Name: Jung Lin (Doris) Lee Student ID: 24253445

CS 189: Introduction to Machine Learning

Homework 2

Due: February 18, 2016 at 11:59pm

Instructions

- Homework 2 is completely a written assignment; no coding involved.
- We prefer that you typeset your answers using the LaTeX template on bCourses. If there is not enough space for your answer, you may continue your answer on the next page. Make sure to start each question on a new page.
- Neatly handwritten and scanned solutions will also be accepted. Make sure your answers are readable!
- Submit a PDF with your answers to the Homework 2 assignment on Gradescope. You should be able to see CS 189/289A on Gradescope when you log in with your bCourses email address. Please make a Piazza post if you have any problems accessing Gradescope.
- While submitting to Gradescope, you will have to select the pages containing your answer for each question.
- The assignment covers concepts in probability, linear algebra, matrix calculus, and decision theory.
- **Start early. This is a long assignment. Some of the material may not have been covered in lecture; you are responsible for finding resources to understand it.**

Problem 1: Expected Value.

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.

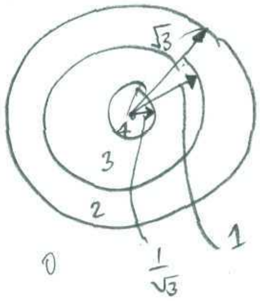
Let X be the distance of the hit from the center (in feet), and let the probability density function of X be

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

always $x>0$ ∵ x is radius can't be negative

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

Solution:



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

$$\langle\gamma\rangle = \int_0^\infty \gamma(x)\, f(x)\, dx$$

$\int \frac{1}{1+x^2}dx = \tan^{-1}x + C$

$$= 4\cdot\frac{2}{\pi}\int_0^{\frac{1}{\sqrt{3}}}\frac{1}{1+x^2}\,dx + 3\cdot\frac{2}{\pi}\int_{\frac{1}{\sqrt{3}}}^{1}\frac{1}{1+x^2}\,dx + 2\cdot\frac{2}{\pi}\int_1^{\sqrt{3}}\frac{1}{1+x^2}\,dx + 0$$

$$= \frac{8}{\cancel{\pi}}\,\frac{\cancel{\pi}}{6} + \frac{6}{\cancel{\pi}}\,\frac{\cancel{\pi}}{12} + \frac{4}{\cancel{\pi}}\,\frac{\cancel{\pi}}{12}$$

$$= \frac{4}{3} + \frac{1}{2} + \frac{1}{3}$$

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

$$\boxed{\text{Expected value of the score of a single shot} = \frac{13}{6}}$$

Problem 2: MLE.

parameter

variable

Assume that the random variable X has the exponential distribution

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

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 $x_5 = 2.6$, generated i.i.d. (i.e., independent and identically distributed).

likelihood ↓

Solution:

joint probability

$$\mathcal{L} = f(x_1, x_2, X_3, X_4, x_5 | \theta) = f(x_1|\theta) f(x_2|\theta) f(x_3|\theta) f(x_4|\theta) f(x_5|\theta)$$

$$= [\theta e^{-0.9\theta}][\theta e^{-1.7\theta}][\theta e^{-0.4\theta}][\theta e^{-0.3\theta}][\theta e^{-2.6\theta}]$$

$$= \theta^5 e^{-5.9\theta}$$

$$\ln \mathcal{L} = \ln(\theta^5) + \ln(e^{-5.9\theta}) = 5 \ln \theta - 5.9\theta$$

∵ log fun is monotonically increasing it is okay to just maximize the log f(x)

Maximize log-likelihood:

$$\frac{\partial \ln \mathcal{L}(\theta)}{\partial \theta} = \frac{5}{\theta} - 5.9 = 0 \Rightarrow \boxed{\theta = 0.8475}$$

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

Problem 3: Positive Definiteness.

