CSE 250B Winter 2019 - Quiz 1

Jan 25, 2019 - 20 Minutes **Name:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Remember that a classifier is just a function that takes a feature vector x in a vector space X and maps it to a discrete label y. Two classifiers C_1 and C_2 are said to be equal if for all $x \in X$, $C_1(x) = C_2(x)$.

Now suppose we have a training dataset S, and we have two 1-nearest neighbor classifiers C_1 and C_2, both of which are trained on S, and both of which use the same tie-breaking rule. State whether C_1 and C_2 are equal in the following two cases for all possible values of S. Briefly justify your answer by either a short proof or justification, or a counterexample.

(1) (5 points) C_1 uses the L_1-distance (to measure distance between training examples and the test point) and C_2 uses the L_2-distance.

C_1 is generally not equal to C_2 in this case. For a brief counterexample, suppose S has two points $(3, 3)$ with label 0 and $(5, 0)$ with label 1. Pick a test point $x = (0, 0)$.

$C_1(x) = 1$ as $\|x - (3, 3)\|_1 = 6 > \|x - (5, 0)\|_1 = 5$. But $C_2(x) = 0$ as $\|x - (3, 3)\|_2 = 3\sqrt{2} = 4.2 < \|x - (5, 0)\|_2 = 5$.

(2) (5 Points) C_1 uses the L_2-distance (to measure distance between training examples and the test point) and C_2 uses the square of the L_2-distance. (Note that the "distance" used in nearest neighbors does not always have to obey the triangle inequality.)

Here, C_1 is equal to C_2. Pick any point x; the closest neighbor of x within the training set S in L_2-distance is going to be the closest neighbor of x within S in the square of the L_2-distance. This means that both C_1 and C_2 will output the same label for x. Since this holds for any test point x, C_1 and C_2 are equal.

CSE 250B Winter 2019 - Quiz 2

Feb 1, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.
(2) Start writing when instructed. Stop writing when your time is up.
(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Remember that in lecture, we talked about the statistical learning framework, where training, validation and test data are all independent samples drawn from the same underlying data distribution. State whether the statistical learning framework assumption applies in the following cases. In each case, if your answer is yes, explain why. If your answer is no, explain what is different between training and test data – μ (the marginal over x), η (the conditional distribution of $y|x$), or something else – and justify your answer.

(1) (5 points) Alice wants to build a classifier that detects if a patient's chest X-ray shows tuberculosis. She gathers a large training data set by collecting all patient chest X-rays from the UCSD hospital, and getting UCSD doctors to label them; this data is then used to build a classifier. The classifier is tested in the Mayo Clinic in New York.

The statistical learning framework does not hold – μ changes between training and test, but η does not. Whether a particular chest x-ray corresponds to tubercolosis or not still remains the same – whether the x-ray comes from a patient in San Diego or New York; however, the types and distribution of patients in the two locations may be different. For example, San Diego may have more patients with a certain type of tubercolosis.

(2) (5 Points) Bob wants to build a classifier that can predict whether a candidate would be a good engineer in his company. For this, he collects historical data of all applicants for engineers, along with who was hired; this is used to build a classifier, which is then tested on future job applications.

The statistical learning framework does not hold here either, as both μ and η are different between training and test. μ changes because the distribution of applicants and their properties changes over time. η is also different – whether a person is hired as an engineer or not often reflects biases of the hiring party, and is not always a good reflection of whether they turn out to be a good engineer or not.

CSE 250B Winter 2019 - Quiz 3

Feb 8, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Alice has collected a dataset of dependent and independent variables $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$; she does linear regression on it and obtains the weight vector w_{Alice}. Bob also collects the same dataset, but while recording the independent variables (the $x^{(i)}$'s) he uses different units. Specifically, Bob's dataset is $\{(z^{(1)}, y^{(1)}), \ldots, (z^{(n)}, y^{(n)})\}$, where for each i, $z^{(i)} = cx^{(i)}$, where $c > 0$ is a scalar. Bob does linear regression on this dataset and obtains a weight vector w_{Bob}.

(1) (5 points) Do Alice and Bob have the same training loss? Is $w_{Alice} = w_{Bob}$? In either case, justify your answer.

