

# HOMework 6: LEARNING THEORY AND GENERATIVE MODELS

10-301/10-601 Introduction to Machine Learning (Fall 2023)

<https://www.cs.cmu.edu/~mgormley/courses/10601/>

OUT: Friday, October 27th

DUE: Friday, November 3rd

TAs: Alisa, Annie, Bhargav, Erin, Sebastian, Markov

Homework 6 covers topics on Learning Theory, MLE/MAP, Naive Bayes, and Generative vs. Discriminative Models. The homework includes multiple choice, True/False, and short answer questions. There will be no consistency points in general, so please make sure to double check your answers to all parts of the questions!

## START HERE: Instructions

- **Collaboration Policy:** Please read the collaboration policy here: <http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html>
- **Late Submission Policy:** For this homework, you will only have 2 late days instead of the usual 3. This allows us to provide feedback before the exam. See the late submission policy here: <http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html>
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. If your scanned submission misaligns the template, there will be a 5% penalty. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader.

## Instructions for Specific Problem Types

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

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

- ☒ Matt Gormley
- ☐ 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?

- ☒ Henry Chai
- ☐ 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?

10-601

10-~~6~~301

## Written Questions (100 points)

### 1 $\text{\LaTeX}$ Bonus Point and Template Alignment (1 points)

1. (1 point) **Select one:** Did you use  $\text{\LaTeX}$  for the entire written portion of this homework?

☐ Yes

☐ No

2. (0 points) **Select one:** I have ensured that my final submission is aligned with the original template given to me in the handout file and that I haven't deleted or resized any items or made any other modifications which will result in a misaligned template. I understand that incorrectly responding yes to this question will result in a penalty equivalent to 2% of the points on this assignment.

**Note:** Failing to answer this question will not exempt you from the 2% misalignment penalty.

☐ Yes

### 2 Learning Theory (19 points)

1. Neural the Narwhal is given a classification task to solve, which he decides to use a decision tree learner with 2 binary features  $X_1$  and  $X_2$ . On the other hand, you think that Neural should not have used a decision tree. Instead, you think it would be best to use logistic regression with 16 real-valued features in addition to a bias term. You want to use PAC learning to check whether you are correct. You first train your logistic regression model on  $N$  examples to obtain a training error  $\hat{R}$ .

- (a) (1 point) Which of the following case of PAC learning should you use for your logistic regression model?

☐ Finite and realizable

☐ Finite and agnostic

☐ Infinite and realizable

☐ Infinite and agnostic

- (b) (2 points) What is the upper bound on the true error  $R$  in terms of  $\hat{R}$ ,  $\delta$ , and  $N$ ? You may use big- $\mathcal{O}$  notation if necessary. Write only the final answer. Your work will *not* be graded.

**Note:** Your answer may not contain any other symbols.

Your Answer

(c) (3 points) **Select one:** You want to argue your method has a lower bound on the true error as compared to the Neural's true error bound. Assume that you have obtained enough data points to satisfy the PAC criterion with the same  $\epsilon$  and  $\delta$  as Neural. Which of the following is true?

- ☐ Neural's model will always classify unseen data more accurately because it only needs 2 binary features and therefore is simpler.
- ☐ You must first regularize your model by removing 14 features to make any comparison at all.
- ☐ It is sufficient to show that the VC dimension of your classifier is higher than that of Neural's, therefore having a lower bound for the true error.
- ☐ It is necessary to show that the training error you achieve is lower than the training error Neural achieves.

2. In lecture, we saw that we can use our sample complexity bounds to derive bounds on the true error for a particular algorithm. Consider the sample complexity bound for the infinite, agnostic case:

$$N = O\left(\frac{1}{\epsilon^2} \left[ \text{VC}(\mathcal{H}) + \log \frac{1}{\delta} \right]\right).$$

(a) (2 points) What is the big- $\mathcal{O}$  bound of  $\epsilon$  in terms of  $N$ ,  $\delta$ , and  $\text{VC}(\mathcal{H})$ ?

**Note:**  $A = \mathcal{O}(B)$  (for some value  $B$ )  $\Leftrightarrow$  there exists a constant  $c \in \mathbb{R}$  such that  $A \leq cB$ .

