Midterm review

Midterm Exam

- **Time / Location**
  - **Time:** Evening Exam
    Thu, March 21st at 6:30pm-8:30pm
  - **Room**: We will contact each student individually with **your room assignment**.
  - **Seats:** There will be **assigned seats**. Please arrive early.
  - Please watch Piazza carefully for announcements regarding room / seat assignments.
- **Logistics**
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Derivations
    - Short answers
    - Interpreting figures
    - Implementing algorithms on paper
  - No electronic devices
  - You are allowed to **bring** one 8½ x 11 sheet of notes (front and back)

Lecture 1

- MLE and MAP for coin flips: estimate the probability Θ of a coin landing on heads
  - MLE:
    - Notion of likelihood function
    - Setting up the objective function
    - Verifying the function is concave
    - Maximizing the objective function

Lecture 1

- MAP:
  - Notion of prior and posterior functions, and their relationship with the likelihood through Bayes rule.
  - Using the Beta prior
  - Setting up the objective function
  - Verifying the function is concave
  - Maximizing the objective function
  - How does the prior affect the final outcome? How can it be interpreted in terms of additional throws?

Lecture 2

- Familiarity with the notion of conjugate priors:
  - In the coin flip MAP example, the beta distribution is the *conjugate prior* of the Bernoulli likelihood function. We also call the prior and the posterior *conjugate priors*
- Probability review:
  - Probability distribution functions (PDF) for continuous variables
  - Expectation, Variance
  - Joint distributions, chain rule, Bayes rule.

Lecture 3

- Decision rules
  - Bayes decision rule
    - Bayes error/unavoidable error
- Definition of generative and discriminative classifiers

- KNN:
  - How to use KNN?
  - Training error vs test error.
  - How to pick K by cross-validation (what is cross-validation)
  - Behavior of the algorithm when K is small and when K is large, in terms of model complexity and risk of overfitting.
  - Behavior of the algorithm when K=1
  - What does overfitting mean?

Lecture 4

- Linear regression
  - Mean squared error statement and derivation
  - Probabilistic statement and derivation, and equivalence of the two solutions (from homework)
- Ridge regression
  - Mean squared error and ridge penalty statement and derivation
  - Probabilistic statement and derivation, and equivalence of the two solutions (from homework)
- Lasso regression:
  - What does it enforce, how do the learned parameters change with increased penalty?

Lecture 4

- Bias variance tradeoff:
  - The expected loss when the model is correctly specified can be decomposed into the sum of three components: bias^2, variance and noise (no need to memorize the derivation from class)
  - You need to know how to compute the bias of an estimator, and what it corresponds to
  - You need to know how to compute the variance of an estimator, and what it corresponds to

Lecture 5 - Naive Bayes

- Conditional independence assumption of naive bayes. How does it it help with learning less parameters?
- Naive Bayes Algorithm with binary X and Y:
  - How is it specified?
  - How to train it?
  - How many parameters does it require for the user to learn?
  - How to use with for prediction?
  - What is chance performance?
  - What happens when the Xs are not conditionally independent? What happens when some of the variables are irrelevant?

Lecture 5 - Naive Bayes

- Naive bayes for text classification:
  - What is the bag of word model?
    - How to formulate it?
    - How to learn it?
  - How to deal with words what we never encounter? What kind of prior can we use?
- Gaussian naive bayes (GNB):
  - What are the assumptions behind GNB?
  - How does it account for continuous variables?

Lecture 6 - Logistic Regression

- Logistic regression:
  - Logistic regression outputs the probability of Y given X.
  - How is the problem setup?
  - It doesn't have an analytical solution so gradient descent/gradient descent has to be used to learn the weights, depending on the formulation of the problem
  - What is block gradient descent? What is stochastic gradient descent? What are their characteristics?
  - What is the learning rate? What are the drawbacks/benefits of a large learning rate or a small learning rate?

