ɔ0B, Winter 2019
.uctor: Kamalika Chaudhuri Name: PID:

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	0	
3	12	
4	12	
5	25	
6	25	
Total:	86	

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 $z_1, \ldots, z_k$, 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?

   $$\Pr(\text{1-NN makes mistake on } x) = \Pr(\text{label of } z_1 \neq \text{label of } x)$$
   $$= 0.1$$

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

   $$\Pr(\text{1-NN makes mistake on } x) = \Pr(y_1 \neq y, y_2 \neq y, y_3 \neq y)$$
   $$+ \Pr(y_1 \neq y, y_2 \neq y, y_3 = y)$$
   $$+ \Pr(y_1 \neq y, y_2 = y, y_3 \neq y)$$
   $$+ \Pr(y_1 = y, y_2 \neq y, y_3 \neq y)$$
   $$= (0.1)(0.2)(0.2) + (0.1)(0.2)(0.8) + (0.1)(0.8)(0.2)$$
   $$+ (0.9)(0.2)(0.2)$$
   $$= 0.072$$

   (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?

   3-NN is more robust

2. [10] Solve the following optimization problem by the substitution method. Write down the KKT conditions for it, and use the solutions to derive the values of the dual variables.

$$\begin{aligned} \min \quad & x+y \\ \text{subject to:} \quad & \\ & x^2+y^2=4 \qquad x^2+y^2-4=0 \\ & x \geq 0 \end{aligned}$$

$$\mathcal{L}(x,y,\lambda) = x+y-\lambda_1(x^2+y^2-4)-\lambda_2 x$$

KKT conditions:

$$\nabla_{(x,y)}\mathcal{L}(x,y,\lambda) = \begin{bmatrix} 1-2\lambda_1 x-\lambda_2 \\ 1-2\lambda_1 y \end{bmatrix} = 0.$$

$$\lambda_2 \geq 0, \quad \lambda_1(x^2+y^2-4)=0, \quad \lambda_2 x = 0$$

$$x^2+y^2-4=0, \quad x \geq 0$$

Solve by substitution:

$x=\sqrt{4-y^2}$ (since $x \geq 0$, no need to consider the other case)

We now want to solve $\min\limits_{y} \sqrt{4-y^2}+y$.

$\Rightarrow y=-2, \quad x=0.$

Solve dual variables:

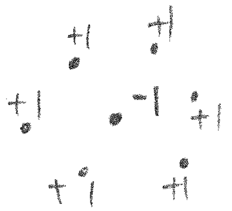
$1-2\lambda_1 y=0 \Rightarrow \lambda_1=-\frac{1}{2}$

$1-2\lambda_1 x-\lambda_2=0 \Rightarrow \lambda_2=1$

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 $(x_1, y_1), \ldots, (x_n, y_n)$, the training error of the 3-nearest neighbor classifier is always zero. You may assume that each x_i is unique.

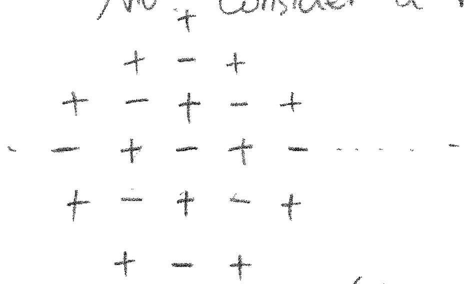
No.



only one point in the training set has label -1.

(b) [4 points] As we increase k, the training error of the k-nearest neighbor classifier always increases.

No. Consider a very large 2-D grid. Adjacent vertices have different label.



$k=5$, training error ≈ 1, almost all points are mis-classfied.

$k=13$, training error ≈ 0. almost all points are correctly classified.

(Another example is that when $k > n$, error should stay the same.)

(c) [4 points] Recall that two classifiers c_1 and c_2 are equal, if $c_1(x) = c_2(x)$ for all x in the domain. If two classifiers c_1 and c_2 have the same training error, then they are equal.

No. Feel free to construct your own counter example.

4. [10 points] Calculate the gradient and the Hessian for the following functions, and use them to determine if they are convex or not. Justify your answer.

(a) [4 points] $f(x) = \sum_{i=1}^{d} x_i \log x_i$.

$$\frac{\partial f(x)}{\partial x_i} = x_i \cdot \frac{1}{x_i} + \log x_i = 1 + \log x_i$$

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \begin{cases} 0 & \text{if } i \neq j \\ \frac{1}{x_i} & \text{if } i = j \end{cases}$$

The Hessian H is a diagonal matrix with $H_{ii} = \frac{1}{x_i}$.

