Final Review

Readings:

Matt Gormley
Lecture 27
December 5, 2016
Reminders

• Final Exam
  – Wed, Dec. 7th, in-class

• Final Exam Review Session
  – Tue, Dec. 6th at 5:30pm
Outline

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

• Exam Details
  – In-class exam on Wed, Dec. 7\textsuperscript{th}
  – 7 problems
  – Format of questions:
    • Multiple choice
    • True / False (with justification)
    • Derivations
    • Short answers
    • Interpreting figures
  – No electronic devices
  – You are allowed to \textbf{bring} one 8½ x 11 sheet of notes (front and back)
Final Exam

• How to Prepare
  – Attend the final recitation session: Tue, Dec. 6th at 5:30pm
  – Review prior year’s exams and solutions (we will post them)
  – Review this year’s homework problems

• Grading scheme drops your lowest homework grade
• If you skipped one, be sure to go back to it and carefully study all the questions
Final Exam

• How to Prepare
  – Attend the final recitation session:
    Tue, Dec. 6th at 5:30pm
  – Review prior year’s exams and solutions
    (we will post them)
  – Review this year’s homework problems
  – Flip through the “What you should know” points
    (see ‘More’ links on ‘Schedule’ page of course website)
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
  – MLE, MAP
  – Optimization

• Classifiers
  – Decision Trees
  – Naïve Bayes
  – Logistic Regression
  – Perceptron
  – SVM

• Regression
  – Linear Regression

• Important Concepts
  – Kernels
  – Regularization and Overfitting
  – Sample Complexity
  – Experimental Design
Topics covered after Midterm

• Supervised Learning
  – Boosting

• Unsupervised Learning
  – K-means / Lloyd’s method
  – Hierarchical clustering
  – PCA
  – EM / GMMs

• Neural Networks
  – Basic architectures
  – Backpropagation
  – Why Deep Nets are hard to train
  – CNNs / RNNs

• Graphical Models
  – Bayesian Networks
  – Factor Graphs
  – HMMs / CRFs
  – Learning and Inference

• Other Learning Paradigms
  – Active Learning
  – Semi-Supervised Learning
  – Reinforcement Learning
  – Collaborative Filtering
SAMPLE QUESTIONS
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.
Sample Questions

1 Topics before Midterm

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

(h) [2 pts.] T or F: The VC dimension of a finite concept class $H$ is upper bounded by $\lceil \log_2 |H| \rceil$. 
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) \( \log(n) \)
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
3 Active Learning

3.1 Linear Separators on the Circle

In this problem you will design an active learning algorithm for finding a consistent linear separator passing through the origin when the data is on the unit circle in 2 dimensions. That is, given a dataset \( S = \{x_1, \ldots, x_n\} \) with \( \|x_i\| = 1 \) for all \( i = 1, \ldots, n \), your goal is to find a consistent classifier of the form \( h(x) = \text{sign}(w^\top x) \). Assume we are in the realizable setting.

(a) [8 pts.] First, suppose that our data lies only on the top half of the circle (e.g., see Figure 3a). In 1–2 sentences, describe an algorithm for finding a consistent linear separator passing through the origin using \( O(\log n) \) label queries. Hint: this problem is very similar to learning a consistent threshold function for data on the real line.

(b) Example data on the top half of the circle.
3 Active Learning

(a) On Figure 5a draw the smallest and largest rectangles that correctly classify the labeled examples. In 1–2 sentences, describe the version space in terms of these two rectangles.

(b) On Figure 5a, mark each point in the region of disagreement with the letter ‘d’. For all other points, mark them with their correct label.

(a) Dataset for parts (a) and (b).
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:
4 Principal Component Analysis

(c) [2 pts.] Assume we apply PCA to a matrix $X \in \mathbb{R}^{n \times m}$ and obtain a set of PCA features, $Z \in \mathbb{R}^{m \times n}$. We divide this set into two, $Z_1$ and $Z_2$. The first set, $Z_1$, corresponds to the top principal components. The second set, $Z_2$, corresponds to the remaining principal components. Which is more common in the training data:

A: a point with large feature values in $Z_1$ and small feature values in $Z_2$

B: a point with large feature values in $Z_2$ and small feature values in $Z_1
Sample Questions

(a) [2 pts.] Write the expression for the joint distribution.

