Question 1: Conditional Probability

• Part (a)

(i) $P(\text{windy}, \text{hits}) = P(\text{hit} \mid \text{windy}) P(\text{windy}) = (0.4)(0.3) = \underline{0.12}$

(ii) $P(\text{hit}) = P(\text{hit}, \text{windy}) + P(\text{hit}, \neg\text{windy}) = P(\text{hit} \mid \text{windy}) P(\text{windy}) + P(\text{hit} \mid \neg\text{windy}) P(\neg\text{windy}) = 0.4(0.3) + 0.7(0.7)$

$= \underline{0.61}$

(iii) $\dfrac{P(\text{hit then not hit}) + P(\text{not hit then hit})}{P(\text{hit}, \text{hit}) + P(\text{hit}, \text{not hit}) + P(\text{not hit}, \text{hit}) + P(\text{not hit}, \text{not hit})} = \dfrac{0.61(0.39) + 0.39(0.61)}{0.61^2 + 2(0.61)(0.39) + 0.39^2} = \underline{0.4758}$

(iv) $P(\neg\text{windy} \mid \text{miss}) = \dfrac{P(\neg\text{windy}, \text{miss})}{P(\text{miss})} = \dfrac{P(\text{miss} \mid \neg\text{windy}) P(\neg\text{windy})}{1 - P(\text{hit})} = \dfrac{(1 - P(\text{hit} \mid \neg\text{windy}))(1 - P(\text{windy}))}{1 - P(\text{hit})} = \dfrac{0.3(0.7)}{0.39}$

$= \underline{0.538}$

• Part (b)

$$P(A \mid B, C) = \frac{P(A, B, C)}{P(B, C)} = \frac{P(C \mid A, B) P(A, B)}{P(C \mid B) P(B)} = \frac{P(C \mid A, B)}{P(C \mid B)} \frac{P(A, B)}{P(B)}$$

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

since we are given that $P(A \mid B, C) > P(A \mid B)$ we can write:

$$\frac{P(C \mid A, B)}{P(C \mid B)} \frac{P(A, B)}{P(B)} > \frac{P(A, B)}{P(B)} \longrightarrow \frac{P(C \mid A, B)}{P(C \mid B)} > 1 \longrightarrow P(C \mid A, B) > P(C \mid B)$$

we can rewrite this as:

$$1 - P(C^c \mid A, B) > 1 - P(C^c \mid B) \longrightarrow \underline{P(C^c \mid A, B) < P(C^c \mid B)} \quad (*)$$

$$P(A \mid B, C^c) = \frac{P(C^c \mid A, B) P(A, B)}{P(C^c \mid B) P(B)} = \frac{P(C^c \mid A, B)}{P(C^c \mid B)} \cdot \frac{P(A, B)}{P(B)} = \frac{P(C^c \mid A, B)}{P(C^c \mid B)} P(A \mid B)$$

since we know $(*)$, and probabilities are positive, $\dfrac{P(C^c \mid A, B)}{P(C^c \mid B)} < 1$. and thus, $P(A \mid B, C^c) < P(A \mid B)$ given that $P(A \mid B, C) > P(A \mid B)$.

Question 2: Positive Definiteness

· Part (a)

(i) $\Rightarrow$ (ii)

$A \succeq 0$

$\underline{x}^T A \underline{x} \geq 0$

$(B\underline{y})^T A (B\underline{y}) \geq 0$ Say that $\underline{x} = B\underline{y}$ where $B\underline{y} \in \mathbb{R}^n$ and B is invertible in $\mathbb{R}^{n\times n}$ since $\underline{x} \in \mathbb{R}^n$

$\underline{y}(B^T A B)\underline{y} \geq 0$ $(*)$

since B is invertible, $\underline{y} \in \mathbb{R}^n$ and since $(*)$ is in the form of the definition of a positive semidefinite matrix, we can see that $B^T A B \succeq 0$ for some invertible matrix $B \in \mathbb{R}^{n\times n}$

(ii) $\Rightarrow$ (i)

$B^T A B \succeq 0$

$\underline{x}^T B^T A B \underline{x} \geq 0$

$(B\underline{x})^T A (B\underline{x}) \geq 0$

$\underline{y}^T A \underline{y} \geq 0$, where $\underline{y} = B\underline{x}$ $(*)$

since $\forall \underline{x} \in \mathbb{R}^n$ $\underline{x}^T B^T A B \underline{x} \geq 0$, we can see that applying the invertible $\mathbb{R}^{n\times n}$ matrix B on $\underline{x}$ maps $\underline{y} \in \mathbb{R}^n$. Therefore, A in $(*)$ is in the form of the definition of a positive semidefinite matrix and $A \succeq 0$.