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 $x^\top Ax$. 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:

a) $$x^T A x = [x_1 \cdots x_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

ex) $n=3$ $$[x_1\ x_2\ x_3] \underset{3\times 3}{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}} \underset{3\times 1}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \begin{bmatrix} a_{11}x_1 + a_{21}x_2 + a_{31}x_3 \\ a_{12}x_1 + a_{22}x_2 + a_{23}x_3 \\ a_{13}x_1 + a_{23}x_2 + a_{33}x_3 \end{bmatrix} \underset{3\times 1}{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}} = \begin{matrix} x_1(a_{11}x_1 + a_{21}x_2 + a_{31}x_3) + \\ x_2(a_{12}x_1 + a_{22}x_2 + a_{23}x_3) + \\ x_3(a_{13}x_1 + a_{23}x_2 + a_{33}x_3) \end{matrix}$$

Generalize to n:

$$= x_1(a_{11}x_1 + \cdots + a_{nn}x_n) + \cdots + x_n(a_{1n}x_1 + \cdots + a_{nn}x_n)$$

$$\boxed{x^T A x = \sum_{i=1\ldots n}\sum_{j=1\ldots n} x_i a_{ij} x_j}$$

b) $x^T A x > 0$ for positive definite matrix:

if $x = e1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$ then $$\underset{1\times n}{[1\ 0\ 0 \cdots]} \underset{n\times n}{\begin{bmatrix} a & & \\ & b & \\ \vdots & & c \ \ddots \end{bmatrix}} \underset{n\times 1}{\begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}} = [a + 0 + 0 + \cdots] \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = a > 0$$

So that means the 1^{st} element of A's diagonal must be positive

ditto the i^{th} element of A's diagonal is positive.

Reference: Math Stackexchange: symmetric positive definite matrix (Aug 8, 2012)

Problem 4: Short Proofs.

A is symmetric in all parts.

(a) Let A be a positive semidefinite matrix. Show that $A + \gamma I$ is positive definite for any $\gamma > 0$.

(b) Let A be a positive definite matrix. Prove that all eigenvalues of A are greater than zero.

(c) Let A be a positive definite matrix. Prove that A is invertible. (Hint: Use the previous part.)

• (d) Let A be a positive definite matrix. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^\top x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^\top B$.)

Solution:

4) a) A is positive semidefinite → $x^T A x \geq 0$

since we already proved in 3b) that all the diagonals of A are positive $(a_i > 0)$

$$A + \gamma I = \begin{bmatrix} a_1 & \cdots & \\ \vdots & \ddots & \\ \vdots & & a_n \end{bmatrix} + \begin{bmatrix} \gamma & 0 & 0 & \cdots & 0 \\ 0 & \gamma & & & \vdots \\ 0 & \cdots & & \ddots & \gamma \end{bmatrix} = \begin{bmatrix} a_1+\gamma & \cdots & & \\ \vdots & a_2+\gamma & & \\ \vdots & & \ddots & a_n+\gamma \end{bmatrix}$$

$a_i + \gamma > 0 \because a_i > 0$ and $\gamma > 0$

since this new matrix have positive diagonals therefore it is positive definite.

b) A is positive definite → $x^T A x > 0$

eigenvalue problem $Ax = \lambda x \Rightarrow (A - \lambda I)x = 0$

multiply x^T on both sides of the eigenvalue equation $x^T A x = x^T \lambda x$

we know that the LHS > 0 $\therefore x^T \lambda x > 0$

$\because x^T x > 0$ because if you have a negative number, you are always multiplying it to itself so the product is a positive number. Likewise for postive number, the product is positive, The sum of positive numbers is positive →

ex) $[-1 \ 5 \ -2] \begin{bmatrix} -1 \\ 5 \\ -2 \end{bmatrix} = (-1)(-1) + (5)(5) + (-2)(-2)$

$\therefore x^T x > 0$ in order for $x^T \lambda x > 0$ to be true

$\underbrace{\therefore \lambda > 0}$

→ all positive eigenvalues

4) c)

According to the invertible matrix theorem (Lay, Sec 2.3, Theorem 8) if the equation $A\vec{x}=0$ has only the trivial solution $x=0$ then the matrix A is invertible.

