CSE 575 Statistical Machine Learning
Final Exam
December 6, 2016

1. Personal info:
   - Name:
   - ASU ID#:
2. There should be 12 numbered pages in this exam (including this cover sheet).
3. THIS IS CLOSED BOOK EXAM!
4. Work efficiently. Some questions are easier, some more difficult. Be sure to give yourself time to answer all of the easy ones, and avoid getting bogged down in the more difficult ones before you have answered the easier ones.
5. You have 110 minutes.
6. Good luck!

Question	Topic	Max. score	Score
1	Short Questions	16	
2	Support Vector Machines	33	
3	Linear Regression	24	
4	Number of Parameters	27	

1 [16 points] Short Questions

1. [3 points] Suppose that we wish to calculate $P(H|E_1, E_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation?

   (a) $P(E_1, E_2)$, $P(H)$, $P(E_1|H)$, $P(E_2|H)$

   (b) $P(E_1, E_2)$, $P(H)$, $P(E_1, E_2|H)$

   (c) $P(H)$, $P(E_1|H)$, $P(E_2|H)$

   Solution: (b)

2. [3 points] Suppose we know that $P(E_1|H, E_2) = P(E_1|H)$ for all values of H, E_1, E_2. Which of the following sets of numbers are sufficient for calculating $P(H|E_1, E_2)$?

   (a) $P(E_1, E_2)$, $P(H)$, $P(E_1|H)$, $P(E_2|H)$

   (b) $P(E_1, E_2)$, $P(E_1, E_2|H)$

   (c) $P(E_1|H)$, $P(E_2|H)$

   Solution: (a)

3. [3 points] For k-NN, we usually chose k to be an odd number. Why?

   Solution: To avoid the cases where we have equal numbers of positive/negative examples.

4. [1 points] [**True or False**] All of the following algorithms are greedy algorithms and they can only guarantee a local optimal solution: (1) K-means, and (2) EM for GMM.

   Solution: True

5. [3 points] If the training data is noise-free, can we use 1-NN as the weak learning algorithm in AdaBoost? Why?

   Solution: No. 1-NN training error is always 0.

6. [3 points] If the training set contains 2 different examples in the 2-dimensional space, what is the reconstruction error of PCA using only 1 principle component?

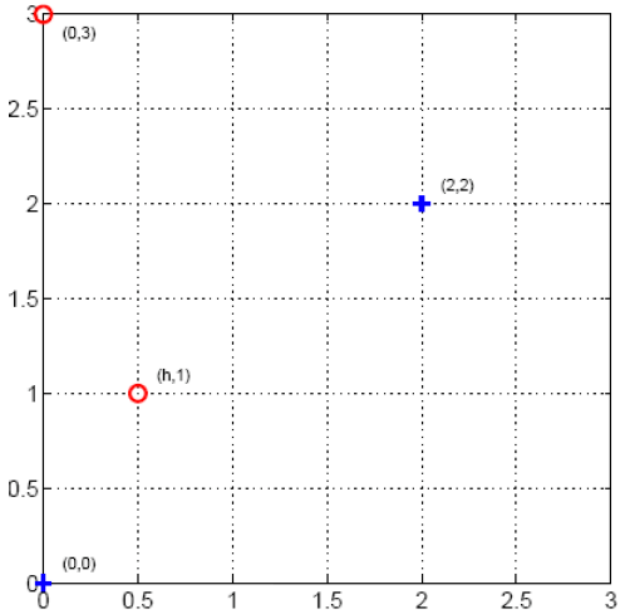
   Solution: 0. The two examples are always on a line, which corresponds to the only principle component.

2 [33 points] Support Vector Machines

1. [7 points] Consider a two-dimensional input space $x = (x_1, x_2)$ ($z = (z_1, z_2)$), and a kernel function defined as follows: $k(x, z) = (x_1 z_1 + x_2 z_2)^3$. Derive the corresponding feature mapping function $\Phi(x)$ that satisfies the following equation: $k(x, z) = \Phi(x) \cdot \Phi(z)$.

   Solution: $\Phi(x) = [x_1^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, x_2^3]^T$.

2. [12 points] Suppose we only have 4 training examples in two-dimensional space (see Figure 1): 2 positive examples at $x_1 = [0, 0]$, $x_2 = [2, 2]$, and 2 negative examples at $x_3 = [h, 1]$, $x_4 = [0, 3]$, where the value of h is between 0 and 3.



- [6 points] How large can h be so that the training examples are still linearly separable?