(i) $\Rightarrow$ (iii)

(1) $A \succeq 0$

$\underline{x}^T A \underline{x} \geq 0$

$\underline{v}^T(\lambda \underline{v}) \geq 0$ using the definition of eigenvalue λ and eigenvector $\underline{v} \in \mathbb{R}^n$: $A\underline{v} = \lambda \underline{v}$, for some λ and its $\underline{v}$

$\lambda \underline{v}^T \underline{v} \geq 0$

$\lambda \|v\|_2^2 \geq 0$ for this to hold true, $\lambda \geq 0$. Therefore, all eigenvalues of A must be non-negative

(iii) $\Rightarrow$ (iv)

$A = U \Lambda U^T$ by the Spectral Theorem for Symmetric Matrices

$A = U \Lambda^{1/2} \Lambda^{1/2} U^T$ since Λ is a diagonal matrix and all $\lambda \geq 0$

$A = U \Lambda^{1/2} (\Lambda^{1/2})^T U^T$

$A = (U\Lambda^{1/2})(U\Lambda^{1/2})^T$

$A = MM^T$, $M = U\Lambda^{1/2}$

Therefore there exists some matrix $M \in \mathbb{R}^{n\times n}$ s.t $A = MM^T$

(iv) $\Rightarrow$ (i)

$A = UU^T$

$\underline{x}^T A \underline{x} = \underline{x}^T U U^T \underline{x}$

$\underline{x}^T A \underline{x} = (U^T \underline{x})^T (U^T \underline{x})$

$\underline{x}^T A \underline{x} = \|U^T x\|_2^2 \geq 0$ Therefore if there exists some matrix $U \in \mathbb{R}^{n\times n}$ s.t $A = UU^T$, then $A \succeq 0$

$\underline{x}^T A \underline{x} \geq 0$

$A \succeq 0$

Therefore, we conclude that (i), (ii), (iii), and (iv) are equivalent.

· Part (b)

(i) $\underline{x}^T I \underline{x} = \underline{x}^T \underline{x} = \|x\|_2^2 > 0$ Therefore $\lambda I \succ 0$ since $\lambda > 0$

$\underline{x}^T (A + \lambda I) \underline{x}$

$\underline{x}^T A \underline{x} + \underline{x}^T \lambda I \underline{x}$

$\underline{x}^T A \underline{x} + \lambda \|x\|_2^2$

since $A \succ 0$, $\underline{x}^T A \underline{x} > 0$ and since $\lambda > 0$, $\lambda \|x\|_2^2 > 0$

Therefore, $\underline{x}^T A \underline{x} + \lambda \|x\|_2^2 > 0$ and $A + \lambda I \succ 0$

(ii) $A - \gamma I \succ 0$

$\underline{x} (A - \gamma I) \underline{x}^T > 0$

$\underline{x} A \underline{x}^T - \underline{x} \gamma I \underline{x}^T > 0$

$\underline{x} A \underline{x}^T - \gamma \|x\|_2^2 > 0$

$\gamma < \dfrac{\underline{x} A \underline{x}^T}{\|x\|_2^2}$

$\gamma > 0$ since $\underline{x} A \underline{x}^T > 0$ and $\|x\|_2^2 > 0$

Therefore there exists some γ s.t $A - \gamma I \succ 0$

(iii) Say that not all the diagonal entries of A are positive s.t the i-th diagonal element $d_i < 0$.

construct $\underline{x}$ as a vector with 1 as its i-th element and 0 for the rest. $\underline{x}$ must still be in $\mathbb{R}^n - \{0\}$

Then we can see that $\underline{x}^T A \underline{x} = d_i$, which is a contradiction. Thus, all the diagonal entries of A must be positive.