Lecture 6 - Logistic Regression

- Logistic regression:
  - Without regularization, the logistic regression problem is ill specified and the magnitude of the weights will tend to infinity as the probability of the training labels is maximized. Regularization offers a tradeoff between weight size and training error
  - How many parameters need to be estimated?
- Logistic regression (LR) vs GNB:
  - If the variances learned for each X_i are made to be equal across classes, GNB learns a linear decision boundary.
  - GNB converges faster to its asymptotic error
  - If the variances are equal across classes and if the X_is are conditionally independent given Y then GNB and LR perform similarly.
  - GNB might perform worse than LR if the data is not conditionally independent but it is not always easy to predict

Lecture 7 - Decision trees

- How to use a decision tree to make predictions
- Entropy, conditional entropy and information gain statements.
- How to use information gain to greedily build a tree (using ID3). This approach optimizes the length of the tree to obtain short trees.
- How to interpret train error / test error plots. What does overfitting correspond to mathematically/graphically?

Lecture 8 - Perceptron and NN

- The perceptron algorithm:
  - Statement, how to train it, how to use it to predict
  - The effect of a misclassified positive sample in training: changes the training prediction in the positive direction.
  - The effect of a misclassified negative sample in training: changes the training prediction in the negative direction.
  - The notion of margin and of a linearly separable training set.

Lecture 9 - Neural Networks

- Neural networks:
  - Stacking layers with non-linear activations allows us to learn complex decision boundaries
  - Sigmoid functions are used instead of sign functions because they are differentiable. They however introduce problems such as the vanishing gradient problem.
  - The steps required to train a neural network: (We saw this in detail in HW and in class)
    - Including how to use back-propagation to estimate the gradient at every parameter of the network.

Lecture 9 - Neural Networks

- Different tricks can be used to optimize neural networks, leading to a large array of decision involving the number of layers, the number of nodes, the choice of different activation gates or of different loss functions.
- Neural networks can be used in order to learn transformations of input data such as images and sounds into a space in which they are characterized by features that are important for the task at hand. This is referred to as representation learning.

Lecture 10/11 - SVMs

- Separable training set - linear SVMs:
  - Problem statement
  - Writing down the Lagrangian then the dual formulation and the solution of the dual problem.
  - How to use a trained SVM to make predictions. What decision boundaries are learned?
  - The dual formulation lets us see that the algorithm relies on the dot product of new test points with the support vectors.
- How to show that k(x,y) is a kernel?
  - Mercer theorem or expressing k(x,y) as a dot product of the same feature map of x and y.

Lecture 10/11 - SVMs

- Kernel SVMs:
  - We can use the kernel trick to allow us to efficiently train and use SVMs with non-linear kernels, allowing us to learn complex decision boundaries
  - Common kernels and resulting decision boundaries (from homework)
- Non-separable training set:
  - We can use SVMs with slack variables to learn decision boundaries (homework)

Lecture 12 - Boosting

- Boosting uses weak learners with accuracy above 0.5 percent to derive a strong learner
- Adaboost algorithm:
  - How does it works for prediction / what does it output?
  - How is it trained?
  - It ends up performing surprisingly well in some scenarios. It could suffer from overfitting sometimes, and sometimes the test error keeps improving even when the training error is zero.

Lecture 13/14 Learning Theory

- Notion of the complexity of the hypothesis space H.
- Theory to relate the number of training examples, the complexity of hypothesis space, training error and true error (as well as how training examples are presented).
- Mathematical formulation of overfitting
- The notion of ε-exhausting a version-space
- The notion of VC dimensions and how to derive it with simple classifiers in the case where H is not finite

Lecture 13/14 Learning Theory

With probability $\geq (1-\delta)$, $(error_{true} - error_{train}) \leq \epsilon$

(1) for all $h \in H$ such that $error_{train} = 0$,

$$\epsilon = \frac{\ln |H| + \ln(1/\delta)}{m}$$

finite H

(2) for all $h \in H$

Agnostic

$$\epsilon = \sqrt{\frac{\ln |H| + \ln(1/\delta)}{2m}}$$

finite H

(3) for all $h \in H$

Agnostic

$$\epsilon = 8\sqrt{\frac{VC(H)(\ln \frac{m}{VC(H)} + 1) + \ln(8/\delta)}{2m}}$$

infinite H

Lecture 13/14 Learning Theory

- In the case where we have consistent classifiers, the number of examples we need grow as a function of $\frac{1}{\epsilon}$
- In the case of agnostic learning (training error is not 0) the number of examples we need grow as a function of $\frac{1}{\epsilon^2}$

Sample Questions

Now we will apply K-Nearest Neighbors using Euclidean distance to a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. A point can be its own neighbor.

