

# CS189: Introduction to Machine Learning

## Homework 2

Due: September 24, 2015 at 11:59pm

### Instructions:

- Homework 2 is completely a written assignment, no coding involved.
- Please write (legibly!) or typeset your answers in the space provided. If you choose to typeset your answers, please use this template file ([hw2.tex](#)), provided on bCourses announcement page. If there is not enough space for your answer, you can continue your answer on a separate page (for example : You might want to append pages in Questions 6,7,8).
- Submit a pdf of your answers to <https://gradescope.com> under Homework 2. A photograph or scanned copy is acceptable as long as it is clear with good contrast. You should be able to see CS 189/289 on gradescope when you login with your primary e-mail address used in bCourses. Please let us know if you have any problems accessing gradescope.
- While submitting to Gradescope, you will have to select the region containing your answer for each of the question. Thus, write the answer to a question (or given part of the question) at one place only.
- Start early and don't wait until last minute to submit the assignment as the submission procedure might take sometime too.

### About the Assignment:

- This assignment tries to refresh the concepts of probability, linear algebra and matrix calculus.
- Questions 1 to 6 are dedicated to deriving fundamental results related to these concepts. You might want to refer your math class textbooks for help.
- Questions 7,8 discuss few applications of these concepts in machine learning.
- Hope you will enjoy doing the assignment !

**Homework Party : Sept 21, 2-4pm in the Wozniak Lounge, SODA 430**

**Problem 1.** A target is made of 3 concentric circles of radii  $1/\sqrt{3}$ , 1 and  $\sqrt{3}$  feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let  $X$  be the distance of the hit from the center (in feet), and let the p.d.f of  $X$  be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot?

**Solution:**

**Problem 2.** Assume that the random variable  $X$  has the exponential distribution

$$f(x|\theta) = \theta e^{-\theta x} \quad x > 0, \theta > 0$$

where  $\theta$  is the parameter of the distribution. Use the method of maximum likelihood to estimate  $\theta$  if 5 observations of  $X$  are  $x_1 = 0.9$ ,  $x_2 = 1.7$ ,  $x_3 = 0.4$ ,  $x_4 = 0.3$ , and  $x_5 = 2.4$ , generated i.i.d. (i.e., independent and identically distributed).

**Solution:**

**Problem 3.** The polynomial kernel is defined to be

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^d$$

where  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , and  $c \geq 0$ . When we take  $d = 2$ , this kernel is called the quadratic kernel.

- (a) Find the feature mapping  $\Phi(\mathbf{z})$  that corresponds to the quadratic kernel.
- (b) How do we find the optimal value of  $d$  for a given dataset?

**Solution:**

**Def:** Let  $A \in \mathbb{R}^{n \times n}$  be a symmetric matrix. We say that  $A$  is positive definite if  $\forall x \in \mathbb{R}^n$ ,  $x^\top Ax > 0$ . Similarly, we say that  $A$  is positive semidefinite if  $\forall x \in \mathbb{R}^n$ ,  $x^\top Ax \geq 0$ .

**Problem 4.** Let  $x = [x_1 \ \cdots \ x_n]^\top \in \mathbb{R}^n$ , and let  $A \in \mathbb{R}^{n \times n}$  be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

- (a) Give an explicit formula for  $x^\top Ax$ . Write your answer as a sum involving the elements of  $A$  and  $x$ .
- (b) Show that if  $A$  is positive definite, then the entries on the diagonal of  $A$  are positive (that is,  $a_{ii} > 0$  for all  $1 \leq i \leq n$ ).

**Solution:**

**Problem 5.** Let  $B$  be a positive semidefinite matrix. Show that  $B + \gamma I$  is positive definite for any  $\gamma > 0$ .

**Solution:**

**Problem 6 : Derivatives and Norms.** Derive the expression for following questions.  
Do not write the answers directly.