Since we have shown in 4b) that all eigenvalues of a positive definite matrix A are positive, so $\lambda=0$ can not be an eigenvalue, that means that for $\{A\vec{x}=\lambda\vec{x}=0\ ;\ \lambda\neq 0\}$ to be true, $\vec{x}=0$ is the only trivial solution.
therefore the matrix A must be invertible.

4) d) A positive definite $\rightarrow x^TAx>0$

Spectral theorem: "for any symetric matrix, there are exactly **n** real eigenvalues and the associated eigenvectors can be chosen to form a orthonormal basis" (EE127a website)

$$A=\sum_{i=1}^{n}\lambda_i u_i u_i^T=U\Lambda U^T \quad ; \quad \Lambda=\text{diag}(\lambda_1 \dots \lambda_n) \text{ whee } \lambda_i>0 \because \text{positive definite matrix have positive eigenvalues}$$

$$\Lambda=Q^TQ=\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix} \Rightarrow Q=\begin{bmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{bmatrix}$$

$$A=UQ^TQU^T=(QU^T)^T(QU^T)=B^TB$$

$$B=QU^T=\begin{bmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{bmatrix}\underbrace{\begin{bmatrix}\leftarrow u_1 \rightarrow \\ \leftarrow u_2 \rightarrow \\ \vdots \\ \leftarrow u_n \rightarrow\end{bmatrix}}_{\text{linearly independent eigenvector}}=$$

$$B=QU^T=\begin{bmatrix}\sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_n}\end{bmatrix}\begin{bmatrix}\leftarrow u_1 \rightarrow \\ \leftarrow u_2 \rightarrow \\ \vdots\end{bmatrix}=\begin{bmatrix}\leftarrow \sqrt{\lambda_1}u_1 \rightarrow \\ \leftarrow \sqrt{\lambda_2}u_2 \rightarrow \\ \vdots\end{bmatrix}$$

$$A=B^TB=\begin{bmatrix}\uparrow & & \uparrow \\ \sqrt{\lambda_1}u_1 & \cdots & \sqrt{\lambda_n}u_n \\ \downarrow & & \downarrow\end{bmatrix}\begin{bmatrix}\leftarrow \sqrt{\lambda_1}u_1 \rightarrow \\ \vdots \\ \leftarrow \sqrt{\lambda_n}u_n \rightarrow\end{bmatrix} \Rightarrow A_{ij}=x_i^Tx_j$$

we know x_i x_j are linearly independent $\because \{u_1 \dots u_n\}$ is orthogonal $\therefore \{u_1 \dots u_n\}$ is linearly independent and we are simply scaling it by a scalar so the linear independence is presened.

$$\underset{\text{eigenvalue equation}}{A\vec{u}=\lambda\vec{u}} \Rightarrow A\vec{x}=\lambda\vec{u}$$

if $\vec{u}$ linearly indepedent then $\vec{x}$ linearly independent

Problem 5: Derivatives and Norm Inequalities.

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 $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial \mathbf{x}}$.

(c) Let $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times n}$. Derive $\frac{\partial \text{Trace}(\mathbf{XA})}{\partial \mathbf{X}}$.

(d) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{n}\|\mathbf{x}\|_2$. (Note that $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.) (Hint: The Cauchy-Schwarz inequality may come in handy.)