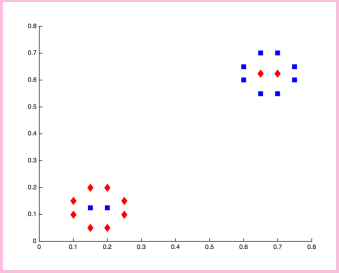


Figure 5

3. **[2 pts]** What value of k minimizes leave-one-out cross-validation error for the dataset shown in Figure 5? What is the resulting error?

23

Sample Questions

3.2 Logistic regression

Given a training set $\{(x_i, y_i), i = 1, \ldots, n\}$ 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

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

(b) [5 pts.] What is the form of the classifier output by logistic regression?

(c) [2 pts.] **Extra Credit:** 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 $\hat{w}$?

24

Sample Questions

2.1 Train and test errors

In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained till convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate test set $\mathcal{D}^{\text{test}}$. You look at the test error, and find that it is very high. You then compute the training error and find that it is close to 0.

1. **[4 pts]** Which of the following is expected to help? Select all that apply.
   (a) Increase the training data size.
   (b) Decrease the training data size.
   (c) Increase model complexity (For example, if your classifier is an SVM, use a more complex kernel. Or if it is a decision tree, increase the depth).
   (d) Decrease model complexity.
   (e) Train on a combination of $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ and test on $\mathcal{D}^{\text{test}}$
   (f) Conclude that Machine Learning does not work.

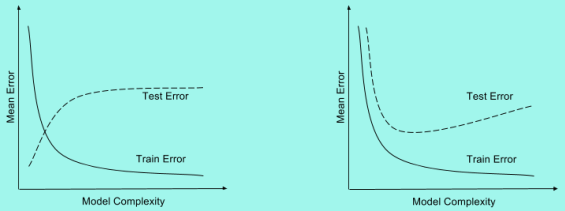
25

Sample Questions

2.1 Train and test errors

In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained till convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate test set $\mathcal{D}^{\text{test}}$. You look at the test error, and find that it is very high. You then compute the training error and find that it is close to 0.

4. **[1 pts]** Say you plot the train and test errors as a function of the model complexity. Which of the following two plots is your plot expected to look like?



26

Sample Questions

4.1 True or False

Answer each of the following questions with **T** or **F** and **provide a one line justification**.

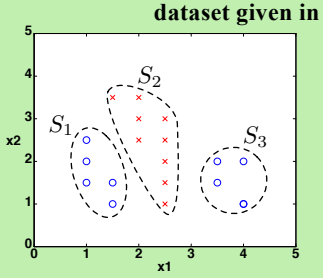
(a) [2 pts.] Consider two datasets $D^{(1)}$ and $D^{(2)}$ where $D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), ..., (x_n^{(1)}, y_n^{(1)})\}$ and $D^{(2)} = \{(x_1^{(2)}, y_1^{(2)}), ..., (x_m^{(2)}, y_m^{(2)})\}$ such that $x_i^{(1)} \in \mathbb{R}^{d_1}$, $x_i^{(2)} \in \mathbb{R}^{d_2}$. Suppose $d_1 > d_2$ and $n > m$. Then the maximum number of mistakes a perceptron algorithm will make is higher on dataset $D^{(1)}$ than on dataset $D^{(2)}$.

27

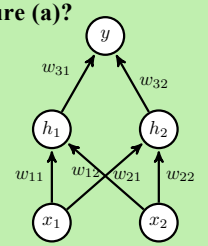
Sample Questions

Neural Networks

Can the neural network in Figure (b) correctly classify the dataset given in Figure (a)?



(a) The dataset with groups S_1, S_2, and S_3.



(b) The neural network architecture

28

Samples Questions

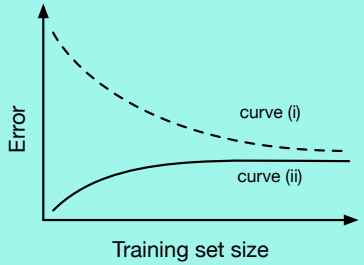
2.1 True Errors

(b) [4 pts.] **T or F:** Learning theory allows us to determine with 100% certainty the true error of a hypothesis to within any $\epsilon > 0$ error.