If X is Alice's data matrix, and Z is Bob's data matrix, then $Z = cX$. $w_{Bob} = (Z^\top Z)^{-1} Z^\top y = (c^2 X^\top X)^{-1} c X^\top y = w_{Alice}/c$; thus when $c \neq 1$, $w_{Alice} \neq w_{Bob}$.

The training loss of Bob is: $\|Zw_{Bob} - y\|^2 = \|cX \cdot w_{Alice}/c - y\|^2 = \|Xw_{Alice} - y\|^2$, which by definition is equal to the training loss of Alice.

(2) (5 Points) Suppose now that Bob records each feature in a different unit; that is, for each coordinate j, $z_j^{(i)} = c_j x_j^{(i)}$ for all i, where $c_j > 0$ is a scalar, and the c_j's are not all equal. Do Alice and Bob still have the same training loss and is $w_{Alice} = w_{Bob}$? If yes, justify your answer. If no, provide an example of a dataset where this is not the case.

The answer will still be the same. Let C be a diagonal matrix whose i-th diagonal entry is c_i. Observe that Bob's data matrix $Z = X \cdot C$. Moreover, for any w, $ZC^{-1}w = Xw$. Hence, the optimal solution to the training loss for Bob $w_{Bob} = C^{-1} w_{Alice}$, and the training loss for Bob is: $\|Zw_{Bob} - y\|^2 = \|Xw_{Alice} - y\|^2$ which is equal to Alice's training loss.

CSE 250B Winter 2019 - Quiz 4

Feb 15, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

You are given below two functions f and g that map d-dimensional vectors x into scalars. For each of these functions, calculate the gradient and the Hessian. Recall that a function is convex if its Hessian is positive semi-definite at all inputs. Use the Hessian to determine whether each function is convex, and justify your answer.

(1) (5 points) $f(x) = e^{-\frac{1}{2}x^\top x}$.

$\nabla f(x) = -x \cdot e^{-\frac{1}{2}x^\top x}$.

$\nabla^2 f(x) = (xx^\top - I) \cdot e^{-\frac{1}{2}x^\top x}$.

$f(x)$ is not convex as the Hessian is not PSD at all x. In particular, if $x = [0, 0, \ldots, 0]$, then the Hessian at x is the matrix $diag(-1, -1, -1, \ldots, -1)$, which is not PSD – if $z = [1, 0, \ldots, 0]$, then $z^\top \nabla^2 f(x) z = -1 < 0$.

(2) (5 Points) Suppose $z^{(i)} \in \mathbb{R}^d$, for $i = 1, \ldots, n$. $g(x) = \sum_{i=1}^n (e^{x^\top z^{(i)}} - x^\top z^{(i)})$.

$\nabla g(x) = \sum_{i=1}^n z^{(i)} \cdot e^{x^\top z^{(i)}} - z^{(i)}$.

$\nabla^2 g(x) = \sum_{i=1}^n z^{(i)} \cdot (z^{(i)})^\top \cdot e^{x^\top z^{(i)}}$.

g is convex as the Hessian is PSD at all x. We prove it as follows.

For any x and $z^{(i)}$, $e^{x^\top z^{(i)}} > 0$; moreover, $z^{(i)} \cdot (z^{(i)})^\top$ is PSD – as for any vector $w \in \mathbb{R}^d$, we have $w^\top z^{(i)} \cdot (z^{(i)})^\top w = \|w^\top z^{(i)}\|^2 \geq 0$.

If c_i are scalars that are > 0 and if A_i are PSD matrices, then $\sum_i c_i A_i$ is also PSD; this is because for any vector w, $w^\top (\sum_i c_i A_i) w = \sum_i c_i w^\top A_i w \geq 0$ as each individual term $c_i w^\top A_i w \geq 0$. Plugging in $c_i = e^{x^\top z^{(i)}}$, and $A_i = z^{(i)} \cdot (z^{(i)})^\top$, we get that the Hessian is PSD at all x.

CSE 250B Winter 2019 - Quiz 5

Feb 22, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Consider the following optimization problem:

$$\begin{aligned} \min \quad & x + 5y \\ \text{subject to:} \quad & xy = 4 \\ & x \geq 0, y \geq 0 \end{aligned}$$

(1) (2 Points) First, solve the optimization problem by using the substitution method (eg, by substituting $y = 4/x$ in the equality constraint.) What is the optimal value of the objective function?

