Final Exam Review
Reminders

• Homework 9: Learning Paradigms
  – Out: Fri, Nov 30
  – Due: Fri, Dec 7 at 11:59pm
Outline

1. Exam Logistics
2. Sample Questions
3. Overview
Final Exam

• **Time / Location**
  – **Time:** Evening Exam
    **Thu, Dec 13 at 1:00pm – 4:00pm**
  – **Room:** We will contact each student individually with your room assignment. The rooms are not based on section.
  – **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)
Final Exam

• How to Prepare
  – Attend (or watch) this final exam review session
  – Review prior year’s exams and solutions
    • We already posted these for the midterm
    • Disclaimer: This year’s 10-601 is not the same as prior offerings, so review both midterm and final
  – Review this year’s homework problems
  – Consider whether you have achieved the “learning objectives” for each lecture / section
  – Attend the Final Exam Recitation (Friday)
Final 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
Final Exam

• Exam Contents
  – ~20% of material comes from topics covered before the midterm exam
  – ~80% of material comes from topics covered after the midterm exam
Topics covered before Midterm

• Foundations
  – Probability, Linear Algebra, Geometry, Calculus
  – MLE
  – Optimization

• Important Concepts
  – Regularization and Overfitting
  – Experimental Design

• Classifiers
  – Decision Tree
  – KNN
  – Perceptron
  – Logistic Regression

• Regression
  – Linear Regression

• Feature Learning
  – Neural Networks
  – Basic NN Architectures
  – Backpropagation

• Learning Theory
  – PAC Learning
Topics covered after Midterm

• Learning Theory
  – PAC Learning

• Generative Models
  – Generative vs. Discriminative
  – MLE / MAP
  – Naïve Bayes

• Graphical Models
  – HMMs
  – Learning and Inference
  – Bayesian Networks

• Reinforcement Learning
  – Value Iteration
  – Policy Iteration
  – Q-Learning
  – Deep Q-Learning

• Unsupervised Learning
  – K-Means
  – PCA

• Other Learning Paradigms
  – SVM (large-margin)
  – Kernels
  – Ensemble Methods / AdaBoost
SAMPLE QUESTIONS

Material Covered Before Midterm Exam
Matching Game

**Goal:** Match the Algorithm to its Update Rule

1. SGD for Logistic Regression
   \[ h_\theta(x) = p(y|x) \]

2. Least Mean Squares
   \[ h_\theta(x) = \theta^T x \]

3. Perceptron
   \[ h_\theta(x) = \text{sign}(\theta^T x) \]

4. \[ \theta_k \leftarrow \theta_k + (h_\theta(x^{(i)}) - y^{(i)}) \]

5. \[ \theta_k \leftarrow \theta_k + \frac{1}{1 + \exp \lambda(h_\theta(x^{(i)}) - y^{(i)})} \]

6. \[ \theta_k \leftarrow \theta_k + \lambda(h_\theta(x^{(i)}) - y^{(i)}) x_k^{(i)} \]

A. 1=5, 2=4, 3=6
B. 1=5, 2=6, 3=4
C. 1=6, 2=4, 3=4
D. 1=5, 2=6, 3=6
E. 1=6, 2=6, 3=6
4. **K-NN [12 pts]**

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](image)

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?
# 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.](image1.png)

![Figure 2: New regression lines for altered data sets $S_{\text{new}}$.](image2.png)

(a) Adding one outlier to the original data set.
Topographical Maps
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 $S^{new}$ 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 $S^{new}$.

Dataset

(c) Adding three outliers to the original data set. Two on one side and one on the other side.
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.
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 $S^{\text{new}}$.

