Name: Jesus Avila **Student ID:** 24084724

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:

This is a discrete random variable. We can use the probability-weighted average to estimate the expected value.

$$E(X) = x_1 p_1 + x_2 p_2 + x_3 p_3 + x_4 p_4 , \quad \text{for } x_1 = 4,\ x_2 = 3,\ x_3 = 2,\ x_1 = 0$$

We can estimate $p_1, p_2,$ & p_3 using the p.d.f of X, and we'll get p_4 using $p_4 = 1-(p_1+p_2+p_3)$

$$p_i = \int_a^b f(x)\,dx = \int_a^b \frac{2}{\pi(1+x^2)}\,dx = \frac{2}{\pi}\tan^{-1}(x)\Big|_a^b = \frac{2}{\pi}\left[\tan^{-1}(b) - \tan^{-1}(a)\right]$$

$$p_1 = \frac{2}{\pi}\left[\tan^{-1}\left(\frac{1}{\sqrt{3}}\right) - \tan^{-1}(0)\right] = \frac{2}{\pi}\left[\frac{\pi}{6} - 0\right] = \frac{2}{\pi}\left(\frac{\pi}{6}\right) = \underline{2\left(\frac{1}{6}\right)}$$

$$p_2 = \frac{2}{\pi}\left\{\tan^{-1}(1) - \tan^{-1}\left(\frac{1}{\sqrt{3}}\right)\right\} = \frac{2}{\pi}\left(\frac{\pi}{4} - \frac{\pi}{6}\right) = 2\left(\frac{1}{4} - \frac{1}{6}\right) = 2\left(\frac{3}{12} - \frac{2}{12}\right) = \underline{2\left(\frac{1}{12}\right)}$$

$$p_3 = \frac{2}{\pi}\left\{\tan^{-1}(\sqrt{3}) - \tan^{-1}(1)\right\} = \frac{2}{\pi}\left(\frac{\pi}{3} - \frac{\pi}{4}\right) = 2\left(\frac{4}{12} - \frac{3}{12}\right) = \underline{2\left(\frac{1}{12}\right)}$$

$$p_4 = 1 - 2\left(\frac{2}{12} + \frac{1}{12} + \frac{1}{12}\right) = 1 - 2\left(\frac{1}{3}\right) = \frac{1}{3} = \underline{2\left(\frac{1}{6}\right)}$$

$$\Rightarrow E(x) = 2\left[4\left(\frac{1}{6}\right) + 3\left(\frac{1}{12}\right) + 2\left(\frac{1}{12}\right) + 0\left(\frac{1}{6}\right)\right] = 2\left(\frac{8+3+2}{12}\right) = \underline{\frac{13}{6}}$$

$$\Rightarrow \boxed{E(x) = \frac{13}{6}}$$

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

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

where θ is the parameter of the distribution. Use the method of maximum likelihood to estimate θ 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:

$$\mathcal{L}(\theta; x_1, \dots x_5) = f(x_1, \dots x_5 | \theta) = \prod_{i=1}^{5} f(x_i | \theta) \qquad \text{(from wikipedia)}$$

$$\Rightarrow \ln \mathcal{L} = \sum_{i=1}^{5} \ln f(x_i | \theta)$$

we can take $\hat{\ell} = \frac{1}{n} \ln \mathcal{L}$ as the average log-likelihood.

Now we find $\hat{\theta}$ that maximizes $\hat{\ell}$.

$$\hat{\ell} = \frac{1}{5} \left[\sum_{i=1}^{5} \ln \left(\hat{\theta} e^{-\hat{\theta} x_i} \right) \right] = \frac{1}{5} \left[\ln(\hat{\theta} e^{-\hat{\theta} x_1}) + \dots \ln(\hat{\theta} e^{-\hat{\theta} x_5}) \right]$$

$$\hat{\ell} = \frac{1}{5} \left[5 \ln \hat{\theta} - \hat{\theta} \left(x_1 + \dots + x_5 \right) \right]$$