Solution:

a) $$\frac{\partial(x^Ta)}{\partial x} = \frac{\partial x}{\partial x}a^T = \boxed{a^T}$$

treat as const

b) $$\underbrace{\frac{\partial(x^TAx)}{\partial x}}_{\text{product rule}} = \frac{\partial(x^TAx)}{\partial x} + \frac{\partial(x^TAx)}{\partial x} \underset{\text{from a)}}{=} (Ax)^T + Ax = \boxed{(A^T+A)x}$$

c)

$$\frac{\partial \text{Trace}(XA)}{\partial X} = \frac{\partial}{\partial X}\left[\text{Tr}\begin{bmatrix} \leftarrow x_1 \rightarrow \\ \leftarrow x_2 \rightarrow \\ \vdots \\ \leftarrow x_n \rightarrow \end{bmatrix}\begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ a_1 & a_2 & \cdots & a_n \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}\right]$$

$$= \frac{\partial}{\partial X}\left[\text{Tr}\begin{bmatrix} x_1^Ta_1 & x_1^Ta_2 & \cdots & x_1^Ta_n \\ \vdots & \ddots & & \\ x_n^Ta_1 & & & x_n^Ta_n \end{bmatrix}\right]$$

$$= \frac{\partial}{\partial X}\left[\sum_i x_i^Ta_i\right]$$

$$= \frac{\partial}{\partial X}(X^TA)$$

$$\boxed{= A^T}$$

5) d) ① $\|x\|_2 \le \|x\|_1$

$$\left[\sqrt{\sum_{i=1}^{n} x_i^2}\right]^2 \le \left[\sum_{i=1}^{n} |x_i|\right]^2$$

$$\sum_{i=1}^{n} x_i^2 \le \underbrace{\sum_{i=1}^{n} |x_i|^2}_{i=j} + 2\underbrace{\sum_{i=1}^{n} |x_i| \sum_{j=1}^{n} |x_j|}_{j\neq j}$$

$$0 \le 2\sum_{i=1}^{n} |x_i| \sum_{j=1}^{n} |x_j|$$

② $\|x\|_1 \le \sqrt{n}\,\|x_2\|$

$$\left[\sum x_i\right]^2 \le \left[\sqrt{n}\sqrt{\sum x_i^2}\right]^2$$

$$\sum |x_i|^2 \le n \sum x_i^2$$

$$\underbrace{\left|\sum x_i y_j\right|^2 \le \sum |x_j|^2 \sum \overset{1}{\cancel{|y_k|}}^2}_{\text{Cauchy-Schwarz inequality}} \qquad n = \sum |y_k|^2 = n \cdot 1$$

$$\uparrow$$
$$y_k = 1$$

in order to get a sum of n with a list of n numbers each y_k must be 1

∵ we have shown that ① & ② are true

$$\boxed{\therefore\ \|x\|_2 \le \|x\|_1 \le \sqrt{n}\,\|x_2\|}$$

Problem 6: Weighted Linear Regression.

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 a weight for each example. More explicitly, we have

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

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

In 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 $\mathbf{y}^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$, where ϵ is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk (cost) function is defined as follows:

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

$$= \omega^T X^T \Lambda X \omega$$
$$-2 Y^T X \omega - Y^T Y$$

(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 weight vector $\mathbf{w}$ that minimizes the risk function obtained in the previous part. You can assume that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is full rank. (Hint: You may use the expression you derived in Question 5(b).)

(c) The L_2 regularized risk function, for $\gamma > 0$, is

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

Rewrite this new risk function in matrix notation as in (a) and solve for $\mathbf{w}$ as in (b).

(d) How does γ affect the regression model? How does this fit in with what you already know about L_2 regularization? (Hint: Observe the different expressions for $\mathbf{w}$ obtained in (b) and (c).)

Solution:

6) a) $R[w] = \sum_{i=1}^{n} \lambda_i \left(w^T x_i - y_i\right)^2$

$$= \| X W - Y \|^2 \Lambda$$

$$= (XW - Y)^T \Lambda (XW - Y)$$

$$= \left[(XW)^T - Y^T\right] \Lambda (XW - Y)$$

$(AB)^T = B^T A^T$

$$= \left[W^T X^T - Y^T\right] \Lambda \left[XW - Y\right]$$

$$= \left[W^T X^T - Y^T\right]\left[\Lambda X W - \Lambda Y\right]$$

$$= W^T X^T \Lambda X W - Y^T \Lambda X W - W^T X^T \Lambda Y + Y^T \Lambda Y$$

b) $\frac{\partial R[w]}{\partial w} = 0 \Rightarrow \frac{\partial}{\partial w}\left[(XW)^T \Lambda X W\right] - \frac{\partial}{\partial w} Y^T \Lambda X W] - \frac{\partial}{\partial w}\left[W^T X^T \Lambda Y\right]$