(iv) Say that $\underline{x} \in \mathbb{R}^n$ and $\underline{x} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

$\underline{x}^T A \underline{x} = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}^T [a_1 \cdots a_n] \begin{bmatrix} \vdots \\ \vdots \end{bmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} > 0$ by definition of $A \succ 0$

Therefore, $\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} > 0$ if $A \succ 0 \in \mathbb{R}^{n \times n}$

Question 3: Derivatives and Norms

- Part (a)

$$\nabla_x (a^T \underline{x}) = \nabla_x \left(\sum a_i x_i\right) = \underline{a}$$

- Part (b)

$$\nabla_x (\underline{x}^T A \underline{x}) = \nabla_x ([\underline{x}^T a_1 \cdots \underline{x}^T a_n] \underline{x}) = \nabla_x \left(\sum_{i=1}^n x_i \sum_{j=1}^n x_j A_{ij}\right) = \nabla_x \left(\sum_{i=1}^n \sum_{j=1}^n x_i x_j A_{ij}\right)$$

$$\begin{bmatrix} 2\underline{x}_1 A_{11} + \sum_{j=2}^n x_j (A_{1j} + A_{j1}) \\ \vdots \\ 2x_n A_{nn} + \sum_{j=1}^{n-1} x_j (A_{nj} + A_{jn}) \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n x_j (A_{1j} + A_{j1}) \\ \vdots \\ \sum_{j=1}^n x_j (A_{nj} + A_{jn}) \end{bmatrix} = (A + A^T)\, \underline{x}$$

if A is symmetric then $A = A^T$. Thus if A is symmetric $\nabla_x (\underline{x}^T A \underline{x}) = 2A\underline{x}$

- Part (c)

$$A^T \underline{x} = [\underline{a}_1^T x_1 \cdots \underline{a}_n^T x_n]$$

$$\text{trace}(A^T X) = \sum_{i=1}^n (A^T X)_{ii} = \sum_{i=1}^n \sum_{j=1}^n A^T_{ij} X_{ij} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} X_{ji} = t \longrightarrow \nabla_x(t) = \begin{bmatrix} \partial t/\partial x_{11} & \cdots & \partial t/\partial x_{1n} \\ \vdots & & \\ \partial t/\partial x_{n1} & & \partial t/\partial x_{nn} \end{bmatrix}$$

$$\nabla_x (\text{trace}(A^T X)) = A$$

- Part (d)

Say $\underline{x} = \begin{bmatrix} a^2 \\ b^2 \end{bmatrix}$ and $\underline{y} = \begin{bmatrix} b^2 \\ a^2 \end{bmatrix}$, $a, b \in \mathbb{R}$

$$f(\underline{x}+\underline{y}) = \left(\sqrt{a^2+b^2} + \sqrt{b^2+a^2}\right)^2 = \left(2\sqrt{a^2+b^2}\right)^2 = 4(a^2+b^2)$$

$$f(\underline{x}) + f(\underline{y}) = \left(\sqrt{a^2} + \sqrt{b^2}\right)^2 + \left(\sqrt{b^2} + \sqrt{a^2}\right)^2 = 2(|a|+|b|)^2$$

clearly $f(\underline{x}+\underline{y}) > f(\underline{x}) + f(\underline{y})$. Counterexample so $f(x)$ is not a norm for vectors $x \in \mathbb{R}^2$.

- Part (e)

$$\|\underline{x}\|_\infty = \max_i |x_i|$$

$$\|\underline{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

It's easy to see that $\|\underline{x}\|_\infty \le \|\underline{x}\|_2$ since we are summing only positive squares for $\|\underline{x}\|_2$, of which $\max_i |x_i|$ is included. The max value of $\|\underline{x}\|_2$ would be a uniform vector $\underline{u} = \begin{bmatrix} u \\ \vdots \\ u \end{bmatrix}$. In this case,

$$\|\underline{u}\|_2 = \sqrt{\sum_{i=1}^n u^2} = \sqrt{n u^2} = \sqrt{n}\, u$$

since u is the max element in $\underline{u}$, $\|\underline{u}\|_2 = \sqrt{n} \max_i |x_i| = \sqrt{n} \|x\|_\infty$

$$\therefore \|x\|_\infty \le \|x\|_2 \le \sqrt{n} \|x\|_\infty$$

- Part (f)