Your Answer

- (b) (2 points) Now, using the definition of  $\epsilon$  (i.e.  $|R(h) - \hat{R}(h)| \leq \epsilon$ ) and your answer to part a, prove that with probability at least  $(1 - \delta)$ :

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N} \left[ \text{VC}(\mathcal{H}) + \log \frac{1}{\delta} \right]}\right).$$

Your Answer

3. (3 points) Consider the hypothesis space of functions that map  $M$  binary attributes to a binary label. A function  $f$  in this space can be characterized as  $f : \{0, 1\}^M \rightarrow \{0, 1\}$ . Neural the Narwhal says that regardless of the value of  $M$ , a hypothesis class containing functions in this space can always shatter  $2^M$  points. Is Neural wrong? If so, provide a counterexample. If Neural is right, briefly explain why in 1-2 *concise* sentences.

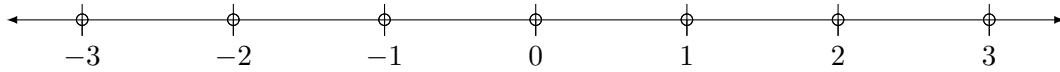
Your Answer

4. Consider an instance space  $\mathcal{X}$  which is the set of real numbers.

(a) (3 points) **Select one:** What is the VC dimension of hypothesis class  $H$ , where each hypothesis  $h$  in  $H$  is of the form “if  $a < x < b$  or  $c < x < d$  then  $y = 1$ ; otherwise  $y = 0$ ”? (i.e.,  $H$  is an infinite hypothesis class where  $a, b, c$ , and  $d$  are arbitrary real numbers).

- ☐ 2  
☐ 3  
☐ 4  
☐ 5  
☐ 6

(b) (3 points) Given the set of points in  $\mathcal{X}$  below, construct a labeling of some subset of the points to show that any dimension larger than the VC dimension of  $H$  by *exactly* 1 is incorrect (e.g. if the VC dimension of  $H$  is 3, only fill in the answers for 4 of the points). Fill in the boxes such that for each point in your example, the corresponding label is either 0 or 1. For points you are not using in your example, write N/A (do *not* leave the answer box blank).



Answer for -3	Answer for -2	Answer for -1	
Answer for 0	Answer for 1	Answer for 2	Answer for 3

### 3 MLE/MAP (32 points)

1. (1 point) **True or False:** Suppose you place a Beta prior over the Bernoulli distribution, and attempt to learn the parameter  $\theta$  of the Bernoulli distribution from data. Further suppose an adversary chooses “bad” but finite hyperparameters for your Beta prior in order to confuse your learning algorithm. As the number of training examples grows to infinity, the MAP estimate of  $\theta$  can still converge to the MLE estimate of  $\theta$ .

☐ True

☐ False

2. (2 points) **Select one:** Let  $\Gamma$  be a random variable with the following probability density function (pdf):

$$f(\gamma) = \begin{cases} 2\gamma & \text{if } 0 \leq \gamma \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose another random variable  $Y$ , which is conditioning on  $\Gamma$ , follows an exponential distribution with  $\lambda = 3\gamma$ . Recall that the exponential distribution with parameter  $\lambda$  has the following pdf:

$$f_{exp}(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the MAP estimate of  $\gamma$  given  $Y = \frac{2}{3}$  is observed?

Your Answer

3. (4 points) Neural the Narwhal found a mystery coin and wants to know the probability of landing on heads by flipping this coin. He models the coin toss as sampling a value from  $\text{Bernoulli}(\theta)$  where  $\theta$  is the probability of heads. He flips the coin three times and the flips turned out to be heads, tails, and heads. An oracle tells him that  $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$ , and *no other values of  $\theta$  should be considered*.

Find the MLE and MAP estimates of  $\theta$ . Use the following prior distribution for the MAP estimate:

$$p(\theta) = \begin{cases} 0.9 & \text{if } \theta = 0 \\ 0.04 & \text{if } \theta = 0.25 \\ 0.03 & \text{if } \theta = 0.5 \\ 0.02 & \text{if } \theta = 0.75 \\ 0.01 & \text{if } \theta = 1 \end{cases}.$$

Again, remember that  $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$ , so the MLE and MAP should also be one of them.

MLE of $\theta$	MAP of $\theta$

4. In a previous homework assignment, you have derived the closed form solution for linear regression. Now, we are coming back to linear regression, viewing it as a statistical model, and deriving the MLE and MAP estimate of the parameters in the following questions.

As a reminder, in MLE, we have

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} p(D|\theta)$$

For MAP, we have

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} p(\theta|D)$$

Assume we have data  $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$ , where  $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_M^{(i)})$ . So our data has  $N$  instances and each instance has  $M$  features. Each  $y^{(i)}$  is generated given  $\mathbf{x}^{(i)}$  with additive noise  $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$ : that is,  $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$  where  $\mathbf{w}$  is the parameter vector of linear regression.