$$= (\Lambda + \Lambda^T)(XW) - (Y^T \Lambda X)^T - X^T \Lambda Y$$

$$= 2X^T \Lambda X W - X^T \Lambda Y - X^T \Lambda Y$$

$$= 2X^T \Lambda X W - 2X^T \Lambda Y = 0$$

$$X^T \Lambda X W = X^T \Lambda Y$$

$$\boxed{\therefore\ W = (X^T \Lambda X)^{-1}(X^T \Lambda Y)}$$

using result from 5 b)

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

Λ is symetric

$\Lambda^T = \Lambda$

& $\frac{\partial (x^T a)}{\partial x} = a^T$

$\frac{\partial w^T A}{dw} = A$

c) $R[w] = \sum_{i=1}^{n} \lambda_i (W^T x_i + y_i) + \gamma \|w\|_2^2$

$$R[w] = w^T X^T \Lambda X w - Y^T \Lambda X w - w^T X^T \Lambda Y + Y^T \Lambda Y + \gamma \cdot w^T w$$

$$\frac{\partial R[w]}{\partial w} = 2X^T \Lambda X W - 2X^T \Lambda Y + \frac{\partial}{\partial w} \gamma\, w^T w] = 0$$

$\frac{d(w^T \bar{w})}{dw} + \frac{d(\bar{w}^T w)}{dw}$

$= \bar{w} + \bar{w}$

$= 2w$

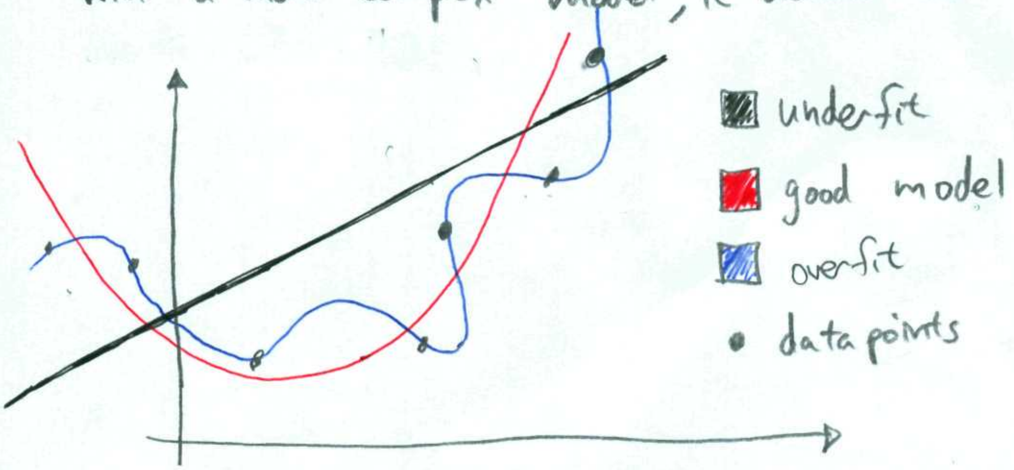
$$2X^T \Lambda X W - 2X^T \Lambda Y + 2w\gamma = 0$$

$$2(X^T \Lambda X w + I w \gamma) - 2X^T \Lambda Y = 0$$

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

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

6) d) As γ increases, we incur a larger penalty, so to minimize the risk function, $||\vec{w}$ must be small, this means that our fitting coefficients would be small and we will tend to underfit. As γ decreases, because our penalty for misclassification is not so large anymore, our weight can be large values. That means that we would be fitting a more complex model since the coefficients on the higher order terms are non-negligible. With a more complex model, it would be more likely to overfit.