Now we take derivitive to find maximum

$$\frac{d\hat{\ell}}{d\hat{\theta}} = \frac{1}{5} \left[\frac{5}{\hat{\theta}} - (x_1 + \dots + x_5) \right] = 0$$

$$\Rightarrow \frac{5}{\hat{\theta}} = x_1 + \dots + x_5 \Rightarrow \boxed{\hat{\theta} = \frac{5}{x_1 + \dots + x_5}} = \frac{5}{0.9 + 1.7 + 0.4 + 0.3 + 2.4}$$

$$\boxed{\hat{\theta} = 0.9}$$

Problem 3. The polynomial kernel is defined to be

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

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathbf{n}}$, 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:

(a) $k(x,y) = \Phi(x) \cdot \Phi(y)$

$k(x,y) = (x^T y + c)^2 = \left(\sum_{i=1}^{n} x_i y_i + c\right)^2$

using the multinomial theorem for $d = 2$ we get

$$\sum_{i=1} x_i^2 y_i^2 + \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\sqrt{2}\, x_i x_j)(\sqrt{2}\, y_i y_j) + \sum_{i=1}^{n} (\sqrt{2c}\, x_i)(\sqrt{2c}\, y_i) + c^2$$

$\Rightarrow$ $\Phi(z) = \Phi(x) \cdot \Phi(y)$, where

$\Phi(x) = [x_n^2, \ldots, x_1^2, \sqrt{2}\, x_n x_{n-1}, \ldots, \sqrt{2}\, x_n x_1, \sqrt{2}\, x_{n-1} x_{n-2}, \ldots, \sqrt{2}\, x_{n-1} x_1, \ldots \sqrt{2}\, x_2 x_1, \sqrt{2c}\, x_n, \ldots \sqrt{2c}\, x_1, c]$

$\Phi(y) = [y_n^2, \ldots, y_1^2, \sqrt{2}\, y_n y_{n-1}, \ldots, \sqrt{2}\, y_n y_1, \sqrt{2}\, y_{n-1} y_{n-2}, \ldots, \sqrt{2}\, y_{n-1} y_1, \ldots \sqrt{2}\, y_2 y_1, \sqrt{2c}\, y_n, \ldots \sqrt{2c}\, x_1, c]$

(b) We can optimize d by performing gradient-descent or the Nelder-Mead simplex algorithm.

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

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

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

(a) Give an explicit formula for $x^\top A x$. 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:

(a) $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $\overset{1 \times n}{x^\top = [x_1 \cdots x_n]}$

$$[\overset{1 \times n}{x^\top} \overset{n \times n}{A}]x = \left[(x_1 a_{11} + \ldots + x_n a_{n1}) \quad (x_1 a_{12} + \ldots + x_n a_{n2}) \quad \cdots \quad (x_1 a_{1n} + \ldots + x_n a_{nn})\right] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \left[\sum_{i=1}^{n} x_i a_{i1} \quad \sum_{i=1}^{n} x_i a_{i2} \quad \cdots \quad \sum_{i=1}^{n} x_i a_{in}\right] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= x_1 \sum_{i=1}^{n} x_i a_{i1} + x_2 \sum_{i=1}^{n} x_i a_{i2} + \ldots + x_n \sum_{i=1}^{n} x_i a_{in}$$

$$\boxed{x^\top A x = \sum_{j=1}^{n} \left(x_j \sum_{i=1}^{n} x_i a_{ij} \right)}$$

(b) Since A is positive definite, then $x^\top A x > 0$ for any non-zero vector $x \in \mathbb{R}^n$. Let's se what happens when we choose x to be the unit vector.

$$x^\top A x = x_1 \sum_{i=1}^{n} x_i a_{i1} + x_2 \sum_{i=1}^{n} x_i a_{i2} + \ldots + x_n \sum_{i=1}^{n} x_i a_{in}$$

$$= x_1 (x_1 a_{11} + x_2 a_{21} + \ldots + x_n a_{n1}) + \ldots + x_n (x_1 a_{1n} + x_2 a_{2n} + \ldots x_n a_{nn})$$

Now, for $x = e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ $\Rightarrow$ Everything will be zero except for a_{11}

$\Rightarrow \quad 0 < e_1^\top A e_1 = a_{11}$

This same thing happens for all e_i, where $i = 1 \ldots n$

$\Rightarrow \quad 0 < e_i^\top A e_i = a_{ii} \quad \Rightarrow a_{ii}$ must be positive