- (a) (2 points) **Select one:** Given this assumption, what is the distribution of  $y^{(i)}$ ?

- ☐  $y^{(i)} \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}^{(i)}, \sigma^2)$
- ☐  $y^{(i)} \sim \mathcal{N}(0, \sigma^2)$
- ☐  $y^{(i)} \sim \text{Uniform}(\mathbf{w}^T \mathbf{x}^{(i)} - \sigma, \mathbf{w}^T \mathbf{x}^{(i)} + \sigma)$
- ☐ None of the above

- (b) (2 points) **Select one:** The next step is to learn the MLE of the parameters of the linear regression model. Which expression below is the correct conditional log likelihood  $\ell(\mathbf{w})$  with the given data?

- ☐  $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐  $\sum_{i=1}^N [\log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐  $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐  $-\log(\sqrt{2\pi\sigma^2}) + \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

- (c) (3 points) **Select all that apply:** Then, the MLE of the parameters is just  $\operatorname{argmax}_{\mathbf{w}} \ell(\mathbf{w})$ . Among the following expressions, select ALL that can yield the correct MLE.

- ☐  $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐  $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐  $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐  $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐  $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐ None of the above



5. Now we are moving on to learn the MAP estimate of the parameters of the linear regression model. Consider the same data  $D$  we used for the previous problem.

(a) (3 points) **Select all that apply:** Which expression below is the correct optimization problem the MAP estimate is trying to solving? Recall that  $D$  refers to the data, and  $\mathbf{w}$  to the regression parameters (weights).

- ☐  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(D, \mathbf{w})$
- ☐  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$
- ☐  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} \frac{p(D, \mathbf{w})}{p(\mathbf{w})}$
- ☐  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$
- ☐  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(\mathbf{w}|D)$
- ☐ None of the above

(b) (3 points) **Select one:** Suppose we are using a Gaussian prior distribution with mean 0 and variance  $\frac{1}{\lambda}$  for each element  $w_m$  of the parameter vector  $\mathbf{w}$ , i.e.  $w_m \sim \mathcal{N}(0, \frac{1}{\lambda})$  ( $1 \leq m \leq M$ ). Assume that  $w_1, \dots, w_M$  are mutually independent of each other. Which expression below is the correct log joint-probability of the data and parameters  $\log p(D, \mathbf{w})$ ? Please show your work below.

- ☐  $\sum_{i=1}^N \left( -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \right) - \sum_{m=1}^M \log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- ☐  $\sum_{i=1}^N \left( -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}) \right) + \sum_{m=1}^M -\log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- ☐  $\sum_{i=1}^N \left( -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}) \right) - \sum_{m=1}^M \log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$
- ☐  $\sum_{i=1}^N \left( -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \right) + \sum_{m=1}^M -\log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$

Work

(c) (2 points) **Select one:** For the same linear regression model with a Gaussian prior on the parameters as in the previous question, maximizing the log posterior probability  $\ell_{MAP}(\mathbf{w})$  gives you the MAP estimate of the parameters. Which of the following is an equivalent definition of  $\max_{\mathbf{w}} \ell_{MAP}(\mathbf{w})$ ?

- ☐  $\max_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$
- ☐  $\min_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$
- ☐  $\max_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \lambda \|\mathbf{w}\|_2^2$
- ☐  $\min_{\mathbf{w}} - \sum_{i=1}^N \frac{1}{2\sigma^2} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 - \frac{\lambda}{2} \|\mathbf{w}\|_2^2$

(d) (2 points) **Select one:** You found a MAP estimator that has a much higher test error than train error using some Gaussian prior. Which of the following could be a possible approach to fixing this?

- ☐ Increase the variance of the prior used
- ☐ Decrease the variance of the prior used

6. (4 points) **Select one:** Suppose now the additive noise  $\epsilon$  is different per datapoint. That is, each  $y^{(i)}$  is generated given  $\mathbf{x}^{(i)}$  with additive noise  $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma_i^2)$ , i.e.  $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$ . Unlike the standard regression model we have worked with until now, there is now an example specific variance  $\sigma_i^2$ . Maximizing the log-likelihood of this new model is equivalent to minimizing the *weighted* mean squared error with which of the following as the weights? Please show your work below.