It is P.S.D. if $x_i > 0$ for all i. (x_i must >0 for $\log x_i$ to be defined)

(b) [4 points] $f(x) = 20x^\top x - 5x^\top \mathbf{1}$. where $\mathbf{1}$ is the all-ones vector.

Recall that the Hessian of $x^T M x = M + M^T$

$20x^T x = 20x^T I x$

Hessian of $20x^T I x = 20(I + I^T) = 40I$

...... $-5x^T 1 = 0$.

Therefore, the Hessian $= 40I$, which is P.S.D.

(c) [4 points] Now suppose that $f(x)$ is convex, and let $g(x) = (f(x))^2$. Write down the gradient and Hessian of $g(x)$ in terms of those of f. If f is convex, is $g(x)$ always convex?

A 1-D example is
$f(x) = x^2 - 1$
$g(x) = (x^2 - 1)^2$.

$$\nabla g(x) = \frac{\partial g(x)}{\partial x} = 2f(x)\frac{\partial f(x)}{\partial x} = 2f(x)\nabla f(x)$$

$$H_g(x) = \frac{\partial^2 g(x)}{\partial x^2} = 2\left[\nabla f(x)(\nabla f(x))^T + f(x) H_f(x)\right]$$

g(x) is not always convex.

Let x be the global minimum of $f(x)$, and $f(x) < 0$

$\nabla f(x) = 0$, $H_f(x)$ is P.S.D., $f(x) H_f(x)$ is not P.S.D because $f(x) < 0$.

⇒ $H_g(x)$ is not P.S.D.

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 $(x_1, y_1), \ldots, (x_n, y_n)$ 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 $(x_1, y_1), \ldots, (x_n, y_n)$ is also linearly separable through the origin.

No, the intuition is that the mean of all x does not provide information about y.

Example. $x_1 = 1$, $y_1 = -1$

$x_2 = 2$, $y_2 = +1$

$x_3 = -3$, $y_3 = -1$

(b) [5 points] Let X and Z be two random variables such that $Z = X - 2$. Then $H(Z) = H(X)$.

Yes. Apply the formula for H and use substitution of variables.

(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.

No, the intuition is that some training point can be more "extreme" than the others. The perceptron may overfit to that point in the first round.

Example: the ground truth labeling function has weight $[1, 1]$.

2 training points $x_1 = [100, 1]$, $y_1 = +1$

$x_2 = [-0.1, 1]$, $y_2 = +1$

$w_0 = 0$

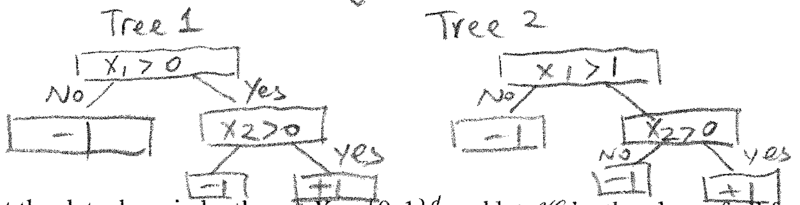
$w_1 = [100, 1]$, $w_2 = [99.9, 2]$, w_2 still classifies x_2 wrong.

(d) [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.

No, one intuition is that there are regions where the training set does not sufficiently represent the underlying distribution.

Alternatively, as long as the two classifiers are not equal on all possible input x, the claim is wrong.

Example



$S = \{([2,1],+1)\}$

$T = \{([0.5,1],+1)\}$

$x \in \mathbb{R}^2$, $y \in \{-1,+1\}$

(e) [5 points] Let the data domain be the set $X = \{0,1\}^d$ and let $\mathscr{H}$ be the class of all functions on X. Then, $|\mathscr{H}| = 2^d$.

No. There are 2^d point in the input domain.
so there 2^{2^d} ways of labeling.

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.

Yes, the feature map ϕ is $\phi(x) = \frac{x}{\|x\|^2}$.

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

No, the intuition is that $K(x,x) = 0 < K(x,z) \ \forall z \neq x$. It doesn't look like a "similarity measure".

$x = [1, 0, \dots 0]$. $z = [2, 0, \cdots 0]$

The kernel matrix is $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, which is not P.S.D.

(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.

No, let $K(x,z) = L(x,z) = 1$ for all $x, z \in \mathbb{R}^d$.

The feature map ϕ is $\phi(x) = 1$

$M(x,z) = -1$ is not a kernel

(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.

Yes. product of kernels is a kernel.

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

$$B = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \tag{1}$$

No. $\det(B) < 0 \Rightarrow B$ is not P.S.D.