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

Solution:

Since B PSM, $\forall x \in \mathbb{R}^n$, $x^T B x \geq 0$

$$B + \gamma I = \begin{bmatrix} b_{11}+\gamma & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22}+\gamma & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & \cdots & \cdots & b_{nn}+\gamma \end{bmatrix}$$

$$x^T(B+\gamma I)x = x_1\left[x_1(b_{22}+\gamma) + x_2 b_{21} + \ldots + x_n b_{n1}\right] + \ldots + x_n\left[x_1 b_{1n} + x_2 b_{2n} + \ldots + x_n(b_{nn}+\gamma)\right]$$

$$= x^T B x + x_1^2\gamma + x_2^2\gamma + \ldots + x_n^2\gamma$$

$$\boxed{= \underbrace{x^T B x}_{\geq 0} + \gamma \underbrace{\left(x_1^2 + x_2^2 + \ldots + x_n^2\right)}_{\text{This will always be positive b/c it's squared}}}$$

$\Rightarrow$ as long as $\gamma > 0$, then $B + \gamma I$ is a positive def. matrix

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 $\frac{\partial \text{Trace}(\mathbf{XA})}{\partial \mathbf{X}}$.

(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 $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$. Here $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.

Solution:

(a) Let $y = x^T a = [x_1 \dots x_n]\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = x_1 a_1 + x_2 a_2 + \dots + x_n a_n$

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \Rightarrow \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \boxed{a}$$

(b) Let $y = x^T A x = x_1 \sum_{i=1}^n x_i a_{i1} + x_2 \sum_{i=1}^n x_i a_{i2} + \dots + x_n \sum_{i=1}^n x_i a_{in}$

$$= x_1(x_1 a_{11} + x_2 a_{21} + \dots x_n a_{n1}) + x_2(x_1 a_{12} + x_2 a_{22} + \dots + x_n a_{n2}) + \dots + x_n(x_1 a_{1n} + x_2 a_{2n} + \dots + x_n a_{nn})$$

$$= (a_{11} x_1^2 + a_{21} x_2 x_1 + \dots + a_{n1} x_n x_1) + (a_{12} x_1 x_2 + a_{22} x_2^2 + \dots + a_{n2} x_n x_2) + \dots + (a_{1n} x_1 x_n + a_{2n} x_2 x_n + \dots + a_{nn} x_n^2)$$

$$\frac{dy}{dx} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial y}{\partial x_1} = (2 a_{11} x_1 + a_{21} x_2 + \dots + a_{n1} x_n) + a_{12} x_2 + \dots + a_{1n} x_n$$

$$= 2 a_{11} x_1 + (a_{21} + a_{12}) x_2 + \dots + (a_{n1} + a_{1n}) x_n = \underline{[2a_{11} \quad (a_{21}+a_{12}) \cdots (a_{n1}+a_{1n})]\, x}$$

similarly,

$$\frac{\partial y}{\partial x_2} = [(a_{12}+a_{21}) \quad 2a_{22} \quad \cdots \quad (a_{n2}+a_{2n})]\, x$$

$$\Rightarrow \frac{dy}{dx} = \begin{bmatrix} [2a_{11} \quad (a_{21}+a_{12}) \cdots (a_{n1}+a_{1n})]\, x \\ [(a_{12}+a_{21}) \quad 2a_{22} \quad \cdots (a_{n2}+a_{2n})]\, x \\ \vdots \\ [(a_{1n}+a_{n1}) \quad (a_{2n}+a_{n2}) \quad \dots 2a_{nn}]\, x \end{bmatrix} = \begin{bmatrix} 2a_{11} & (a_{21}+a_{12}) & \cdots & (a_{n1}+a_{1n}) \\ (a_{12}+a_{21}) & 2a_{22} & \cdots & (a_{n2}+a_{2n}) \\ & \vdots & & \\ (a_{1n}+a_{n1}) & (a_{2n}+a_{n2}) & \cdots & 2a_{nn} \end{bmatrix} x$$

$$\boxed{\frac{dy}{dx} = (A + A^T)x}$$

Problem 6 Continued

(c) $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & \ddots & & \vdots \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$, $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & & \vdots \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$