- ☐  $1/y^{(i)}$
- ☐  $1/\sigma_i^2$
- ☐  $1/\|\mathbf{x}^{(i)}\|_2^2$

Work

7. (4 points) **Select one:** MAP estimation with what prior is equivalent to  $\ell_1$  regularization? Please show your work below.

Note:

- The pdf of a uniform distribution over  $[a, b]$  is  $f(x) = \frac{1}{b-a}$  if  $x \in [a, b]$  and 0 otherwise.
  - The pdf of an exponential distribution with rate parameter  $a$  is  $f(x) = a \exp(-ax)$  for  $x > 0$ .
  - The pdf of a Laplace distribution with location parameter  $a$  and scale parameter  $b$  is  $f(x) = \frac{1}{2b} \exp\left(-\frac{|x-a|}{b}\right)$  for all  $x \in \mathbb{R}$ .
- ☐ Uniform distribution over  $[-1, 1]$
- ☐ Uniform distribution over  $[-\mathbf{w}^T \mathbf{x}^{(i)}, \mathbf{w}^T \mathbf{x}^{(i)}]$
- ☐ Exponential distribution with rate parameter  $a = \frac{1}{2}$
- ☐ Exponential distribution with rate parameter  $a = \mathbf{w}^T \mathbf{x}^{(i)}$
- ☐ Laplace distribution with location parameter  $a = 0$
- ☐ Laplace distribution with location parameter  $a = \mathbf{w}^T \mathbf{x}^{(i)}$

Work

## 4 Naïve Bayes (31 points)

1. The following dataset describes several features of a narwhal and then whether or not it has an Instagram account.

Color	Size	Has Instagram?
Rainbow	Small	N
Cyan	Small	N
Cyan	Small	Y
Cyan	Medium	Y
Rainbow	Medium	N
Fuchsia	Medium	Y
Fuchsia	Large	Y
Cyan	Large	Y

Neural the Narwhal is cyan and medium-sized. We would like to determine whether he has an Instagram account, using the Naïve Bayes assumption to estimate the following probabilities.

- (a) (1 point) Under the Naïve Bayes assumption, how many parameters do we need to estimate?

Your Answer

- (b) (1 point) If we don't use the Naïve Bayes assumption (i.e. we cannot factorize  $P(X|Y)$ ), how many parameters do we need to estimate?

Your Answer

- (c) (3 points) What is the generative story for this data? Name the distributions for the features and labels. Let  $Y$  represent whether or not a narwhal has an Instagram account and  $X_1$  and  $X_2$  represent Color and Size, respectively.

$Y$	$X_1 Y$	$X_2 Y$

- (d) (1 point) Suppose we use Maximum Likelihood Estimation to train our model. What is our estimate of  $\theta_{X_1 = \text{Rainbow}, Y = N}$ ?

Your Answer

(e) (1 point) What is our MLE estimate of  $\theta_{X_1 = \text{Fuchsia}, Y = N}$ ?

Your Answer

(f) (1 point) What is the probability that a narwhal is cyan, medium-sized, and has an Instagram account? Round the answer to the fourth decimal place, e.g. 0.1234.

Your Answer

(g) (1 point) What is the probability that a narwhal is cyan, medium-sized, and does *not* have an Instagram account? Round the answer to the fourth decimal place, e.g. 0.1234.

Your Answer

(h) (1 point) **Select one:** Does Neural the Narwhal have an Instagram account?

- ☐ Yes
- ☐ No

(i) (1 point) Give a test data point for which a Naïve Bayes model trained via MLE will never predict the correct label. The example data point must use feature values that are already present in the data set (i.e do not create new values for colors, sizes, or has instagram in your example)

Your Answer

2. (3 points) **Select all that apply:** Gaussian Naïve Bayes in general can learn non-linear decision boundaries. Consider the simple case where we have just one real-valued feature  $X_1 \in \mathbb{R}$  from which we wish to infer the value of label  $Y \in \{0, 1\}$ . The corresponding generative story would be:

$$Y \sim \text{Bernoulli}(\phi)$$

$$X_1|Y \sim \text{Gaussian}(\mu_y, \sigma_y^2)$$

where the parameters are the Bernoulli parameter  $\phi$  and the class-conditional Gaussian parameters  $\mu_0, \sigma_0^2$  and  $\mu_1, \sigma_1^2$  corresponding to  $Y = 0$  and  $Y = 1$ , respectively.

