John Semerdjian

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:

$$g(x) = \begin{cases} 4 & 0 \le x < \frac{1}{\sqrt{3}} \\ 3 & \frac{1}{\sqrt{3}} \le x < 1 \\ 2 & 1 \le x < \sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$

$$E[X] = \int g(x) f(x) dx$$

$$= \int_0^{\frac{1}{\sqrt{3}}} 4 \cdot \frac{2}{\pi(1+x^2)} dx + \int_{\frac{1}{\sqrt{3}}}^{1} 3 \cdot \frac{2}{\pi(1+x^2)} dx + \int_1^{\sqrt{3}} 2 \cdot \frac{2}{\pi(1+x^2)} dx$$

$$= \frac{8}{\pi}\left(\tan^{-1}\frac{1}{\sqrt{3}} - \tan^{-1} 0\right) + \frac{6}{\pi}\left(\tan^{-1} 1 - \tan^{-1}\frac{1}{\sqrt{3}}\right) + \frac{4}{\pi}\left(\tan^{-1}\sqrt{3} - \tan^{-1} 1\right)$$

$$= \frac{8}{\pi}\left(\frac{\pi}{6} - 0\right) + \frac{6}{\pi}\left(\frac{\pi}{4} - \frac{\pi}{6}\right) + \frac{4}{\pi}\left(\frac{\pi}{3} - \frac{\pi}{4}\right)$$

$$= \frac{13}{6}$$

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:

likelihood function

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{(-\theta \sum_{i=1}^{n} x_i)} = \theta^n e^{-\theta n \bar{x}}$$

Take the derivative of log of likelihood function and set equal to 0

$$\frac{d}{d\theta} \log(L(\theta)) = \frac{d}{d\theta}(n \log(\theta) - \theta n \bar{x})$$

$$= \frac{n}{\theta} - n\bar{x}$$

$$n\bar{x} = \frac{n}{\theta}$$

$$\theta = \frac{1}{\bar{x}} = \frac{1}{\left(\frac{0.9 + 1.7 + 0.4 + 0.3 + 2.4}{5}\right)} = 0.877$$

Problem 3. The polynomial kernel is defined to be

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

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

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

b.) You can find the optimal value of d for a give dataset through cross-validation.

4. **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

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

(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.) $x^T A x$

$$\sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} a_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

$= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & \ddots & \vdots \\ A_{N1} & \cdots & A_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$

$1 \times N$ $\quad N \times N$ $\quad N \times 1$

b.) Using a unit vector such as $\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ whose all values are 0 except one value as one, we can use it as a means of extracting all the diagonal elements of A. We can shift the position of 1 in the unit vector sequentially to achieve this. Since $x^T A x$ is > 0, the diagonal elements would have to be positive or else the definition of the positive definite matrix would not hold.

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

Solution:

$B + \gamma I > 0$ if $\gamma > 0$

$$x^T(B + \gamma I)x = (x^T B + x^T \gamma I)x = x^T B x + x^T \gamma I x$$

$$= \underbrace{x^T B x}_{\geq 0} + \underbrace{\gamma \|x\|_2^2}_{> 0} > 0$$

> 0 since x can't equal 0

since B is P.S.D.

Thus, $B + \gamma I$ is P.D.

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

Solution:

a.) $$\frac{\partial(x^Ta)}{\partial x} = \begin{bmatrix} \frac{\partial(x^Ta)}{\partial x_1} & \frac{\partial(x^Ta)}{\partial x_2} & \cdots & \frac{\partial(x^Ta)}{\partial x_N} \end{bmatrix} = \begin{bmatrix} a_1^T \cdots a_N^T \end{bmatrix} = \boxed{a}$$

b.) $$x^TAx = \begin{bmatrix} x_1 \cdots x_N \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{N1} & \cdots & & a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_1 + a_{21}x_2 + \cdots + a_{N1}x_N & \cdots & a_{1N}x_1 + a_{2N}x_2 + \cdots + a_{NN}x_N \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

$$= \begin{bmatrix} a_{11}x_1^2 + \cdots + a_{NN}x_N^2 + (a_{12}+a_{21})x_1x_2 + \cdots + (a_{1N}+a_{N1})x_1x_N \\ + (a_{23}+a_{32})x_2x_3 + \cdots + (a_{2N}+a_{N2})x_2x_N + \cdots + (a_{N,N-1}+a_{N-1,N})x_Nx_{N-1} \end{bmatrix}$$

Take derivative w.r.t. $\vec{x}$

$$= \frac{\partial(x^TAx)}{\partial x} = \begin{bmatrix} 2a_{11}x_1 + \cdots + (a_{N1}+a_{1N})x_N \\ (a_{12}+a_{21})x_1 + \cdots + (a_{N2}+a_{2N})x_N \\ \vdots \\ (a_{1N}+a_{N1})x_1 + \cdots + 2a_{NN}x_N \end{bmatrix} = \begin{bmatrix} 2a_{11} & (a_{21}+a_{12}) & \cdots \\ (a_{12}+a_{21}) & 2a_{22} & \\ \vdots & & \ddots \\ (a_{1N}+a_{N1}) & \cdots & 2a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