$$XA = \begin{bmatrix} (x_{11}a_{11} + x_{12}a_{21} + \dots + x_{1n}a_{n1}) & \cdots & & \vdots \\ \vdots & (x_{21}a_{12} + x_{22}a_{22} + \dots + x_{2n}a_{n2}) & & \vdots \\ \vdots & \vdots & \ddots & \\ \cdots & \vdots & & (x_{n1}a_{1n} + x_{n2}a_{2n} + \dots + x_{nn}a_{nn}) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{n} x_{1i}a_{i1} & \cdots & \vdots \\ \vdots & \sum_{i=1}^{n} x_{2i}a_{i2} & \vdots \\ \vdots & \ddots & \sum_{i=1}^{n} x_{ni}a_{in} \end{bmatrix}$$

$$\text{Trace}(XA) = \sum_{i=1}^{n} x_{1i}a_{i1} + \sum_{i=1}^{n} x_{2i}a_{i2} + \dots + \sum_{i=1}^{n} x_{ni}a_{in}$$

Now, Let $y = \text{Trace}(XA) \Rightarrow \frac{dy}{dX} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & & & \vdots \\ \frac{\partial y}{\partial x_{n1}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nn}} \end{bmatrix}$

$\frac{\partial y}{\partial x_{11}} = a_{11} \Rightarrow$ Similar for all $\frac{\partial y}{\partial x_{ij}}$

$$\Rightarrow \boxed{\frac{\partial(\text{Trace}(XA))}{\partial X} = A}$$

$y = a^T A b$

(d) Let $\underset{m\times n}{X} = \underset{m\times n}{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ A_{m1} & \cdots & & A_{mn} \end{bmatrix}$, $a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$, $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$

(wrote A instead of X, will change at end)

$$\text{Let } y = a^T A b = \underset{1\times n}{[a^T A_{i1} \;\; a^T A_{i2} \cdots a^T A_{in}]}\,\underset{n\times 1}{b}$$

$$= b_1 a^T A_{i1} + b_2 a^T A_{i2} + \dots + b_n a^T A_{in}$$

$$= b_1(a_1 A_{11} + a_2 A_{21} + \dots a_m A_{m1}) + b_2(a_1 A_{12} + a_2 A_{22} + \dots + a_m A_{m2}) + b_n(a_1 A_{1n} + a_2 A_{2n} + \dots + a_m A_{mn})$$

$\frac{\partial y}{\partial A} = \begin{bmatrix} \frac{\partial y}{\partial A_{11}} & \frac{\partial y}{\partial A_{12}} & \cdots & \frac{\partial y}{\partial A_{1n}} \\ \vdots & \ddots & & \vdots \\ \frac{\partial y}{\partial A_{m1}} & \cdots & & \frac{\partial y}{\partial A_{mn}} \end{bmatrix}$, $\frac{\partial y}{\partial A_{11}} = a_1 b_1$, $\frac{\partial y}{\partial A_{12}} = a_1 b_2$, $\frac{\partial y}{\partial A_{21}} = a_2 b_1$

$$\Rightarrow \frac{\partial y}{\partial A_{ij}} = a_i b_j$$

$$\Rightarrow \frac{\partial y}{\partial A} = \underset{m\times n}{\begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & & & \vdots \\ \vdots & & & \\ a_mb_1 & \cdots & & a_mb_n \end{bmatrix}} = \underset{m\times 1}{a}\,\underset{1\times n}{b^T}$$

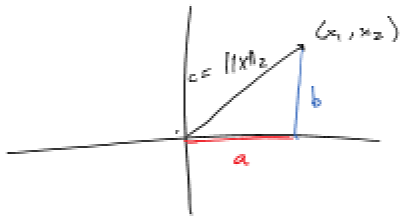
Now switching back to $X = A$

$$\Rightarrow \boxed{\frac{\partial(a^T X b)}{\partial X} = a b^T}$$

Prove

(e) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$ $\quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $\quad \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots x_n^2}$

$\|x\|_1 = |x_1| + |x_2| + \dots + |x_n|$



$\|x\|_2 = \sqrt{x_1^2 + x_2^2}$

= distance from origin to point (x_1, x_2), or the hypotenuse of our triangle.