A linear decision boundary in one dimension can be described by a rule of the form:

$$\text{if } X_1 > c \text{ then } Y = 1, \text{ else } Y = 0$$

where  $c$  is a real-valued threshold and  $k \in \{0, 1\}$ .

Is it possible in this simple one-dimensional case to construct a Gaussian Naïve Bayes classifier with a decision boundary that cannot be expressed by a rule in the above form? (Hint: Think about what a Gaussian distribution looks like for one random variable)

- ☐ Yes, this can occur if the Gaussians are of equal means and unequal variances.
  - ☐ Yes, this can occur if the Gaussians are of unequal means and equal variances.
  - ☐ Yes, this can occur if the Gaussians are of unequal means and unequal variances.
  - ☐ None of the above
3. (4 points) Gaussian Naïve Bayes has a linear objectives in some cases. Consider a case where we have features  $x_1, x_2$  and a binary label  $y$ . Below is the quadratic decision boundary derived in recitation Q5.

$$0 = C + \frac{(x_1 - \mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2 - \mu_{21})^2}{2\sigma_{21}^2} - \frac{(x_1 - \mu_{10})^2}{2\sigma_{10}^2} - \frac{(x_2 - \mu_{20})^2}{2\sigma_{20}^2}$$

- (a) (2 points) Clearly state the assumption/conditions under which Gaussian Naïve Bayes model learns a linear decision boundary

Your Answer

- (b) (2 points) Using the assumptions in part (a), re-derive the decision boundary as a linear combination of  $x_1, x_2$ . Any additive constants NOT depending on  $x_1, x_2$  can be folded into  $C$ . Please simplify the coefficients for  $x_1, x_2$ , but do NOT fold into a constant. You do not need to show work, just provide the final answer.

Your Answer

4. (2 points) **Select all that apply:** Select all possible decision boundaries that can be produced by a Gaussian Naïve Bayes classifier. The shaded region is assigned class 1 and the unshaded regions is assigned class 0. (*Hint: Recall the conclusion of the proof given in recitation.*)

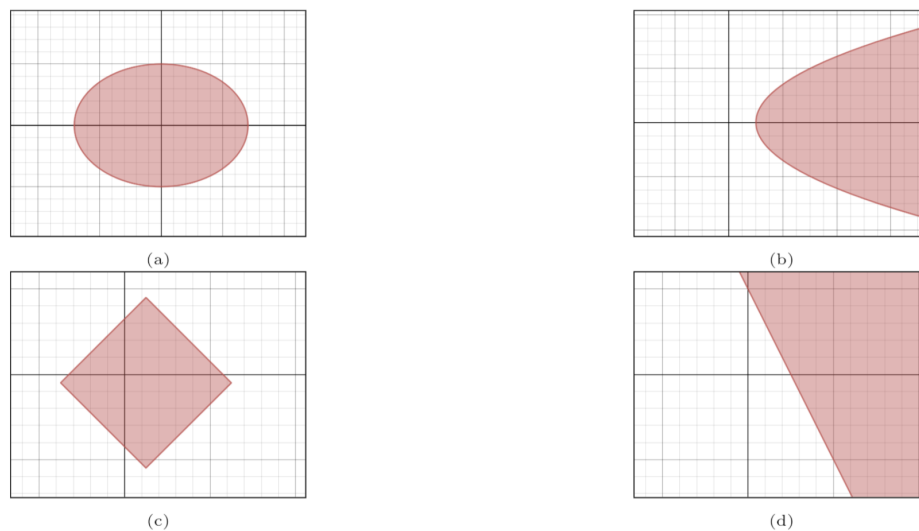


Figure 1: Decision Boundaries

- ☐ (a)
- ☐ (b)
- ☐ (c)
- ☐ (d)
- ☐ None of the above

5. Suppose we want to extend the Naïve Bayes model to time series data. Formally, a training data point consists of a binary label  $Y$  and  $D$  sequentially ordered observations of a binary outcome where  $X_1$  occurs before  $X_2$ , which occurs before  $X_3$  and so on all the way down to the final observation  $X_D$ ; each feature  $X_d$  is binary.