29

Samples Questions

2.2 Training Sample Size



(a) [8 pts.] Which curve represents the training error? **Please provide 1–2 sentences of justification**.

(b) [4 pt.] In one word, what does the gap between the two curves represent?

30

Sample Questions

5 Learning Theory [20 pts.]

(a) [3 pts.] **T or F**: It is possible to label 4 points in $\mathbb{R}^2$ in all possible 2^4 ways via linear separators in $\mathbb{R}^2$.

(d) [3 pts.] **T or F**: The VC dimension of a concept class with infinite size is also infinite.

(f) [3 pts.] **T or F**: Given a realizable concept class and a set of training instances, a consistent learner will output a concept that achieves 0 error on the training instances.

31

Sample Questions

1.2 Maximum Likelihood Estimation (MLE)

Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. We are going to derive the MLE for θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \theta) = \theta^X (1 - \theta)^{1-X}.$$

(a) [2 pts.] Derive the likelihood, $L(\theta; X_1, \ldots, X_n)$.

(c) **Extra Credit:** [2 pts.] Derive the following formula for the MLE: $\hat{\theta} = \frac{1}{n}(\sum_{i=1}^{n} X_i)$.

32

Sample Questions

1.3 MAP vs MLE

Answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** In the limit, as n (the number of samples) increases, the MAP and MLE estimates become the same.

Sample Questions

1.1 Naive Bayes

You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a 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 **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

(c) [2 pts.] **T or F:** $P(\texttt{height}|\texttt{sex},\texttt{hair}) = P(\texttt{height}|\texttt{sex})$.

Sample Questions

4.3 Analysis

(a) [4 pts.] In one or two sentences, describe the benefit of using the Kernel trick.

(b) [4 pt.] The concept of margin is essential in both SVM and Perceptron. Describe why a large margin separator is desirable for classification.

Sample Questions

(c) [4 pts.] **Extra Credit:** Consider the dataset in Fig. 4. Under the SVM formulation in section 4.2(a),

(1) Draw the decision boundary on the graph.
(2) What is the size of the margin?
(3) Circle all the support vectors on the graph.

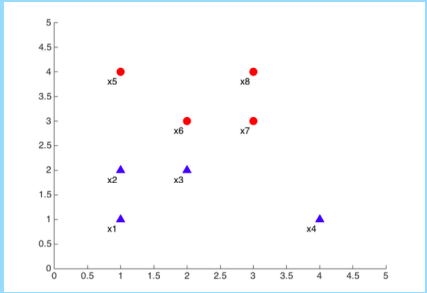


Figure 4: SVM toy dataset

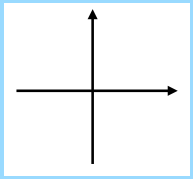
Sample Questions

3. **[Extra Credit: 3 pts.]** One formulation of soft-margin SVM optimization problem is:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^\top x_i) \geq 1 - \xi_i \quad \forall i = 1, ..., N$$
$$\xi_i \geq 0 \quad \forall i = 1, ..., N$$
$$C \geq 0$$

where (x_i, y_i) are training samples and $\mathbf{w}$ defines a linear decision boundary.

Derive a formula for ξ_i when the objective function achieves its minimum (No steps necessary). Note it is a function of $y_i\mathbf{w}^\top x_i$. Sketch a plot of ξ_i with $y_i\mathbf{w}^\top x_i$ on the x-axis and value of ξ_i on the y-axis. What is the name of this function?



Sample Questions

1 Topics before Midterm

(a) [2 pts.] **T or F**: Naive Bayes can only be used with MLE estimates, and not MAP estimates.

(b) [2 pts.] **T or F**: Logistic regression cannot be trained with gradient descent algorithm.

(d) [2 pts.] **T or F**: Leaving out one training data point will always change the decision boundary obtained by perceptron.

38

Sample Questions

1 Topics before Midterm

(e) [2 pts.] **T or F**: The function $K(\mathbf{x}, \mathbf{z}) = -2\mathbf{x}^T\mathbf{z}$ is a valid kernel function.

8. **[2 pts]** With an infinite supply of training data, the trained Naïve Bayes classifier is an optimal classifier.

Circle one: True False
One line justification (only if False):

39