Now, $\|x\|_1$ is just the sum of the sides of the triangle except the hypotenuse.

Now, let's work on first inequality: $\|x\|_2 \leq \|x\|_1$

Since $\|x\|_2$ is the hypotenuse, it can never be larger or less than the sum of its sides. This is the triangle inequality: $c \leq a + b$.

$$\Rightarrow \|x\|_2 \leq \sqrt{x_1^2} + \sqrt{x_2^2}$$

$$\|x\|_2 \leq |x_1| + |x_2|$$

$$\underline{\|x\|_2 \leq \|x\|_1}$$

$$|\vec{x} \cdot \vec{y}| \leq \|\vec{x}\|_2 \|\vec{y}\|_2$$

$$|\vec{x} \cdot \vec{y}| = \|\vec{x}\|_2 \|\vec{y}\|_2 \text{ only when } \vec{x} = c\vec{y}$$

Proof: $\|x\|_1 \leq \sqrt{n}\,\|x\|_2$

$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$|x \cdot y| \leq \|x\|_2 \|y\|_2$ $\quad , \quad u = \dfrac{x}{\|x\|_2}$

$|x^T y| \leq \|x\|_2 \|y\|_2$

$|x_1 y_1 + \dots + x_n y_n| \leq \|x\|_2 \|y\|_n$

$x i = x$

$|x \cdot i| \leq \|x\|_2 \|i\|$

$|x^T i| \leq \sqrt{n}\,\|x\|_2$

$|x_1 + \dots + x_n| \leq \sqrt{n}\,\|x\|_2$

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 $\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(n)})^T \end{bmatrix}$, $\mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(n)} \end{bmatrix}$ and $\mathbf{\Lambda} = \text{diag}(\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(n)})$
where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $\mathbf{y}^{(i)} \in \mathbb{R}$, and $\lambda^{(i)} > 0 \quad \forall \;\; i \in \{1 \ldots n\}$. $\mathbf{X}$, $\mathbf{Y}$ 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 w which best satisfies the following equation $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$ where ϵ is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk function is defined as follows:

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

(a) Write this risk function $R[\mathbf{w}]$ in matrix notation, i.e., in terms of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Lambda}$ and $\mathbf{w}$.

(b) 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).

(c) 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 L2 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.

(d) 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 7 Continued

(a) First we find an expression for our ϵ in vector form; that is, $\epsilon \in \mathbb{R}^n$. Since ϵ is composed of X, Y, w, then we need to find an expression for $w^T x^{(i)}$. Y is already in the correct form.

$\Rightarrow$ X is $n \times d$, and w is $d \times 1$

$\Rightarrow$ Xw will yield an $n \times 1$ vector.

$$\Rightarrow \quad \epsilon = Xw - Y = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix}$$

- Next, the elements are scaled by $\lambda^{(i)}$. Our goal is to get the following vector: $\begin{bmatrix} \lambda^{(1)} \epsilon^{(1)} \\ \vdots \\ \lambda^{(n)} \epsilon^{(n)} \end{bmatrix}$

Now, Λ is $n \times n$ and ϵ is $n \times 1$. This means we can do the multiplication $\Lambda \epsilon$. This yields our goal vector since Λ is a diagonal:

$$\begin{bmatrix} \lambda^{(1)} & 0 & \cdots & 0 \\ \vdots & \lambda^{(2)} & & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda^{(n)} \end{bmatrix} \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(n)} \end{bmatrix} = \begin{bmatrix} \lambda^{(1)} \epsilon^{(1)} \\ \vdots \\ \lambda^{(n)} \epsilon^{(n)} \end{bmatrix}$$

- Lastly, our goal is to get the following:

$$\lambda^{(1)}(\epsilon^{(1)})^2 + \dots + \lambda^{(n)}(\epsilon^{n})^2$$

We arrive at this by simply multiplying by ϵ^T:

$$\Rightarrow \quad \lambda^{(1)}(\epsilon^{(1)})^2 + \dots + \lambda^{(n)}(\epsilon^{n})^2 = \epsilon^T \Lambda \epsilon$$

$$\Rightarrow \quad \boxed{R[w] = \sum_{n=1}^{n} \lambda^{(i)} \left(\epsilon^{(i)}\right)^2 = (Xw - Y)^T \Lambda (Xw - Y)}$$

(b) We can find the minimum by taking the derivative of R and setting it equal to zero. This w will give us the minimum of R if R is convex. Our optimization function has the form $f(x) = x^T A x$, which is convex if A is positive semidefinite. Here, $A = \Lambda$. Λ is a diagonal $\Rightarrow$ symmetric matrix.