  Solution: 1.

- [6 points] What is the margin achieved by the maximum margin boundary when $h = 0.5$?

  Solution: $\frac{\sqrt{2}}{8}$. They can get half the points for $\frac{\sqrt{2}}{4}$.

3. [8 points] Given n training examples (x_i, y_i), where x_i is the input feature, y_i is the class label, $i = 1, \ldots, n$. Let $n \times n$ matrix $\boldsymbol{A}$ denote the kernel matrix, i.e., $A(i, j) = k(x_i, x_j)$. Prove that $\boldsymbol{A}$ is semi-positive definite.
Hint: Matrix $\boldsymbol{A}$ is semi-positive definite if and only if for any n-dimensional vector f, we have $f'\boldsymbol{A}f \geq 0$.

   Solution: $A = [\Phi(x_1), \ldots, \Phi(x_n)]^T[\Phi(x_1), \ldots, \Phi(x_n)]$.

4. [6 points] Using the notation in the last question, let ξ_i denote the slack variable for x_i. Please provide an upper bound on the training error using ξ_i, $i = 1, \ldots, n$.

   Solution: $\sum_i \xi_i$.

3 [24 points] Linear regression

Suppose that we have a data set where the i^{th} example has one real-valued input feature x_i and one real-valued output y_i. We use linear regression to model the data.

1. [4 points] We start with a small subset as the training set, and gradually increase the size of the training set. How would the training error change in general, and why?

   Solution: First increases, and then stabilizes. Deduct 1 point if they miss the 'stabilizes' part.

2. [4 points] As we gradually increase the size of the training set, how would the true error of the linear regression model change in general, and why?

   Solution: First decreases, and then stabilizes. Deduct 1 point if they miss the 'stabilizes' part.

3. [5 points] For the linear regression model trained on infinite amount of data, how would the training error compare with the true error, and why?

   Solution: The same. Training with infinite amount of data is equivalent to training with the knowledge of the underlying distribution. So the training error is the same as the true error.

4. [5 points] Let M_1 denote the linear regression model using the one-dimensional input feature x_i alone; let M_k denote the linear regression model using $x_i, (x_i)^2, \ldots, (x_i)^k$, where $k \geq 1$ is a positive integer. Given the same training set, how would the training error change with increasing value of k, and how would the true error change with increasing value of k? Please draw a figure to how your answer, where the x-axis corresponds to k, and the y-axis corresponds to the true error of the associated model M_k.

   Solution: The training error decreases with increasing value of k; the true error first decreases, and then increases. The figure should demonstrate the general trend.

5. [4 points] Using the notation in the last question, how do you pick the optimal value of k that corresponds to the optimal model M_k?

   Solution: Cross validation. LOOCV is also ok.

6. [2 points] Let β denote the coefficient vector in the linear regression model. Provide 1 common regularizer of β.

   Solution: $\|\beta\|^2$.

4 [27 points] Number of Parameters

In this question, we consider a data set with 3 categorical input features (A, B, and C), and one categorical output Y.

1. [5 points] Suppose that you are using Naive Bayes classifier for predicting Y. Draw the Bayes Network that represents the Naive Bayes classifier.

   Solution: Y is the parent of A, B, and C.

2. [5 points] Suppose that the arity of A, B, and C is 3, and Y is binary. How many parameters are needed in the Naive Bayes classifier?

   Solution: 13.

3. [2 points] In your Bayes Network, if you add one more feature D as the child of A, how would the new Bayes Network look like?

   Solution: In addition to the Bayes Network from the first question, add D as the child of A.

4. [5 points] For the new Bayes Network in the last question, please list all the conditional independence statements according to the local Markov assumption.

   Solution: $A \perp\!\!\!\perp B|Y$, $A \perp\!\!\!\perp C|Y$, $B \perp\!\!\!\perp C|Y$, $B \perp\!\!\!\perp D|A, Y$, $C \perp\!\!\!\perp D|A, Y$.

5. [5 points] Please decompose the joint probability $P(A, B, C, D, Y)$ according to the new Bayes Network.

   Solution: $P(A, B, C, D, Y) = P(Y)P(A|Y)P(B|Y)P(C|Y)P(D|A)$.

6. [2 points] Suppose that D is binary. How many parameters are needed in the new Naive Bayes classifier?

   Solution: 16.

7. [3 points] Suppose that D follows univariate Gaussian distribution. How many parameters are needed in the new Naive Bayes classifier?

   Solution: 19.