
Practice Final Exam

Instructions: read these first!

Do not open the exam, turn it over, or look inside until you are told to begin.

Switch off cell phones and other potentially noisy devices.

Write your *full name* on the line at the top of this page. Do not separate pages.

You may refer to a calculator and your cheat-sheet. You may not refer to any other printed material, and *any other computational device* (such as laptops, phones, iPads, friends, enemies, pets, lovers).

Read questions carefully. Show all work you can in the space provided.

Where limits are given, write no more than the amount specified.

The rest will be ignored.

Avoid seeing anyone else's work or allowing yours to be seen.

Do not communicate with anyone but an exam proctor.

If you have a question, raise your hand.

When time is up, stop writing.

You can see your graded final in the student affairs office starting Mon Mar 27.

Question	Points	Score
1	12	
2	12	
3	12	
4	10	
5	25	
6	25	
Total:	96	

1. [12 points] In this question, we will again look at how k -nearest neighbor classifiers can be robust to noise. Suppose we have two labels 0 and 1. We are given a test point x , and its k nearest neighbors $z_1, \dots, z_k$, where z_i is the i -th closest neighbor of x (so z_1 is the closest neighbor, z_2 is the second closest neighbor and so on).

Suppose that the probability that the label of z_i is not equal to the label of x is p_i . Also assume that all distances are unique, so the i -th closest neighbor of x is unique for all i .

Answer the following questions:

- (a) [3 points] Now, suppose, for this question that $p_1 = 0.1$, and $p_i = 0.2$ for $i > 1$. What is the probability that the 1-nearest neighbor classifier makes a mistake on x ?

- (b) [6 points] In the setting of part (a), what is the probability that the 3-nearest neighbor classifier makes a mistake on x ?

- (c) [3 points] Based on your calculation, what can you conclude about the relative robustness of 1 and 3-nearest neighbor classifiers in this case?

2. [12 points] Recall that the projection of a vector x onto another vector y is defined as the vector $\frac{\langle x, y \rangle}{\|y\|^2} y$. Suppose $d > 1$ and suppose that $u_1 \neq u_2$ are two $d \times 1$ vectors such that $\|u_1\| = \|u_2\| = 1$, and u_1 and u_2 are *not orthogonal* to each other. Let x be a third non-zero vector.

Alice and Bob have been asked to take projections of x onto u_1 and u_2 . Alice first projects x onto u_1 and then projects the result onto u_2 to get the vector a . Bob first projects x onto u_2 and then projects the result onto u_1 to get the vector b .

(a) [3 points] Is $a = b$ always?

(b) [5 points] Is $\|a\| = \|b\|$ for all x ? Justify your answer if this is the case; if it is not the case, provide a counter-example.

(c) [4 points] Does your answer change when u_1 and u_2 are orthogonal?

3. [12 points] For each of the following statements, say whether they are correct or incorrect. Justify your answer.
- (a) [4 points] For every training data set $(x_1, y_1), \dots, (x_n, y_n)$, the training error of the 3-nearest neighbor classifier is always zero. You may assume that each x_i is unique.
- (b) [4 points] As we increase k , the training error of the k -nearest neighbor classifier always increases.
- (c) [4 points] If u and v are any two orthogonal unit vectors, then $\|u + v\| = 1$.

4. [10 points] Suppose you are given the following feature vectors:

$$x_1 = (0, 4), x_2 = (0, 2), x_3 = (2, 0), x_4 = (4, 0), x_5 = (5, 0)$$

Their corresponding labels are

$$y_1 = 1, y_2 = -1, y_3 = -1, y_4 = 1, y_5 = 1$$

Consider four weak learners h_1, h_2, h_3, h_4 , which are: $h_i(x) = \text{sign}(\langle w_i, x \rangle + b_i)$, for $i = 1, 2, 3, 4$. The parameters are $w_1 = w_2 = (1, 0)$, $b_1 = -1$, $b_2 = -3$, $w_3 = w_4 = (0, 1)$, $b_3 = -1$, $b_4 = -3$.

Using these weak learners, we will now create an ensemble classifier H using boosting. In each round, we will pick the best weak learner – that is, the learner that has the lowest classification error. Write down the weak learners h_t picked in the first three rounds of boosting as well as the corresponding α_t values.

5. [25 points] State whether each of the following statements is true or false. If it is true, provide a brief justification or proof; if it is false, provide a counterexample or a justification.
- (a) [5 points] Suppose we are given a data set $(x_1, y_1), \dots, (x_n, y_n)$ that is linearly separable. If the feature vectors have zero mean (that is, if $\frac{1}{n} \sum_{i=1}^n x_i = 0$), then, the data set $(x_1, y_1), \dots, (x_n, y_n)$ is also linearly separable through the origin.
- (b) [5 points] Let X and Z be two random variables such that $Z = X - 2$. Then $H(Z) = H(X)$.
- (c) [5 points] If the training data is linearly separable, then running perceptron on one pass of the data is guaranteed to learn a classifier with zero training error.

-
- (d) [5 points] The decision boundary of averaged perceptron is always a hyperplane, no matter what the training data is.
- (e) [5 points] Two decision trees that have the same training error on a training dataset S also have the same test error on the same test dataset T , no matter what S and T are.

6. [25 points] State whether each of the following statements is true or false. If it is true, provide a brief justification or proof; if it is false, provide a counterexample or a justification.

(a) [5 points] x, z are real d -dimensional vectors. Then the function $K(x, z) = \frac{\langle x, z \rangle}{\|x\|^2 \|z\|^2}$ is a kernel.

(b) [5 points] x, z are real d -dimensional vectors. Then the function $K(x, z) = \|x - z\|^2$ is a kernel.

(c) [5 points] If $K(x, z)$ and $L(x, z)$ are kernels, then $M(x, z) = 2K(x, z) - 3L(x, z)$ is a kernel, no matter what K and L are.

- (d) [5 points] If $K(x, z)$ and $L(x, z)$ are kernels, then $M(x, z) = K(x, z)L(x, z)^2$ is also a kernel, no matter what K and L are.

- (e) [5 points] The matrix B is a valid kernel matrix:

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \quad (1)$$