Let $z(w) = Xw - Y$

$$\Rightarrow R = z^T \Lambda z.$$

$$\Rightarrow \frac{\partial R}{\partial w} = \frac{\partial z}{\partial w}\frac{\partial R}{\partial z}$$

From 6(b), $\frac{\partial R}{\partial z} = (\Lambda + \Lambda^T)z = 2\Lambda z$, since Λ is a diagonal matrix.

$$\overset{d \times n}{\frac{\partial z}{\partial w}} = \begin{bmatrix} \frac{\partial z_1}{\partial w_1} & \frac{\partial z_2}{\partial w_1} & \cdots & \frac{\partial z_n}{\partial w_1} \\ \frac{\partial z_1}{\partial w_2} & \frac{\partial z_2}{\partial w_2} & \cdots & \frac{\partial z_n}{\partial w_2} \\ & \vdots & & \\ \frac{\partial z_1}{\partial w_d} & \frac{\partial z_2}{\partial w_d} & \cdots & \frac{\partial z_n}{\partial w_d} \end{bmatrix} \Rightarrow z = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ X_{21} & & & X_{2d} \\ \vdots & & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nd} \end{bmatrix}\begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \begin{bmatrix} X_{11}w_1 + \dots + X_{1d}w_d \\ \vdots \\ X_{n1}w_1 + \dots + X_{nd}w_d \end{bmatrix} - \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\frac{\partial z_1}{\partial w_1} = X_{11}, \quad \frac{\partial z_2}{\partial w_1} = X_{21}$$

$$\frac{\partial z_1}{\partial w_2} = X_{12}$$

$$\Rightarrow \frac{\partial z}{\partial w} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & & \\ \vdots & & \ddots & \\ X_{1d} & \cdots & & X_{nd} \end{bmatrix} = \underline{X^T}$$

$$\Rightarrow \frac{\partial R}{\partial w} = \left(\frac{\partial z}{\partial w}\right)\left(\frac{\partial R}{\partial z}\right) = \left(X^T\right)\left(2\Lambda z\right) = 2X^T\Lambda(Xw - Y) = 0$$

$$\Rightarrow \overset{d \times n}{X^T}\overset{n \times n}{\Lambda}\overset{n \times 1}{(Xw - Y)} = 0 \Rightarrow X^T\Lambda X w - X^T \Lambda Y = 0$$

$$\Rightarrow X^T \Lambda X w = X^T \Lambda Y$$, Assume that $X^T\Lambda X$ is a full-rank matrix

$\Rightarrow$ the square matrix $X^T\Lambda X$ is invertible

$$\Rightarrow \boxed{w = (X^T \Lambda X)^{-1} X^T \Lambda Y.}$$

c) For a), we had $\lambda^{(1)}(\epsilon^{(1)})^2 + \ldots + \lambda^{(n)}(\epsilon^{(n)})^2 = \epsilon^T \Lambda \epsilon$

Now $R = \lambda^{(1)}(\epsilon^{(1)})^2 + \ldots + \lambda^{(n)}(\epsilon^{(n)})^2 + \gamma[(w^{(1)})^2 + \ldots + (w^{(d)})^2]$

Lets put $\gamma \|w\|_2^2$ in matrix form:

$$\gamma \|w\|_2^2 = \gamma[(w^{(1)})^2 + \ldots + (w^{(d)})^2] = \gamma \sum_{i=1}^{d} (w^{(i)})^2$$

Previously, we saw that $\sum \lambda^{(i)} \epsilon^{(i)})^2 = \epsilon^T \Lambda \epsilon$

$$\Rightarrow \gamma \|w\|_2^2 = \gamma w^T w$$

$$\Rightarrow R = (X_w - Y)^T \Lambda (X_w - Y) + \gamma w^T w$$

$$\Rightarrow \boxed{R = (X_w - Y)^T \Lambda (X_w - Y) + \gamma w^T I_d w}$$

For b), we also have to prove that R is convex.