Problem 7: 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

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

(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 $\lambda_r > \lambda_s$? Is this consistent with your intuition?

Solution:

a) Choose class i if $P(\omega_i|x) > P(\omega_j|x)$ makes sense because if your classifier is good the probability of it chosing the right class (class i in this case) must be higher than chosing the wrong class. risk= penalty

(i.e it is likely to make the right choice)

risk (certain about choices) = risk (uncertain about choices)

risk (getting it right) + risk (getting it wrong) = λr P(uncertain) & P(uncertain) = 1 because if you're uncertain you must get the uncertainty penalty

$$0 \cdot \cancelto{0}{P(\text{right})} + \lambda s \; P(\text{wrong}) = \lambda r$$

$$\lambda s \left[1 - P(\omega_i|x) \right] = \lambda r$$

$$1 - P(\omega_i|x) = \frac{\lambda r}{\lambda s}$$

$\therefore$ risk is minimized when $P(\omega_i|x) \geq 1 - \frac{\lambda r}{\lambda s}$

Continued

7) a) We are either certain about our result (in which case we'd get $\{0, \lambda_r\}$) or we are doubtful (and we'd always get the $\{\lambda_s\}$ penalty).

risk = (penalty from making that choice) · (Probabilty of making that choice)

Risk is also minimized if we chose doubt otherwise

b) If $\lambda_r = 0$, then risk is minimized when $P(\omega_i|x) \geq 0$ that means that it is always better to be uncertain to minimize the loss function because thee is no cost to being uncertain.

$N(\mu_i;\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Problem 8: Gaussians.

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with $P(\omega_1) = P(\omega_2) = 1/2$. Here, the classes are ω_1 and ω_2. For this problem, we have $\mu_2 \geq \mu_1$.

(a) Find the optimal Bayes decision boundary (i.e., find x such that $P(\omega_1 \mid x) = P(\omega_2 \mid x)$). What is the corresponding decision rule?

(b) Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-z^2/2} dz$$

where $a = \frac{\mu_2 - \mu_1}{2\sigma}$. The Bayes error is the probability of misclassification:

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2) P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1) P(\omega_1).$$

Solution:

a) Bayes theorem $P(x \mid \omega_i) = \frac{P(\omega_i \mid x) P(x)}{P(\omega_i)} \Rightarrow P(\omega_i \mid x) = \frac{P(x \mid \omega_i) P(\omega_i)}{P(x)}$

$$P(\omega_1 \mid x) = \frac{P(x \mid \omega_1) P(\omega_1)}{P(x)} = P(\omega_2 \mid x) = \frac{P(x \mid \omega_2) P(\omega_2)}{P(x)}$$

$$P(x \mid \omega_1) \cancel{P(\omega_1)} = P(x \mid \omega_2) \cancel{P(\omega_2)} \quad \because P(\omega_1) = P(\omega_2) = 1/2$$

$$P(x \mid \omega_1) = P(x \mid \omega_2)$$

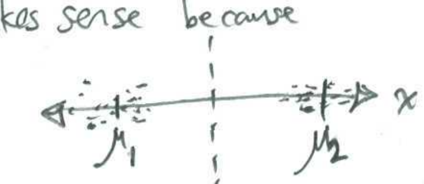
$$\cancel{\frac{1}{\sigma\sqrt{2\pi}}} e^{-\frac{(x-\mu_1)^2}{2\sigma}} = \cancel{\frac{1}{\sigma\sqrt{2\pi}}} e^{-\frac{(x-\mu_2)^2}{2\sigma}}$$