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

6(.) $\text{Trace}(XA) = \text{Trace}\begin{bmatrix} -\vec{x}_1- \\ \vdots \\ -\vec{x}_n- \end{bmatrix}\begin{bmatrix} | & & | \\ \vec{a}_1 & \cdots & \vec{a}_n \\ | & & | \end{bmatrix}$

$$= \text{Trace}\begin{bmatrix} \vec{x}_1^T\vec{a}_1 & \cdots & \vec{x}_1^T\vec{a}_n \\ \vdots & \ddots & \vdots \\ \vec{x}_n^T\vec{a}_1 & \cdots & \vec{x}_n^T\vec{a}_n \end{bmatrix}$$

$$= \sum_{i=1}^{m} x_{1i}a_{i1} + \cdots + \sum_{i=1}^{m} x_{ni}a_{in}$$

$$\frac{\partial \text{Trace}(XA)}{\partial X_{ij}} = a_{ji} \Rightarrow \boxed{A^T}$$

6d.) $\dfrac{\partial(a^TXb)}{\partial X}$ $\quad [a_1 \cdots a_n]\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{mn} \end{bmatrix}\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$

$$\frac{\partial \sum_{j=1}^{n}\sum_{i=1}^{n} a_i x_{ij} b_j}{\partial X} \Rightarrow \begin{bmatrix} a_1b_1 & \cdots & a_1b_N \\ \vdots & \ddots & \vdots \\ a_mb_1 & \cdots & a_mb_N \end{bmatrix} = \boxed{ab^T}$$

e.) Prove $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\,\|x\|_2$

$$\|x\|_2^2 = \sum_{i=1}^{n} x_i^2 \qquad \|x\|_1^2 = \left(\sum_{i=1}^{n} |x_i|\right)^2 = x_1^2 + x_2^2 + 2x_1x_2 \dots 2x_nx_{n-1}$$

$$= \sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \quad i \neq j$$

$$\sqrt{\sum_{i=1}^{n} x_i^2} \leq \sqrt{\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \quad \checkmark$$

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

where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$, 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:

a) $R[w] = \sum_{i=1}^{n} \lambda^{(i)} (w^T x^{(i)} - y^{(i)})^2$

$= (Xw - y)^T \Lambda (Xw - Y)$

7b.) $w = ?$

$$R[w] = (Xw - y)^T \Lambda (Xw - y)$$

$$= (w^T X^T - y^T)\Lambda (Xw - y)$$

$$= (w^T X^T \Lambda - y^T \Lambda)(Xw - y)$$

$$= w^T X^T \Lambda Xw - w^T X^T \Lambda y - y^T \Lambda Xw + y^T \Lambda y$$

Take derivative and set = 0 wrt w

$$0 = \frac{2 w^T X^T \Lambda Xw}{dw} - \frac{2\, 2 x^T \Lambda Xw}{2w} + \frac{2 y^T \Lambda y}{\partial w}$$

$B = x^T A x$

$C = y^T \Lambda x$

$$0 = \frac{2 w^T B w}{2w} - 2 \frac{d C w}{\partial w}$$

$$= (B + B^T) w - 2 c^T$$

$$2 (y^T \Lambda x)^T = \left(x^T A x + (x^T A x)^T\right) w$$

$$2 (y^T \Lambda x)^T = 2 (x^T \Lambda x) w$$

$$\boxed{w = (x^T \Lambda x)^{-1} x^T \Lambda y}$$

7c.) $R[w] = \sum_{i=1}^{n} \lambda^i (w^T x^i - y^i)^2 + \gamma \|w\|_2^2$

Same as previous problem except we add the derivative of the L_2 regularization

$$0 = X^T \Lambda X w - X^T \Lambda y + \gamma I w$$

$$X^T \Lambda y = X^T \Lambda X w + \gamma I w$$

$$X^T \Lambda y = (X^T \Lambda X + \gamma I) w$$

b.) $$\boxed{w = (X^T \Lambda X + \gamma I)^{-1} X^T \Lambda y}$$

a.) $$\boxed{R[w] = (Xw - y)^T \Lambda (Xw - y) + \lambda w^T w}$$

7d.) Regularization is used to prevent the weights from growing too large, and thus not generalizing well to new data. It is a means of reducing model complexity. By adding L_2 regularization, you penalize large weights.

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:

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

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

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

(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:

a.) The risk of classifying a new data point as class i is:

$$R(a_i|x) = \lambda_s (1 - P(w_i|x))$$

The risk of classifying a new data point as doubt is:

$$R(a_{c+1}|x) = \lambda_r \sum P(w_j|x) = \lambda_r$$

If it's better to choose doubt, then the following must be true for all classes:

$$\frac{\lambda_r}{\lambda_s(1 - P(w_i|x))} < 1 \Rightarrow 1 - \frac{\lambda_r}{\lambda_s} > P(w_i|x)$$

If the opposite of the inequality above does not hold true for a particular class, i, then you would not choose doubt.

b.) Always choose doubt because there is no cost ($\lambda_r = 0$). If doubt is greater than misclassification, always choose a label / guess.