Midterm Exam Review
Reminders

• Midterm Exam (Evening Exam)
  – Tue, Mar. 07 at 7:00pm – 9:30pm
  – See Piazza for details about location
Outline

• Midterm Exam Logistics
• Sample Questions
• Classification and Regression: The Big Picture
• Q&A
MIDTERM EXAM LOGISTICS
Midterm Exam

• Logistics
  – Evening Exam
    Tue, Mar. 07 at 7:00pm – 9:30pm
  – 8-9 Sections
  – Format of questions:
    • Multiple choice
    • True / False (with justification)
    • Derivations
    • Short answers
    • Interpreting figures
  – No electronic devices
  – You are allowed to bring one 8½ x 11 sheet of notes (front and back)
Midterm Exam

• How to Prepare
  – Attend the midterm review session: Thu, March 2 at 6:30pm (PH 100)
  – Attend the midterm review lecture Mon, March 6 (in-class)
  – Review prior year’s exam and solutions (we’ll post them)
  – Review this year’s homework problems
Midterm Exam

• **Advice (for during the exam)**
  – Solve the easy problems first (e.g. multiple choice before derivations)
    • if a problem seems extremely complicated you’re likely missing something
  – Don’t leave any answer blank!
  – If you make an assumption, write it down
  – If you look at a question and don’t know the answer:
    • we probably haven’t told you the answer
    • but we’ve told you enough to work it out
    • imagine arguing for some answer and see if you like it
Topics for Midterm

• Foundations
  – Probability
  – MLE, MAP
  – Optimization

• Classifiers
  – KNN
  – Naïve Bayes
  – Logistic Regression
  – Perceptron
  – SVM

• Regression
  – Linear Regression

• Important Concepts
  – Kernels
  – Regularization and Overfitting
  – Experimental Design
SAMPLE QUESTIONS
Sample Questions

1.4 Probability
Assume we have a sample space \( \Omega \). Answer each question with \( T \) or \( F \).

(a) [1 pts.] \( T \) or \( F \): If events \( A \), \( B \), and \( C \) are disjoint then they are independent.

\[ P(A \cup B) \leq P(A) + P(B) \]

(b) [1 pts.] \( T \) or \( F \): \( P(A|B) \propto \frac{P(A)P(B|A)}{P(A|B)} \). (The sign ‘\( \propto \)’ means ‘is proportional to’).
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.

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?
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 $\theta$. 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, $L(\theta; X_1, \ldots, X_n)$.

(b) [2 pts.] Derive the following formula for the log likelihood:

$$`L(\theta; X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i \log(\theta) + \sum_{i=1}^{n} X_i \log(1 - \theta).$$

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


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 ∈ {male, female}
- height ∈ [0,300] centimeters
- hair ∈ {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.

(b) [2 pts.] **T or F:** Since there is not a similar number of men and women in the dataset, Naive Bayes will have high test error.

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

(d) [2 pts.] **T or F:**
Sample Questions

3.1 Linear regression

Consider the dataset $S$ plotted in Fig. 1 along with its associated regression line. For each of the altered data sets $S_{\text{new}}$ plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An observed data set and its associated regression line.

Figure 2: New regression lines for altered data sets $S_{\text{new}}$.

Dataset

(a) Adding one outlier to the original data set.
Sample Questions

3.1 Linear regression

Consider the dataset $S$ plotted in Fig. 1 along with its associated regression line. For each of the altered data sets $S^{\text{new}}$ plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An observed data set and its associated regression line.

Figure 2: New regression lines for altered data sets $S^{\text{new}}$.

(c) Adding three outliers to the original data set. Two on one side and one on the other side.

Dataset

Figure 3: New data set $S^{\text{new}}$. Two on one side and one on the other side.
Sample Questions

### 3.1 Linear regression

Consider the dataset $S$ plotted in Fig. 1 along with its associated regression line. For each of the altered data sets $S_{\text{new}}$ plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An observed data set and its associated regression line.

Figure 2: New regression lines for altered data sets $S_{\text{new}}$.

(d) Duplicating the original data set.

(a) Old and new regression lines.
(b) Old and new regression lines.
(c) Old and new regression lines.
Sample Questions

3.1 Linear regression

Consider the dataset $S$ plotted in Fig. 1 along with its associated regression line. For each of the altered data sets $S^{\text{new}}$ plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression line</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: An observed data set and its associated regression line.

Figure 2: New regression lines for altered data sets $S^{\text{new}}$.

Figure 3: New data set and adding four points that lie on the trajectory of the original regression line.

(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.
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 \( \hat{w} \)
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 $D_{\text{train}}$, and tested on a separate test set $D_{\text{test}}$. 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 $D_{\text{train}}$ and $D_{\text{test}}$ and test on $D_{\text{test}}$
   
   (f) Conclude that Machine Learning does not work.

2. **[5 pts]** Explain your choices.

3. **[2 pts]** What is this scenario called?

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?
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 $D_{\text{train}}$, and tested on a separate test set $D_{\text{test}}$. 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?

![Graphs showing train and test errors as a function of model complexity.](image)
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 \( D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), \ldots, (x_n^{(1)}, y_n^{(1)})\} \) and \( D^{(2)} = \{(x_1^{(2)}, y_1^{(2)}), \ldots, (x_m^{(2)}, y_m^{(2)})\} \) 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 \( D^{(2)} \).
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.
(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:

\[
\min_w \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t. } y_i(w^T x_i) \geq 1 - \xi_i \quad \forall i = 1, \ldots, N \\
\xi_i \geq 0 \quad \forall i = 1, \ldots, N \\
C \geq 0
\]

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

Derive a formula for \(\xi_i\) when the objective function achieves its minimum (No steps necessary). Note it is a function of \(y_i w^T x_i\). Sketch a plot of \(\xi_i\) with \(y_i w^T x_i\) on the x-axis and value of \(\xi_i\) on the y-axis. What is the name of this function?
The Big Picture

CLASSIFICATION AND REGRESSION
Classification and Regression: The Big Picture

Whiteboard

– Decision Rules / Models (probabilistic generative, probabilistic discriminative, perceptron, SVM, regression)

– Objective Functions (likelihood, conditional likelihood, hinge loss, mean squared error)

– Regularization (L1, L2, priors for MAP)

– Update Rules (SGD, perceptron)

– Nonlinear Features (preprocessing, kernel trick)
Q&A