(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.
# Robotic Farming

<table>
<thead>
<tr>
<th></th>
<th>Deterministic</th>
<th>Probabilistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Is this a picture of a wheat kernel?</td>
<td>Is this plant drought resistant?</td>
</tr>
<tr>
<td>(binary output)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>How many wheat kernels are in this picture?</td>
<td>What will the yield of this plant be?</td>
</tr>
<tr>
<td>(continuous output)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multinomial Logistic Regression

polar bears
sea lions
sharks
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 \( \hat{w} \)?
Handcrafted Features

\[ p(y|x) \propto \exp(\Theta_y) \]
Example: Linear Regression

**Goal:** Learn $y = \mathbf{w}^T f(x) + b$ where $f(.)$ is a polynomial basis function.

true “unknown” target function is linear with negative slope and gaussian noise.
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.
2.1 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?
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 $D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), \ldots, (x_n^{(1)}, y_n^{(1)})\}$ 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

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.
Apply the backpropagation algorithm to obtain the partial derivative of the mean-squared error of \( y \) with the true value \( y^* \) with respect to the weight \( w_{22} \) assuming a sigmoid nonlinear activation function for the hidden layer.
Material Covered After Midterm Exam

SAMPLE 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.
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?
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.
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:

- \( \text{sex} \in \{\text{male, female}\} \)
- \( \text{height} \in [0, 300] \) centimeters
- \( \text{hair} \in \{\text{brown, black, blond, red, green}\} \)
- 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: \( P(\text{height} | \text{sex, hair}) = P(\text{height} | \text{sex}) \).
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
4.2  Multiple Choice

(a) [3 pt.] If the data is linearly separable, SVM minimizes $\|w\|^2$ subject to the constraints $\forall i, y_i w \cdot x_i \geq 1$. In the linearly separable case, which of the following may happen to the decision boundary if one of the training samples is removed? Circle all that apply.

- Shifts toward the point removed
- Shifts away from the point removed
- Does not change
3. [Extra Credit: 3 pts.] One formulation of soft-margin SVM optimization problem is:

$$\min_w \frac{1}{2} kw^2 + C \sum_{i=1}^{N} \xi_i$$

s.t. \( y_i(w^T x_i) \geq 1 - \xi_i \quad 8i = 1, \ldots, N \)

\( \xi_i \geq 0 \quad 8i = 1, \ldots, N \)

\( C \geq 0 \)

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?
2 K-Means Clustering

(a) [3 pts] We are given $n$ data points, $x_1, ..., x_n$ and asked to cluster them using K-means. If we choose the value for $k$ to optimize the objective function how many clusters will be used (i.e. what value of $k$ will we choose)? **No justification required.**

(i) 1 (ii) 2 (iii) $n$ (iv) $\log(n)$
Samples Questions

2.2 Lloyd’s algorithm

Circle the image which depicts the cluster center positions after 1 iteration of Lloyd’s algorithm.

Figure 2: Initial data and cluster centers
2.2 Lloyd’s algorithm

Circle the image which depicts the cluster center positions after 1 iteration of Lloyd’s algorithm.

Figure 2: Initial data and cluster centers
Sample Questions

4 Principal Component Analysis [16 pts.]

(a) In the following plots, a train set of data points $X$ belonging to two classes on $\mathbb{R}^2$ are given, where the original features are the coordinates $(x, y)$. For each, answer the following questions:

(i) [3 pt.] Draw all the principal components.

(ii) [6 pts.] Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

Dataset 1:

Dataset 2:
Sample Questions

4 Principal Component Analysis

(i) T  F 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.

(ii) T  F The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

(iii) T  F Subsequent principal components are always orthogonal to each other.
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.
Course Level Objectives

You should be able to...

1. Implement and analyze existing learning algorithms, including well-studied methods for classification, regression, structured prediction, clustering, and representation learning
2. Integrate multiple facets of practical machine learning in a single system: data preprocessing, learning, regularization and model selection
3. Describe the formal properties of models and algorithms for learning and explain the practical implications of those results
4. Compare and contrast different paradigms for learning (supervised, unsupervised, etc.)
5. Design experiments to evaluate and compare different machine learning techniques on real-world problems
6. Employ probability, statistics, calculus, linear algebra, and optimization in order to develop new predictive models or learning methods
7. Given a description of a ML technique, analyze it to identify (1) the expressive power of the formalism; (2) the inductive bias implicit in the algorithm; (3) the size and complexity of the search space; (4) the computational properties of the algorithm; (5) any guarantees (or lack thereof) regarding termination, convergence, correctness, accuracy or generalization power.
Q&A