By Cauchy-Schwartz: $|\langle \underline{x}, \underline{y} \rangle| \le \|\underline{x}\|_2 \|\underline{y}\|_2$

$$|\langle \underline{x}, \underline{1} \rangle| \le \|x\|_2 \|\underline{1}\|_2$$

$$\sum_{i=1}^n 1 \cdot |x_i| \le \sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n 1}$$

$$\|\underline{x}\|_1 \le \sqrt{n} \|\underline{x}\|_2$$

$$\|x\|_1^2 = \left(\sum_{i=1}^n |x_i|\right)^2 = (|x_1| + |x_2| + \cdots + |x_n|)^2$$

$$\|x\|_2^2 = \sum_{i=1}^n x_i^2 = (x_1^2 + x_2^2 + \cdots + x_n^2)$$

clearly $\|\underline{x}\|_1^2 \ge \|\underline{x}\|_2^2$ and since $\|\underline{x}\|_1 \ge 0$, then $\|\underline{x}\|_1 \ge \|\underline{x}\|_2$

$$\therefore \|\underline{x}\|_2 \le \|\underline{x}\|_1 \le \sqrt{n} \|x\|_2$$

Question 4 : Eigenvalues

· Part (a)

$\underline{x} = \alpha_1 \underline{v}_1 + \cdots + \alpha_n \underline{v}_n$ since eigenvalues span the vector space

$\underline{x}^T A \underline{x} = (\alpha_1 \underline{v}_1^T + \cdots + \alpha_n \underline{v}_n^T) A (\alpha_1 \underline{v}_1 + \cdots + \alpha_n \underline{v}_n)$

$= \alpha_1^2 \underline{v}_1^T A \underline{v}_1 + \cdots + \alpha_n^2 \underline{v}_n^T A \underline{v}_n$

$= \alpha_1^2 \underline{v}_1^T (\lambda_1 \underline{v}_1) + \cdots + \alpha_n^2 \underline{v}_n^T (\lambda_n \underline{v}_n)$ since $A\underline{v} = \lambda \underline{v}$

$= \alpha_1^2 \lambda_1 \|\underline{v}_1\|_2 + \cdots + \alpha_n^2 \lambda_n \|\underline{v}_n\|_2$

$\max_{\|\underline{x}\|_2=1} \underline{x}^T A \underline{x} = \alpha_i^2 \lambda_i + \cdots + \alpha_n^2 \lambda_n$

Since we are constrained by $\|\underline{x}\|_2 = 1$, $\sum_{i=1}^{n} |\alpha_i| = 1$

clearly, to maximize $\underline{x}^T A \underline{x}$ we should set $\alpha_i = 1$ where $\lambda_{max}(A) = \lambda_i$ so that $\underline{x} = \underline{v}_i$

$\therefore \max_{\|\underline{x}\|_2=1} \underline{x}^T A \underline{x} = \lambda_{max}(A)$

· Part (b)

We use a similar logic as in 4(a)

$\min_{\|\underline{x}\|_2=1} \underline{x}^T A \underline{x} = \alpha_1^2 \lambda_1 + \cdots + \alpha_n^2 \lambda_n > 0$ since λ and α_i^2 are non-negative (by 2(a))

since we are constrained by $\|\underline{x}\|_2 = 1$, $\sum_{i=1}^{n} |\alpha_i| = 1$

clearly, to minimize $\underline{x}^T A \underline{x}$ we should set $\alpha_i = 1$ where $\lambda_{min}(A) = \lambda_i$ so that $\underline{x} = \underline{v}_i$

$\therefore \min_{\|\underline{x}\|_2=1} \underline{x}^T A \underline{x} = \lambda_{min}(A)$

· Part (c)

Both optimization problems in (a) and (b) aren't convex programs. because their constraint $\|\underline{x}\|_2 = 1$ isn't convex. That is, $\|x\|_2 = 1$ creates a circle, which fails the 2 points test.