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., \( R, S, E, A \in \{0, 1\} \).

![Directed graphical model]

Figure 5: Directed graphical model for problem 5.
(b) [2 pts.] How many parameters, i.e., entries in the CPT tables, are necessary to describe the joint distribution?

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., \( R, S, E, A \in \{0, 1\} \).

![Diagram](image.png)

Figure 5: Directed graphical model for problem 5.
(d) [2 pts.] Is $S$ marginally independent of $R$? Is $S$ conditionally independent of $R$ given $E$? Answer yes or no to each question and provide a brief explanation why.

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying ($S$), being well-rested ($R$), doing well on the exam ($E$), and getting an A grade ($A$). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

![Directed graphical model for problem 5.](image)

Figure 5: Directed graphical model for problem 5.
Sample Questions

5 Graphical Models

(f) [3 pts.] Give two reasons why the graphical models formalism is convenient when compared to learning a full joint distribution.
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
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
Sample Questions

Neural Networks

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.

(b) The neural network architecture
OVERVIEW
Whiteboard

- Overview #1: Learning Paradigms
- Overview #2: Recipe for ML
MATRIX MULTIPLICATION IN MACHINE LEARNING
Recovering latent factors in a matrix

\[ \begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{matrix} \quad \times \quad \begin{matrix} a_1 & a_2 & \ldots & a_m \\ b_1 & b_2 & \ldots & b_m \end{matrix} \quad \approx \quad \begin{matrix} v_{11} \\ \vdots \\ v_{1n} \\ \vdots \\ v_{nm} \end{matrix} \]

\[ V[i,j] = \text{user } i's \text{ rating of movie } j \]
... is like Linear Regression ....

\[ Y[i,1] = \text{instance } i\text{'s prediction} \]
.. for many outputs at once...

... where we also have to find the dataset!

Y[I,j] = instance i’s prediction for regression task j
... vs PCA

Minimize squared error reconstruction error and force the “prototype” users to be orthogonal \( \Rightarrow \) PCA

\[ V[i,j] = \text{user i's rating of movie j} \]
... vs autoencoders & nonlinear PCA

• Assume we would like to learn the following (trivial?) output function:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000001</td>
<td>0000001</td>
</tr>
<tr>
<td>0000010</td>
<td>0000010</td>
</tr>
<tr>
<td>0000101</td>
<td>0000100</td>
</tr>
<tr>
<td>0001000</td>
<td>0001000</td>
</tr>
<tr>
<td>0010000</td>
<td>0010000</td>
</tr>
<tr>
<td>0100000</td>
<td>0100000</td>
</tr>
<tr>
<td>1000000</td>
<td>1000000</td>
</tr>
</tbody>
</table>

• Using the following network:

• With linear hidden units, how do the weights match up to \( W \) and \( H \)?

Flashback to NN lecture.....

Slide from William Cohen
... vs k-means

indicators for r clusters

\[
\begin{pmatrix}
0 & 1 \\
1 & 0 \\
. & . \\
. & . \\
0 & 1 \\
1 & 0 \\
. & . \\
. & . \\
x_n & y_n \\
\end{pmatrix}
\]

cluster means

\[
\begin{pmatrix}
a_1 & a_2 & \ldots & a_m \\
b_1 & b_2 & \ldots & b_m \\
\vdots & \vdots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
\end{pmatrix}
\]

original data set

\[
\begin{pmatrix}
v_{11} & \ldots \\
\ldots & \ldots \\
v_{ij} & \ldots \\
\ldots & \ldots \\
v_{nm} \\
\end{pmatrix}
\]
Recovering latent factors in a matrix

\[
\begin{align*}
\text{r} & \quad \text{m movies} \\
x_1 & \quad y_1 \\
x_2 & \quad y_2 \\
\vdots & \quad \vdots \\
x_n & \quad y_n \\
\text{W} & \\
\text{H} & \quad \text{am} \\
b_1 & \quad b_2 \\
\vdots & \quad \vdots \\
b_m & \\
\text{V[i,j]} & = \text{user i's rating of movie j}
\end{align*}
\]