Since the first term is convex, we need only prove that the second is also convex, since the sum of two convex functions yields a convex function.

Like before, $w^T I_d w$ is convex if I_d is semidefinite. This is the identity matrix, which means it's in rref with all positive values for its diagonal $\Rightarrow$ this function is convex $\Rightarrow$ R is convex.

Let $R = P + Q$, $P = (X_w - Y)^T \Lambda (X_w - Y)$, $Q = \gamma w^T I_d w$

$\Rightarrow \frac{\partial R}{\partial w} = \frac{\partial P}{\partial w} + \frac{\partial Q}{\partial w}$. $\frac{\partial P}{\partial w} = 2X^T \Lambda (X_w - Y)$, from b)

$$\frac{\partial Q}{\partial w} = \begin{bmatrix} \frac{\partial Q}{\partial w_1} \\ \vdots \\ \frac{\partial Q}{\partial w_d} \end{bmatrix}$$, since $\gamma w^T I_d w$ is a scalar

$$\Rightarrow Q = \gamma[w_1^2 + w_2^2 + \ldots + w_d^2]$$

$$\Rightarrow \frac{dQ}{dw_i} = \gamma(2w_i) = 2\gamma w_i$$

$$\Rightarrow \frac{\partial Q}{\partial w} = \begin{bmatrix} 2\gamma w_1 \\ \vdots \\ 2\gamma w_d \end{bmatrix} = 2\gamma w$$

$$\Rightarrow \frac{\partial R}{\partial w} = 2X^T \Lambda (Xw - Y) + 2\gamma w = 0$$

$$\Rightarrow X^T \Lambda X w - X^T \Lambda Y + \gamma w = 0$$

$$\Rightarrow X^T \Lambda X w + \gamma w = X^T \Lambda Y$$

$$\Rightarrow (X^T \Lambda X + \gamma) w = X^T \Lambda Y.$$

since $X^T \Lambda X$ is assumed to be full-rank, adding a positive constant will still keep its full-rank property

$$\Rightarrow \boxed{w = (X^T \Lambda X + \gamma I_d)^{-1} X^T \Lambda Y}$$

$\Rightarrow X^T \Lambda X + \gamma$ is invertible for $\gamma > 0$

d) For b) we got $w_1 = (X \Lambda X)\ X \Lambda$

both are $n \times 1$

For c) we got $w_2 = (X^T \Lambda X + \gamma I_d)^{-1} X^T \Lambda Y$

Let $w_0 = X^T \Lambda Y$ and $D = X^T \Lambda X$, which is a $d \times d$ matrix

$$\Rightarrow w_1 = D^{-1} w_0$$

$$w_2 = (D + \gamma I_d)^{-1} w_0$$

$\Rightarrow$ The main difference between w_1 & w_2 is how much we'll scall w_0. As $\gamma \uparrow$, the second term γI_d dominates the invertible portion. This means that $w_2 \simeq (\gamma I_d)^{-1} w_0 = \frac{w_0}{\gamma}$.

$\Rightarrow$ We scale w_0 by $\frac{1}{\gamma}$, effectively reducing the complexity of our model.

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

classes: 1, 2... c
doubt cat.: 1+c

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

where λ_r is the loss incurred for choosing doubt and λ_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 $i \in \{1, 2, \ldots, c+1\}$ is

Risk of classifying x as class i

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

Probab. of classifying x to class j w/ weight ω_j

(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:

$$R(\alpha_j|x) = \ell(\cdot\cdot)P(\omega_1|x) + \ldots + \ell(\cdots)P(\omega_c|x)$$

Suppose we want to find risk for classifying x as class 1:

$$R(\alpha_1|x) = 0 + \lambda_s [P(\omega_2|x) + \ldots + P(\omega_c|x)]$$

$\Rightarrow$ It's just the prob. of class. point x as any other class.

(1) choose class i if the prob. of classifying point x as i is greater than the prob. of classifying x as any of the other probabilities.