· Part (d)

$A = V \Lambda V^T$

$A^2 = (V \Lambda V^T)(V \Lambda V^T) = V \Lambda^2 V^T$

since Λ is a diagonal matrix of eigenvalues, Λ^2 is the diagonal matrix of eigenvalues squared. Thus if λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2. as the diagonal of Λ^2 are the eigenvalues of A^2.

$\therefore \lambda_{max}(A^2) = \lambda_{max}(A)^2$ and $\lambda_{min}(A^2) = \lambda_{min}(A)^2$

· Part (e)

$\|A\underline{x}\|_2 = \sqrt{(A\underline{x})^T(A\underline{x})} = \sqrt{\underline{x}^T A^T A \underline{x}} = \sqrt{\underline{x}^T A^2 \underline{x}} = \sqrt{\underline{x}^T}\, A \sqrt{\underline{x}}$, which is in the form $y^T A y$ where $y = \sqrt{\underline{x}}$

by 4(a) we can see that $\max_{\|y\|_2=1} y^T A y = \lambda_{max}(A)$ thus $\|A\underline{x}\|_2 \leq \lambda_{max}(A)$

by 4(b) we can see that $\min_{\|y\|_2=1} y^T A y = \lambda_{min}(A)$ thus $\|A\underline{x}\|_2 \geq \lambda_{min}(A)$

$\therefore \lambda_{min}(A) \leq \|A\underline{x}\|_2 \leq \lambda_{max}(A)$

· Part (f)

$\lambda_{min}(A) \leq \|Ay\|_2 \leq \lambda_{max}(A)$. from 4(e) where $y = \frac{1}{\|\underline{x}\|_2} \underline{x}$, $\|x\|_2 \neq 0$ and $\|y\|_2 = 1$

$\lambda_{min}(A) \leq \frac{1}{\|\underline{x}\|_2} \|A\underline{x}\|_2 \leq \lambda_{max}(A)$

$\lambda_{min}(A)\|\underline{x}\|_2 \leq \|A\underline{x}\|_2 \leq \lambda_{max}(A)\|\underline{x}\|_2$

Question 5: Gradient Descent

· Part (a)

the first-order optimality conditions → setting the gradient of the objective function to 0 then solving for $\underline{x}$

$\nabla_x (\frac{1}{2}\underline{x}^T A \underline{x} - \underline{b}^T \underline{x})$

$\frac{1}{2}(2A\underline{x}^*) - \underline{b} = 0$

$x^* = A^{-1}\underline{b}$

$\nabla_x^2 (\frac{1}{2}\underline{x}^T A \underline{x} - \underline{b}^T \underline{x}) = A \succeq 0$ since its eigenvectors are non-negative ($0 < \lambda_{min}(A)$ and by 2(a))

Therefore the optimization problem is convex.

· Part (b)

$x_{i+1} \leftarrow x_i - \eta \nabla_x f(x)$

$x_{i+1} \leftarrow x_i - 1 \cdot \nabla_x (\frac{1}{2}\underline{x}^T A \underline{x} - \underline{b}^T \underline{x})$

$x_{i+1} \leftarrow x_i - (A\underline{x}_i - \underline{b})$

$x_{i+1} \leftarrow (I - A)\underline{x}_i - \underline{b}$

· Part (c)

$x^{(k)} - \underline{x}^* = (I-A)\underline{x}^{(k-1)} - \underline{b} - \underline{x}^*$

Say $\underline{x}^* = x^{(\infty)}$ since we have a quadratic program as our loss function

$\underline{x}^{(k)} - \underline{x}^{(\infty)} = (I-A)\underline{x}^{(k-1)} - \underline{b} - ((I-A)\underline{x}^{(\infty)} - \underline{b})$

$\underline{x}^{(k)} - \underline{x}^{(\infty)} = (I-A)\underline{x}^{(k-1)} + (I-A)\underline{x}^{(\infty)}$

$\underline{x}^{(k)} - \underline{x}^{(\infty)} = (I-A)(\underline{x}^{(k-1)} - \underline{x}^{(\infty)})$

$\underline{x}^{(k)} - \underline{x}^* = (I-A)(\underline{x}^{(k-1)} - \underline{x}^*)$

