CS189, HW2

Completed by: Matthew Wu

1. Conditional Probability

In the following questions, **show your work**, not just the final answer.

(a) The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

(i) on a given shot there is a gust of wind and she hits her target.

(ii) she hits the target with her first shot.

(iii) she hits the target exactly once in two shots.

(iv) there was no gust of wind on an occasion when she missed.

Solution: For this problem, let $W = w$ mean that there is wind, and $T = t$ mean that the archer hits the target.

(i) $P(w) * P(t|w) = 0.3 * 0.4 = 0.12$

(ii) $P(T = t) = P(\neg w) * P(t|\neg w) + P(w) * P(t|w) = 0.3 * 0.4 + 0.7 * 0.7 = 0.12 + 0.49 = 0.61$

(iii) $P(t) * P(\neg t) + P(\neg t) * P(t) = 2 * P(t) * P(\neg t) = 2 * 0.39 * 0.61 = 0.4758$

(iv) $P(\neg w|\neg t) = \frac{P(\neg w \cap \neg t)}{P(\neg t)} = \frac{0.7*0.3}{0.39} = 0.5385$

(b) Let A, B, C be events. Show that if

$$P(A|B, C) > P(A|B)$$

then

$$P(A|B, C^c) < P(A|B),$$

where C^c denotes the complement of C. Assume that each event on which we are conditioning has positive probability.

Solution:
$P(A|B) = \frac{P(A|B,C)P(B|C)P(C)+P(A|B,C^c)P(B|C^c)P(C^c)}{P(B|C)P(C)+P(B|C^c)P(C^c)}$
Let $p_1 = P(B|C)P(C)$ and $p_2 = P(B|C^c)P(C^c)$
$P(A|B) = \frac{P(A|B,C)p_1+P(A|B,C^c)p_2}{p_1+p_2}$
$p_1P(A|B) + p_2P(A|B) = p_1P(A|B, C) + p_2P(A|B, C^c)$
$p_2(P(A|B) - P(A|B, C^c)) = p_1(P(A|B, C) - P(A|B))$
Since we are told $P(A|B, C) > P(A|B)$, we know the right side of the equation above is positive. This means that the left side of the equation must also be positive. This means that
$P(A|B) - P(A|B, C^c) > 0 \Rightarrow P(A|B, C^c) < P(A|B)$ $\square$.

2. Positive Definiteness

Definition. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix.

- We say that A is **positive definite** if $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top Ax > 0$. We denote this with $A \succ 0$.
- Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n$, $x^\top Ax \geq 0$. We denote this with $A \succeq 0$.

(a) For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, prove that all of the following are equivalent.

(i) $A \succeq 0$.

(ii) $B^\top AB \succeq 0$, for some invertible matrix $B \in \mathbb{R}^{n\times n}$.

(iii) All the eigenvalues of A are nonnegative.

(iv) There exists a matrix $U \in \mathbb{R}^{n\times n}$ such that $A = UU^\top$.

(Suggested road map: (i) $\Leftrightarrow$ (ii), (i) $\Rightarrow$ (iii) $\Rightarrow$ (iv)$\Rightarrow$ (i). For the implication (iii) $\Rightarrow$ (iv) use the *Spectral Theorem for Symmetric Matrices*.

Solution:
Suppose (i) is true. Then, $B^\top AB \succeq 0$ for $B = I$, the identity matrix. $\therefore$ (i) $\Rightarrow$ (ii).
Suppose (ii) is true. Then there is some invertible matrix B such that $\forall x \in \mathbb{R}^n$, $x^\top B^\top ABx \succeq 0$. However, $x^\top B^\top = (Bx)^\top$. $\therefore (Bx)^\top A(Bx) \succeq 0$. Let $Bx = y$, where y is another matrix in $\mathbb{R}^n$. Since B is invertible, there is a one to one correspondence between x and y, which means $\forall y \in \mathbb{R}^n$, $y^\top Ay \geq 0 \Rightarrow A \succeq 0$. $\therefore$ (i) $\Leftrightarrow$ (ii).