- (a) Let  $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$ . Derive  $\frac{\partial(\mathbf{x}^T \mathbf{a})}{\partial \mathbf{x}}$ .
- (b) Let  $\mathbf{A}$  be a  $n \times n$  matrix and  $\mathbf{x}$  be a vector in  $\mathbb{R}^n$ . Derive  $\frac{\partial(\mathbf{x}^T \mathbf{A} \mathbf{x})}{\partial \mathbf{x}}$ .
- (c) Let  $\mathbf{A}, \mathbf{X}$  be  $n \times n$  matrices. Derive  $\frac{\partial \text{Trace}(\mathbf{X} \mathbf{A})}{\partial \mathbf{X}}$ .
- (d) Let  $\mathbf{X}$  be a  $m \times n$  matrix,  $\mathbf{a} \in \mathbb{R}^m$  and  $\mathbf{b} \in \mathbb{R}^n$ . Derive  $\frac{\partial(\mathbf{a}^T \mathbf{X} \mathbf{b})}{\partial \mathbf{X}}$ .
- (e) Let  $\mathbf{x} \in \mathbb{R}^n$ . Prove that  $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n} \|\mathbf{x}\|_2$ . Here  $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$  and  $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ .

**Solution:**

**Problem 7 : Application of Matrix Derivatives.**

Let  $\mathbf{X}$  be a  $n \times d$  data matrix,  $\mathbf{Y}$  be the corresponding  $n \times 1$  target/label matrix and  $\mathbf{\Lambda}$  be the diagonal  $n \times n$  matrix containing weight of each example. Expanding them, we

have  $\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(n)})^T \end{bmatrix}$ ,  $\mathbf{Y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$  and  $\mathbf{\Lambda} = \text{diag}(\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(n)})$

where  $\mathbf{x}^{(i)} \in \mathbb{R}^d$ ,  $y^{(i)} \in \mathbb{R}$ , and  $\lambda^{(i)} > 0 \quad \forall \quad i \in \{1 \dots n\}$ .  $\mathbf{X}$ ,  $\mathbf{Y}$  and  $\mathbf{\Lambda}$  are fixed and known.

In the remaining parts of this question, we will try to fit a weighted linear regression model for this data. We want to find the value of weight vector  $\mathbf{w}$  which best satisfies the following equation  $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$  where  $\epsilon$  is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk function is defined as follows:

$$\begin{aligned} R[\mathbf{w}] &= \sum_{i=1}^n \lambda^{(i)} (\epsilon^{(i)})^2 \\ &= \sum_{i=1}^n \lambda^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2 \end{aligned}$$

- Write this risk function  $R[\mathbf{w}]$  in matrix notation, i.e., in terms of  $\mathbf{X}$ ,  $\mathbf{Y}$ ,  $\mathbf{\Lambda}$  and  $\mathbf{w}$ .
- Find the value of  $\mathbf{w}$ , in matrix notation, that minimizes the risk function obtained in Part (a). You can assume that  $\mathbf{X}^T \mathbf{\Lambda} \mathbf{X}$  is full rank matrix. Hint: You can use the expression derived in Q-6(b).
- What will be the answer for questions in Parts (a) and (b) if you add  $L_2$  regularization (i.e., shrinkage) on  $\mathbf{w}$ ? The  $L_2$  regularized risk function, for  $\gamma > 0$ , is

$$R[\mathbf{w}] = \sum_{i=1}^n \lambda^{(i)} (\mathbf{w}^T \mathbf{x}^{(i)} - y^{(i)})^2 + \gamma \|\mathbf{w}\|_2^2$$

Hint: You can make use of the result in Q-5.

- What role does the regularization (i.e., shrinkage) play in fitting the regression model and how? You can observe the difference in expressions for  $\mathbf{w}$  obtained in Parts (c) and (d), and argue.

**Solution:**



**Problem 8: Classification.** Suppose we have a classification problem with classes labeled  $1, \dots, c$  and an additional doubt category labeled as  $c + 1$ . Let the loss function be the following:

$$\ell(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \dots, c\} \\ \lambda_r & \text{if } i = c + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where  $\lambda_r$  is the loss incurred for choosing doubt and  $\lambda_s$  is the loss incurred for making a misclassification. Note that  $\lambda_r \geq 0$  and  $\lambda_s \geq 0$ .

Hint : The risk of classifying a new datapoint as class  $i \in \{1, 2, \dots, c + 1\}$  is

$$R(\alpha_i|x) = \sum_{j=1}^{j=c} \ell(f(x) = i, y = j)P(\omega_j|x)$$

- (a) Show that the minimum risk is obtained if we follow this policy: (1) choose class  $i$  if  $P(\omega_i|x) \geq P(\omega_j|x)$  for all  $j$  and  $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$ , and (2) choose doubt otherwise.
- (b) What happens if  $\lambda_r = 0$ ? What happens if  $\lambda_r > \lambda_s$ ?

**Solution:**