative filtering
- None of the above

ACD

2. [5pts] Select all that apply:

- Using matrix factorization, we can embed both users and items in the same space
- Using matrix factorization, we can embed either solely users or solely items in the same space, as we cannot combine different types of data
- In a rating matrix of users by books that we are trying to fill up, the best-known solution is to fill the empty values with 0s and apply PCA, allowing the dimensionality reduction to make up for this lack of data
- Alternating minimization allows us to minimize over two variables
- Alternating minimization avoids the issue of getting stuck in local minima
- If the data is multidimensional, then overfitting is extremely rare
- Nearest neighbor methods in recommender systems are restricted to using euclidian distance for their distance metric
- None of the above

AD

Filling empty values with 0s is not ideal since we are assuming data values that are not necessarily true. Thus, we cannot apply PCA when there is missing values.

Alternating minimization can still get stuck at a local minimum.

Both euclidian distance and cosine similarity are valid metrics.

3. [5pts] Your friend Duncan wants to build a recommender system for his new website DuncTube, where users can like and dislike videos that are posted there. In order to build his system using collaborative filtering, he decides to use Non-Negative Matrix Factorization. What is an issue with Duncan's approach, and what could he change about the website *or* the algorithm in order to fix it?

Since Duncan's website incorporates negative responses directly, NNMF can't be used to model these sorts of responses (since NNMF enforces that both the original and the factored matrices are all non-negative). To fix this, Duncan would either have to remove the dislike option from his website, OR use a different matrix factorization algorithm like SVD.

4. **[3pts]** You and your friends want to build a movie recommendation system based on collaborative filtering. There are three websites (A, B and C) that you decide to extract users rating from. On website A, the rating scale is from 1 to 5. On website B, the rating scale is from 1 to 10. On website C, the rating scale is from 1 to 100. Assume you will have enough information to identify users and movies on one website with users and movies on another website. Would you be able to build a recommendation system? And briefly explain how would you do it?

Yes. We would be able to do it. First, Normalize the ratings score within certain range. (E.g. re-scale each dataset ratings to a 0-1 range). After that, combine users ratings of the three websites by matching movies and users. With users rating, we could conduct Matrix Factorization to predict the missing ratings for users. (Or Neighborhood method)

5. **[3pts]** What is the difference between collaborative filtering and content filtering?

Content filtering assumes access to side information about items and content filtering does not.

## 11 GMM and EM

1. Which of the quantities are updated in the E-step when EM algorithm is used to fit a Gaussian mixture model?

Select all that apply.

- Mixture probabilities (sometimes denoted as  $\pi$ ).
- Assignment probabilities (sometimes denoted as  $y$  or  $z$ ).
- Cluster centers (sometimes denoted as  $\mu$ ).
- Cluster variances (sometimes denoted as  $\sigma$  or  $\Sigma$ ).
- None of the above

Assignment probabilities



## 2. Bernoulli Mixture Models

There are two coins  $\mathcal{C}_1$  and  $\mathcal{C}_2$  with unknown probabilities of heads, denoted by  $q_1$  and  $q_2$  respectively.  $\mathcal{C}_1$  is chosen with probability  $\tau$  and  $\mathcal{C}_2$  is chosen with probability  $1 - \tau$ . The chosen coin is flipped once and the result is recorded. The experiment is repeated five times and the final result is recorded as  $X = \{H, H, T, H, T\}$ . You wish to use the EM algorithm to estimate the parameters  $\theta = [q_1, q_2, \tau]$ . Suppose  $Z$  denotes the hidden variable where  $z^{(i)} = 1$  if  $\mathcal{C}_1$  was tossed for the  $i$ -th flip and 0 if  $\mathcal{C}_2$  was tossed. You start off with an initial guess of  $q_1^{(0)} = \frac{1}{4}$ ,  $q_2^{(0)} = \frac{2}{3}$ , and  $\tau^{(0)} = \frac{1}{2}$ .

In the following questions, please write your answer as an **exact irreducible fraction**  $\frac{a}{b}$  where  $a, b$  are co-prime integers. For example, if your answer is 0.4125, please write

it as  $\frac{33}{80}$ . The horizontal bar for the fraction  $\frac{\square}{\square}$  is already present in the answer box.

- (a) [1 pt] **E step:** Compute  $p(z^{(i)} = 1 \mid x^{(i)} = H, \theta^{(0)})$

$\frac{3}{11}$

- (b) [1 pt] **E step:** Compute  $p(z^{(i)} = 1 \mid x^{(i)} = T, \theta^{(0)})$

$\frac{9}{13}$

- (c) [2 pt] **M step:** Compute  $\tau^{(1)}$  Hint:  $\theta^{(t+1)} = \arg \max_{\theta} \mathbb{E}_{Z \mid X, \theta^{(t)}} [\log p(X, Z \mid \theta)]$

$\frac{63}{143}$

- (d) [2 pt] **M step:** Compute  $q_1^{(1)}$

$\frac{13}{35}$

- (e) [2 pt] **M step:** Compute  $q_2^{(1)}$

$$\frac{39}{50}$$

This is an optional workbox for the above questions. Feel free to leave this blank, we will only grade this to assign partial credit in the case your above answers are incorrect. If you choose to add work please label it clearly.