· Part 5 (d)

$A\cdot\underline{v} = \lambda \underline{v}$

$\underline{v} - A\underline{v} = \underline{v} - \lambda\underline{v}$

$(I-A)\underline{v} = (1-\lambda)\underline{v}$ since we are given that $0 < \lambda_{min}(A)$ and $\lambda_{max}(A) < 1$, $1-\lambda > 0$

Since I is the identity matrix with a diagonal of ones $\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \\ \vdots & & & 1 \end{bmatrix}$ and $1-\lambda_i > 0$, then all the eigenvalues of $I-A$ are positive and it's easy to see that $I-A$ is also symmetric (I identity Matrix consists of 1's on diagonal so it doesn't affect the symmetry along the diagonal).

$\underline{x}^{(k)} - \underline{x}^* = (I-A)(\underline{x}^{(k-1)} - \underline{x}^*)$ from 5(c)

$\underline{x}^{(k)} - \underline{x}^* = By$, $B = I-A$, $y = \underline{x}^{(k-1)} - \underline{x}^*$

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 = \|By\|_2 \leq \lambda_{max}(B)\|y\|_2$ by 4(f)

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \leq \lambda_{max}(I-A)\|\underline{x}^{(k-1)} - \underline{x}^*\|_2$

since all λ_i of $I-A$ are in the range $(0,1)$, then

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \leq \rho\|\underline{x}^{(k-1)} - x^*\|_2$, $0 < \rho < 1$

Part 5(e)

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \le \rho \|\underline{x}^{(k-1)} - \underline{x}^*\|_2$ by 5(d)

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \le \rho \|\underline{x}^{(k-1)} - \underline{x}^*\|_2 \le \rho^2 \|\underline{x}^{(k-2)} - \underline{x}^*\|_2$, using the recurrence relation in 5(d)

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \le \rho^k \|\underline{x}^{(0)} - \underline{x}^*\|_2$.

since we want to find the k where we are $\varepsilon > 0$ close to $\underline{x}^*$

$\|\underline{x}^{(k)} - \underline{x}^*\|_2 \le \rho^k \|\underline{x}^{(0)} - \underline{x}^*\|_2 \geq \varepsilon$

$k \log \rho \le \log \varepsilon - \log \|\underline{x}^{(0)} - \underline{x}^*\|_2$

$$k \geq \frac{\log \varepsilon - \log \|\underline{x}^{(0)} - \underline{x}^*\|_2}{\log \rho}$$ since $0 < \rho < 1$ so $\log \rho < 1$

Part 5(f)

By 5(e) we know that it takes $k \geq \dfrac{\log \varepsilon - \log \|\underline{x}^{(0)} - \underline{x}^*\|_2}{\log \rho}$ steps to get $\varepsilon > 0$ close to $\underline{x}^*$

It takes $(2n-1)n$ for matrix-vector multiplication. Thus the running time is $k(2n-1)(n) = \dfrac{\log \varepsilon - \log \|x^{(0)} - x^*\|_2}{\log \rho}(2n-1)n$

or $\dfrac{\log \varepsilon - \log \|\underline{x}^{(0)} - \underline{x}^*\|_2}{\log \rho} O(n^2)$

Question 6(a): Classification

Part (a)

$$f(x) = \begin{cases} i & \text{if } P(Y=i\mid x) \geq P(Y=j\mid x)\ \forall j \text{ and } P(Y=i\mid x) \geq 1-\lambda_r/\lambda_s \\ c+1 & \end{cases}$$

case 1: if $f(x) = c+1$

$$R(f(x)=i\mid x) = \sum_{j=1}^{c} L(f(x)=i, y=j)P(Y=j\mid x)$$

$$= \lambda_r \sum_{j=1}^{c} P(Y=j\mid x)$$

$$= \lambda_r$$

Now say we have some policy $g: \mathbb{R}^d \to \{1, \dots, c+1\}$. We will consider when $g(x) \in \{1, \dots, c\}$

$$R(g(x)=i\mid x) = \sum_{j=1}^{c} L(g(x)=i, y=j)P(Y=j\mid x)$$

$$= \left[\sum_{j=1, j\neq i}^{c} L(g(x)=i, y=j)P(Y=j\mid x)\right] + L(g(x)=i, y=i)P(Y=i\mid x)$$