You decide to modify the Naïve Bayes assumption such that a feature  $X_d$  is conditionally independent of all other features given the label  $Y$  and the previous feature  $X_{d-1}$ ; the first feature  $X_1$  is conditionally independent of all other features given just the label  $Y$ . The corresponding Generative story would be:

$$\begin{aligned} Y &\sim \text{Bernoulli}(\phi) \\ X_1|Y &\sim \text{Bernoulli}(\theta_{1,y}) \\ X_d|X_{d-1}, Y &\sim \text{Bernoulli}(\theta_{d,x,y}) \end{aligned}$$

where the parameters are the Bernoulli parameter  $\phi$  and, the class-conditional Bernoulli parameter  $\theta_1$ , and the class-conditional Bernoulli parameters  $\theta_{d,x,y}$  for  $X_d|X_{d-1} = x, Y = y$ .

- (a) (2 points) Write down the expression for the joint distribution  $P(X, Y)$  under your new Naïve Bayes model. You don't need to plug in the values of the pdf, you can leave it as probability expressions. You must use the modified Naïve Bayes assumption described above

Your Answer

- (b) (1 point) How many parameters do you need to learn in order to make predictions using this new Naïve Bayes model? Write your answer in terms of  $D$ .

Your Answer

- (c) (2 points) Suppose we train this model via MLE. In at most 2 sentences, describe how we would estimate  $\phi$

Your Answer

- (d) (2 points) In at most 2 sentences, describe how we would estimate  $\theta_{d,x,y}$  for  $2 \leq d \leq D$  using MLE.



Your Answer

## 5 Generative vs. Discriminative Models (15 points)

1. (2 points) **Select all that apply:** Which of the following models are discriminative?

- ☐ Logistic Regression with L2 regularization
- ☐ Logistic Regression without L2 regularization
- ☐ Naive Bayes
- ☐ KNN
- ☐ Decision Trees
- ☐ None of the above

2. Consider the following dataset  $D$  for 1D Logistic Regression and Gaussian Naive Bayes. There are nine points with label  $y = 1$  at  $x = 1$ , one point with label  $y = 1$  at  $x = 2$ , nine points with label  $y = 0$  at  $x = -1$ , one point with label  $y = 0$  at  $x = -2$ . In addition, we have another dataset  $D'$ , which is an exact duplicate of  $D$ , except in addition, we have two outliers, one point with label  $y = 0$  at  $x = 10$ , and one point with label  $y = 1$  at  $x = -2$ . For the following questions, **round your answer to 3 decimal places**.

(a) (1 point) What is the decision rule for Logistic Regression on  $D$ ? In other words, find the  $C$  such that Logistic Regression will predict  $y = 1$  for  $x \geq C$ .

Your Answer

(b) (1 point) What is the decision rule for Naive Bayes on  $D$ ?

Your Answer

(c) (1 point) We ran logistic regression on  $D'$ , and learned that the bias term is  $\beta_0 = -0.0515$ , and the weight is  $\beta_1 = 0.1859$ . What is the decision rule for this dataset (with the outliers)?

Your Answer

(d) (5 points) What are the parameters for our Gaussian Naive Bayes classifier on  $D'$ , if we train via MLE?

(a)  $P(Y = 1)$

(b) mean for  $P(X|Y = 1)$

(c) variance for  $P(X|Y = 1)$

(d) mean for  $P(X|Y = 0)$

(e) variance for  $P(X|Y = 0)$

(a)	(b)	(c)	(d)
(e)			

(f) (2 points) What is the new decision rule for Naive Bayes on  $D'$ ? Note you might get two numbers, since the decision rule is a bounded interval. Please find  $C_1$  and  $C_2$  such that Naive Bayes will predict  $y = 1$  for  $C_1 \leq x \leq C_2$ .

Your Answer

(g) (3 points) Based on your answers to the previous questions, which one of Logistic Regression and Naive Bayes is more resistant to outliers? Compare the decision boundaries on  $D$  and  $D'$  for Logistic Regression and Naive Bayes. Explain your reasoning in 2-3 sentences.

Your Answer

## 6 Google Colab (2 points)

Please refer to the notebook [here](#) as well as the google colab setup [video](#) for the following questions. **Please upload a screenshot of the commands being run and the outputs**, so that we can verify that you have successfully setup google colab for the next homework. When running the commands below, make sure that you are connected to a GPU.

1. (1 point) Upload a screenshot of the output of mounting your drive onto your notebook. On the left hand side, please show the folders in your drive (It is ok to just show the MyDrive folder, we just need to see it). In your screenshot, also include that you are connected to a GPU, which you will see on the right hand side of the page.

Your Answer



2. (1 point) Upload a screenshot of running the following commands: `torch.cuda.is_available()`, `torch.cuda.device_count()`, and `torch.cuda.get_device_name(0)`.

Your Answer



## 7 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found [here](#).

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer