CS534 Machine Learning - Spring 2013

Final Exam

Name: ____________________

• You have 110 minutes.
• There are 6 questions (8 pages including cover page). If you get stuck on one question, move on to others and come back to the difficult one later.
• If the question asks for explanation, you must provide explanation, otherwise no point will be given.
• The exam is open book and notes, but no computer, cell phone, and internet.
• Good luck!
1. (24 points) True/False questions. No need to explain your answer but each incorrect T/F answer will get -1 point. Providing no answer will get zero point.

   a. Both PCA and Spectral Clustering perform eigen-decomposition of different matrices. However, the size of these two matrices are the same.

   b. When the data is sparse, ISOMAP may fail to find the underlying manifold.

   c. The dimensionality of the feature map generated by polynomial kernels (e.g., \( K(x, y) = (1+x\cdot y)^d \)) is polynomial with respect to the power \( d \) of the polynomial kernel.

   d. There is at least one set of 4 points in \( R^3 \) that can be shattered by the hypothesis space of all linear classifiers in \( R^3 \).

   e. The log-likelihood of the data will always increase through successive iterations of the Expectation Maximization algorithm.

   f. In AdaBoost the weights of the misclassified examples go up by the same multiplicative factor in one iteration.

   g. The more hidden layers a neural network has, the better it predicts desired outputs for new inputs that it was not trained with.

   h. When clustering with Gaussian Mixture Models with \( K \) Gaussians, we can select the best \( K \) as follows: run GMM with different \( K \) values and pick the one that leads to the highest data likelihood \( p(x|\theta) \).
2. (20 pts) **Learning theory**

a. (4 pts) Consider the hypothesis space of decision stumps with binary split. What is its VC-dimension? Briefly justify your answer.

b. (4 pts) Consider the hypothesis space of unlimited-depth decision trees. What is its VC-dimension? Briefly justify your answer.

Consider linear regression using polynomial basis functions of order $M$. The expected loss can be decomposed into three parts, the bias, the variance and the noise.

c. (6 pts) If we increase the order of the polynomial basis function (e.g., from quadratic to cubic, or higher order polynomials), will it increase, decrease or have no impact on the three terms? Briefly explain.

**Bias:**

**Variance:**

**Noise:**

d. (6 pts) If we increase the training set size, will it increase, decrease or have no impact on each of the three terms? Briefly explain.

**Bias:**

**Variance:**

**Noise:**
3. [12 points] **Dimension Reduction.** In a supervised classification task we have a total of 10 features, among which only three are useful for predicting the target variable, the other features are pure random noise with very high variance. What complicates matters even worse is that the three features when considered individually show no predictive power, and only work when considered together. Consider each of the following dimension reduction techniques and indicate whether it can successfully identify the relevant dimensions. Briefly explain why.

- Feature selection with greedy forward search

- Feature selection with greedy backward search

- Linear discriminant analysis

- Principle Component Analysis

a. Consider a data set $D = x_1, x_2, ..., x_{10}$. We apply Adaboost with decision stump. Let $w_1, \ldots, w_{10}$ be the weights of the ten training examples respectively.

i. (4pts) In the first iteration, $x_1, x_2$ and $x_3$ are misclassified. Please provide the weights after update and normalization in the first iteration.

ii. (6pts) In the second iteration, $x_3$ and $x_4$ are misclassified. Please provide the weights after update and normalization at the end of this iteration.

b. (4pts) If the training set contains noise in class labels, which ensemble learning method do you expect to be hurt more by the label noise, boosting or bagging? Why?
5. [15 pts] **Clustering: K-Means, GMM and Spectral**

   a. [3 points] Why is it important to randomly restart $k$-means multiple times?

   b. [4 points] Apply Gaussian Mixture Models to cluster the following data into two clusters. Consider two different GMM models. The first variant restricts the covariance matrix $\Sigma$ of both clusters to be diagonal, whereas the second variant makes no such restriction. Please mark out what you think the final clusters (Gaussians) would look like for each variant in the figure.

   ![Diagram of two clusters](image1)

   (a) Diagonal $\Sigma$ (b) Unrestricted $\Sigma$

   c. [4 pts] Consider the following clustering problem where we want to partition the data into two clusters. We consider clustering with both the min-cut and normalized-cut objectives as defined in class. Please mark out the clustering results for both objectives.

   ![Diagram of two clusters](image2)

   (a) Min-cut (b) Normalized cut
d. [4 points] Apply K-means to the following 2-d data set for $k = 2$. Start with $\mu_1(0) = (0, 0)$, and $\mu_2(0) = (4, 3)$. Use the figures provided below to mark the means and the cluster membership after (1) the first iteration and (2) after the algorithm converges. Use as many figures as needed to work through iterations. Simply mark out the first iteration and the final result.
6. (10pts) Expectation maximization. Imagine a machine learning class where the probability that a
student gets an A grade is \( P(A) = \frac{1}{2} \), a B grade \( P(B) = \mu \), a C grade \( P(C) = 2\mu \), and a D grade
\( \frac{1}{2} - 3\mu \). Given a set of students we observe that there are \( c \) students who got C and \( d \) students
who got D. We also know that a total of \( h \) students got either A or B, but we don’t observe \( a \), the
total number of A students, and \( b \), the total number of B students. We only observe their sum, which
is \( h \). The goal is to apply EM to obtain a maximum likelihood estimation of \( \mu \).

– (5 pts) Expectation. Which of the following formula computes the expected value of \( a \) and \( b \) given
\( \mu \)?

\[
\hat{a} = \frac{1/2}{1/2 + h} \mu \quad \hat{b} = \frac{\mu}{1/2 + h} \mu
\]

\[
\hat{a} = \frac{1/2}{1/2 + \mu} h \quad \hat{b} = \frac{\mu}{1/2 + \mu} h
\]

\[
\hat{a} = \frac{\mu}{1/2 + \mu} h \quad \hat{b} = \frac{1/2}{1/2 + \mu} h
\]

\[
\hat{a} = \frac{1/2}{1 + \mu^2} h \quad \hat{b} = \frac{\mu}{1 + \mu^2} h
\]

– (5 pts) Maximization. Given the expected value of \( a \) and \( b \), which of the following formula
computes the maximum likelihood estimate of \( \mu \)? Hint: compute the MLE of \( \mu \) assuming that
unobserved variables are replaced by their expectation,

\[
\hat{\mu} = \frac{h - a + c}{6(h - a + c + d)}
\]

\[
\hat{\mu} = \frac{h - a + d}{6(h - 2a - d)}
\]

\[
\hat{\mu} = \frac{h - a}{6(h - 2a + c)}
\]

\[
\hat{\mu} = \frac{2(h - a)}{3(h - a + c + d)}
\]