$$= \sum_{j=1, j\neq i}^{c} \lambda_s P(Y=j\mid x) + 0\, P(Y=i\mid x)$$

$$= \lambda_s(1-P(Y=i\mid x))$$

We can rewrite this as $P(Y=i\mid x) = 1 - \frac{1}{\lambda_s} R(g(x)=i\mid x)$

Since our policy f chose $c+1$, we know that $P(Y=i\mid x) < 1-\lambda_r/\lambda_s$ because there must be a greatest subset s.t. $\exists i \in \{1, c\}$ $P(Y=i\mid x) \geq P(Y=j\mid x)\ \forall j$

$$1 - \frac{1}{\lambda_s} R(g(x)=i\mid x) < 1 - \lambda_r/\lambda_s$$

$$R(g(x)=i\mid x) < \lambda_r$$

Thus, for some policy g to classify $i \in \{1, \dots, c\}$, $R(g(x)=i\mid x) < \lambda_r$ and therefore, to classify with the doubt class $c+1$, $R(g(x)=i\mid x) \geq \lambda_r$.

$\therefore$ f obtains minimal risk for $c+1$

case 2: if $f(x) \in \{1, \dots, c\}$

$$R(f(x)=i \mid x) = \sum_{j=1}^{c} L(f(x)=i, y=j) P(Y=j \mid x)$$

$$= \left[\sum_{j=1, j\neq i}^{c} L(f(x)=i, y=j) P(Y=j \mid x) \right] + L(f(x)=i, y=i) P(Y=i \mid x)$$

$$= \sum_{j=1, j\neq 1}^{c} \lambda_s P(Y=j \mid x) + 0$$

$$= \lambda_s (1 - P(Y=i \mid x))$$

We can rewrite this as $P(Y=i \mid x) = 1 - \frac{1}{\lambda_s} R(f(x)=i \mid x)$

since our policy f chose i, we know that $P(Y=i \mid x) \geq 1 - \lambda_r / \lambda_s$

$$1 - \frac{1}{\lambda_s} R(f(x)=i \mid x) \geq 1 - \lambda_r / \lambda_s$$

$$R(f(x)=i \mid x) < \lambda_r$$

Now say we have some policy $g: \mathbb{R}^d \to \{1, \dots, c\}$ and check when $g(x) = c+1$

$$R(g(x)=i \mid x) = \sum_{j=1}^{c} L(g(x)=i, y=j) P(Y=j \mid x)$$

$$= \lambda_r \sum_{j=1}^{c} P(Y=j \mid x)$$

$$= \lambda_r$$

Thus we can see that to classify as c+1, $R(g(x)=i \mid x) \geq \lambda_r$. Thus, our policy f is minimal in casifying $i \in \{1, \dots, c\}$

We conclude then that f obtains minimum risk

Part (b)

If $\lambda_r = 0$, then it is beneficial to choose the doubt class c+1 since there would be no loss for choosing doubt, just like for choosing the correct class. If $\lambda_r > \lambda_s$ then we would never classify doubt because we would have a higher loss for choosing doubt than choosing an incorrect class. That is, we would always classify and never choose doubt. This is consistent with intuition since we have shown in 6(a) that to choose doubt $R(g(x)=i \mid x) \geq \lambda_r$

Question 7: Gaussian Classification

· Part (a)

$P(\omega_1 | x) = P(\omega_2 | x)$ definition of Bayes' optimal decision boundary

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

$$\frac{\frac{1}{2}P(x|\omega_1)}{P(x)} = \frac{\frac{1}{2}P(x|\omega_2)}{P(x)}$$