Solution: $x = 2\sqrt{5}$, $y = 2/\sqrt{5}$. Optimal solution value is $2\sqrt{5} + 2\sqrt{5} = 4\sqrt{5}$.

(2) (8 Points) Write down the Lagrangean for the optimization problem. Write down all the KKT conditions. Use the KKT conditions and the optimal solution to solve for the values of the Lagrangean multipliers. Where needed, justify your answer.

The Lagrangean is: $L(x, \lambda, \nu) = x + 5y - \lambda_1 x - \lambda_2 y + \nu(xy - 4)$. The KKT conditions are:

$$1 - \lambda_1 + \nu y = 0 \tag{1}$$

$$5 - \lambda_2 + \nu x = 0 \tag{2}$$

$$xy = 4 \tag{3}$$

$$x \geq 0, \qquad y \geq 0 \tag{4}$$

$$\lambda_1 \geq 0, \qquad \lambda_2 \geq 0 \tag{5}$$

$$\lambda_1 x = 0 \tag{6}$$

$$\lambda_2 y = 0 \tag{7}$$

$$\nu(xy - 4) = 0 \tag{8}$$

Due to complementary slackness, and because $x, y \geq 0$ at the optimal solution, $\lambda_1 = \lambda_2 = 0$. Moreover, $\nu = -1/y^* = -\sqrt{5}/2$.

CSE 250B Winter 2019 - Quiz 6

Mar 1, 2019 - 20 Minutes **Name:** **PID:**

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Let $v_1, \ldots, v_k$ be k vectors in $\mathbb{R}^d$. Recall that the subspace spanned by $v_1, \ldots, v_k$ is defined as the set of all vectors $\sum_{i=1}^k c_i v_i$ where c_i's are *any* scalars. The cone spanned by $v_1, \ldots, v_k$ is defined as the set of all vectors $\sum_{i=1}^k c_i v_i$ where the c_i's are *positive* scalars.

Suppose we are given training data with binary ±1 labels, and we are using the Perceptron algorithm to train a classifier $sign(w^\top x + b)$ on training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$.

State whether following statements are true or false, and justify your answer in each case.

(1) (4 points) Does w lie in the subspace spanned by $x^{(1)}, \ldots, x^{(n)}$?

True. From the perceptron algorithm, note that we begin with $w_1 = 0$, and update w_t as $w_t = w_t + y^{(t)} x^{(t)}$; this means that the w at the end can be written as:

$$w = \sum_{j \in I} y^{(j)} x^{(j)}$$

where I is the set of indices where updates are made. Thus w is a linear combination of the $x^{(i)}$'s and therefore it lies in the subspace spanned by the $x^{(i)}$'s.

(2) (3 points) Does w always lie in the cone spanned by $x^{(1)}, \ldots, x^{(n)}$?

False. This is because some of the $y^{(i)}$'s may be -1.

(3) (3 points) Does w lie in the cone spanned by $y^{(1)}x^{(1)}, \ldots, y^{(n)}x^{(n)}$?

True – for this, $c_i = 1$.

CSE 250B Winter 2019 - Quiz 7

Mar 8, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.

(2) Start writing when instructed. Stop writing when your time is up.

(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

State whether the following functions are kernels or not. If they are kernels, write down the corresponding feature space. If they are not, provide a proof. In both cases, x and z are real d-dimensional vectors.

(1) (5 points) $K(x, z) = \|x\|^2 + \|z\|^2$. Not a kernel. To show this, we pick $x = [0]$, $z = [2]$. The kernel matrix is:

$$M = \begin{bmatrix} 0 & 4 \\ 4 & 4 \end{bmatrix}$$

For a column vector $u = [u_1, u_2]$, $u^\top Mu = 8u_1u_2 + 4u_2^2 < 0$ for $u_2 = 1$ and $u_1 = -1$. Thus the kernel matrix is not positive semidefinite and hence K is not a kernel.

(2) (5 points) $K(x, z) = e^{\|x\|^2 + \|z\|^2}$.

$K(x, z)$ is a kernel, with feature space $\phi(x) = [e^{\|x\|^2}]$.