$$-\frac{(x-\mu_1)^2}{2\sigma} = -\frac{(x-\mu_2)^2}{2\sigma}$$

$$(x-\mu_1)^2 = (x-\mu_2)^2$$

$$x - \mu_1 = -(x - \mu_2) \Rightarrow \boxed{x = \frac{\mu_2 + \mu_1}{2}}$$

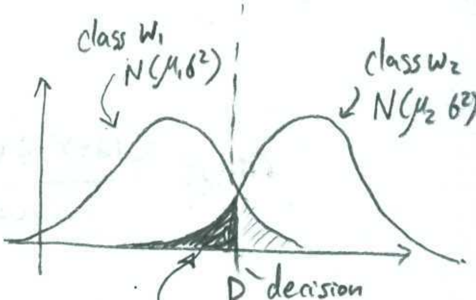
this makes sense because



optimal decision boundary lie in between these two means

8) b) $P_e = P(x=\omega_1 | \omega_2) P(\omega_2) + P(x=\omega_2 | \omega_1) P(\omega_1)$

$$= \frac{1}{2}\left[P(x=\omega_1 | \omega_2) + P(x=\omega_2 | \omega_1)\right]$$



the area under the curve here is where you have misclassification. (ie. the underlying dist. is class W_2 but you chose W_1)

$D = \frac{\mu_2 + \mu_1}{2}$ decision boundary from 8 a)

$$= \frac{1}{2}\left[\int_D^\infty N(\mu_1, \sigma^2)\, dx + \int_{-\infty}^D N(\mu_2, \sigma^2)\, dx\right]$$

$$= \frac{1}{2}\left[\frac{1}{\sigma\sqrt{2\pi}}\int_D^\infty e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}\, dx + \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^D e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}\, dx\right]$$

(Since we want to re arrange it to the given form - let $z = \frac{x-\mu_1}{\sigma} \to dz = \frac{dx}{\sigma} \Rightarrow dx = \sigma dz$)

$$= \frac{1}{2}\left[\underbrace{\frac{1}{\cancel{\sigma}\sqrt{2\pi}}\int_{z(D)}^{z(\infty)} e^{-\frac{z^2}{2}}\, \cancel{\sigma} dz} + \underbrace{\frac{1}{\cancel{\sigma}\sqrt{2\pi}}\int_{z(-\infty)}^{z(D)} e^{-\frac{z^2}{2}}\, \cancel{\sigma} dz}\right]$$

to determine the boundary, we plug in to $z(x) = \frac{x-\mu_1}{\sigma}$ and $z = \frac{x-\mu_2}{\sigma}$

$$z(D) = \frac{D-\mu_1}{\sigma} = \frac{\frac{\mu_2+\mu_1}{2} - \mu_1}{\sigma} = \frac{\frac{\mu_2}{2} - \frac{\mu_1}{2}}{\sigma} = \frac{\mu_2-\mu_1}{2\sigma} = a$$

$z(\infty) = \infty$

likewise for the second term: let $z = \frac{-x+\mu_2}{\sigma} \to dx = \sigma dz.$

$$z(D) = \frac{D-\mu_2}{\sigma} = \frac{\frac{\mu_2+\mu_1}{2} - \mu_1}{\sigma} = \frac{\frac{\mu_2}{2} - \frac{\mu_1}{2}}{\sigma} = \frac{\mu_2-\mu_1}{2\sigma} = a$$

$z(-\infty) = +\infty$

$$= \frac{1}{2\sqrt{2\pi}}\left[\int_a^\infty e^{-\frac{z^2}{2}}\, dz - \int_\infty^a e^{\frac{z^2}{2}}\, dz\right]$$

$$= \frac{1}{2\sqrt{2\pi}}\left[\int_a^\infty e^{-\frac{z^2}{2}}\, dz + \int_a^\infty e^{-\frac{z^2}{2}}\, dz\right]$$

$$= \frac{1}{\cancel{2}\sqrt{2\pi}}\, \cancel{2}\int_a^\infty e^{-\frac{z^2}{2}}\, dz$$

$$\boxed{P_e = \frac{1}{\sqrt{2\pi}}\int_a^\infty e^{-\frac{z^2}{2}}\, dz}$$