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

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

$$x^2 - 2\mu_1 x + \mu_1^2 = x^2 - 2\mu_2 x + \mu_2^2$$

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

$$x_b = \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)} = \frac{\mu_2 + \mu_1}{2}$$ decision boundary

decision rule: choose class ω_2 if $x > x_b$ else ω_1 since $\mu_2 > \mu_1$

· Part (b)

$$P_e = P(\omega_1|\omega_2)P(\omega_2) + P(\omega_2|\omega_1)P(\omega_1)$$

$$P(\omega_1|\omega_2) = P(x < \tfrac{\mu_1+\mu_2}{2} \mid x \sim N(\mu_2, \sigma^2)) = P(\tfrac{x-\mu_2}{\sigma} < \tfrac{\mu_1-\mu_2}{2\sigma})$$

$$P(\omega_2|\omega_1) = P(x \geq \tfrac{\mu_1+\mu_2}{2} \mid x \sim N(\mu_2, \sigma^2)) = 1 - P(\tfrac{x-\mu_1}{\sigma} < \tfrac{\mu_2-\mu_1}{2\sigma}) = 1 - \Phi(\tfrac{\mu_2-\mu_1}{2\sigma}) = P(\omega_1|\omega_2)$$

$$P_e = \tfrac{1}{2}P(\omega_1|\omega_2) + \tfrac{1}{2}P(\omega_2|\omega_1) = P(\omega_1|\omega_2) = 1 - \Phi(\tfrac{\mu_2-\mu_1}{2\sigma})$$

$$= 1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-z^2/2}dz = \frac{1}{\sqrt{2\pi}}\int_{a}^{\infty} e^{-z^2/2}dz, \quad a = \frac{\mu_2-\mu_1}{2\sigma}$$

Question 8: Maximum Likelihood Estimation

Multinomial Distribution: $P(k_1, k_2, k_3) = \frac{3!}{\prod_{i=1}^{3}(k_i!)} \prod_{i=1}^{3} p_i^{k_i}$

$$\ell(p_1, p_2, p_3) = \ln(P(k_1, k_2, k_3)) = \ln(3!) - \sum_{i=1}^{3} \ln(k_i!) + \sum_{i=1}^{3} k_i \ln p_i$$

$$\mathcal{L}(p_1, p_2, p_3, \lambda) = \ell(p_1, p_2, p_3) + \lambda(1 - \sum_{i=1}^{3} p_i)$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial p_1} = 0 \rightarrow \frac{k_1}{p_1} - \lambda = 0 \rightarrow p_1 = \frac{k_1}{\lambda} \\ \frac{\partial \mathcal{L}}{\partial p_2} = 0 \rightarrow \frac{k_2}{p_2} - \lambda = 0 \rightarrow p_2 = \frac{k_2}{\lambda} \\ \frac{\partial \mathcal{L}}{\partial p_3} = 0 \rightarrow \frac{k_3}{p_3} - \lambda = 0 \rightarrow p_3 = \frac{k_1}{\lambda} \\ \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \rightarrow 1 - (p_1 + p_2 + p_3) = 0 \rightarrow p_1 + p_2 + p_3 = 1 \end{cases}$$

$$\frac{k_1}{\lambda} + \frac{k_2}{\lambda} + \frac{k_3}{\lambda} = 1$$

$\lambda = k_1 + k_2 + k_3 = n$ since k_i are counts and we are summing counts from all classes $\{1, 2, 3\}$

$\therefore p_1 = \frac{k_1}{n}, \quad p_2 = \frac{k_2}{n}, \quad p_3 = \frac{k_3}{n}$

$$H = \nabla_p^2 = \begin{bmatrix} -k_1/p_1^2 & 0 & 0 \\ 0 & -k_2/p_2^2 & 0 \\ 0 & 0 & -k_3/p_3^2 \end{bmatrix}$$

for $\underline{x} \in \mathbb{R}^3$

$$\underline{x}^T H \underline{x} = [-\frac{k_1}{p_1^2} x_1 \quad -\frac{k_2}{p_2^2} x_2 \quad -\frac{k_3}{p_3^2} x_3] \underline{x}$$

$$\underline{x}^T H \underline{x} = [-\frac{k_1}{p_1^2} x_1^2 \quad -\frac{k_2}{p_2^2} x_2^2 \quad -\frac{k_3}{p_3^2} x_3^2]$$

$\forall x \in \mathbb{R}^3 - \{0\}$, $\underline{x}^T H \underline{x} < 0$ since $k_i > 0$, probabilities are positive, and we square x_i. Therefore, the Hessian is negative definite and $\mathcal{L}$ is concave.