Suppose (i) is true. Assume that there is a negative eigenvalue λ for A. Then there is some nonzero vector x such that $Ax = \lambda x$. This gives us $x^\top(\lambda x) = \lambda x^\top x$. Since $x \neq \vec{0}$, $x^\top x > 0$. However λ is negative, which means $\lambda x^\top x = x^\top Ax < 0$, which contradicts $A \succeq 0$. $\therefore$ (i) $\Rightarrow$ (iii).
Suppose (iii) is true. Since A is symmetric, by the spectral theorem for symmetric matrices, there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ and an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ such that $A = VDV^\top$. Since all eigenvalues of A are nonnegative, all terms of D are nonnegative. We can construct another matrix E where $E_{ij} = \sqrt{D_{ij}}$. We have $EE = D$ and $E^\top = E$. This gives us $A = VEE^\top V^\top$. Let $U = VE$. Then we have $A = UU^\top$. $\therefore$ (iii) $\Rightarrow$ (iv).
Suppose (iv) is true. $\forall x \in \mathbb{R}^n$, $x^\top Ax = x^\top UU^\top x = (U^\top x)^\top (U^\top x) \geq 0 \Rightarrow A \succeq 0$. $\therefore$ (iv) $\Rightarrow$ (i).
$\therefore$ (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii) $\Leftrightarrow$ (iv). $\square$

(b) For a symmetric positive definite matrix $A \succ 0 \in \mathbb{R}^{n \times n}$, prove the following.

(i) For every $\lambda > 0$, we have that $A + \lambda I \succ 0$.

(ii) There exists a $\gamma > 0$ such that $A - \gamma I \succ 0$.

(iii) All the diagonal entries of A are positive; i.e. $A_{ii} > 0$ for $i = 1, \ldots, n$.

(iv) $\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$, where A_{ij} is the element at the i-th row and j-th column of A.

Solution:

(i) $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top (A + \lambda I)x = x^\top Ax + x^\top \lambda I x = x^\top Ax + \lambda x^\top x$.
We know $x^\top Ax > 0$, we know $\lambda > 0$, and we know $x^\top x > 0$.
$\therefore x^\top Ax + \lambda x^\top x > 0 \Rightarrow A + \lambda I \succ 0$.

(ii) Suppose that λ is an eigenvalue of A. Then there is some nonzero vector x such that $Ax = \lambda x$. For x, this gives us $x^\top Ax = x^\top \lambda x = \lambda x^\top x$. Since $x^\top x > 0$, it must be the case that $\lambda > 0$. This means that all eigenvalues of A are greater than 0.
Suppose we have an eigenvector x of A with eigenvalue λ. Consider what happens to this eigenvector in $A - \gamma I$.
$(A - \gamma I)x = Ax - \gamma I x = \lambda x - \gamma x = (\lambda - \gamma)x$
x is still an eigenvector of $A - \gamma I$, but the new eigenvalue for this eigenvector is $\lambda - \gamma$.
Let λ_{min} be the smallest eigenvalue of A. Let $\gamma = \lambda_{min}/2$. Then, all the eigenvalues of $A - \gamma I$ are still positive. This implies $A - \gamma I \succ 0$.

(iii) Assume that for at least one i where $1 \leq i \leq n$, we have $A_{ii} \leq 0$. Consider the vector x where $x_i = 1$ and $\forall j \in \{1 \ldots n\} - \{i\}$, $x_j = 0$. Let $Ax = y$. We have $y_i = A_{ii} \leq 0$. This gives us $x^\top Ax = x^\top y = A_{ii} \leq 0$. This means, $A \not\succ 0$. However, this is a contradiction.
$\therefore \forall i \in \{1, \ldots, n\}$, $A_{ii} > 0$.

(iv) Consider the vector x where $\forall i \in \{1, \ldots, n\}$, $x_i = 1$. Let $Ax = y$.
$\forall i \in \{1, \ldots, n\}$, $y_i = \sum_{j=1}^{n} A_{ij}$.
$\therefore x^\top y = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}$. Since $\forall x \in \mathbb{R}^n - 0$, $x^\top Ax > 0$, it must be the case that $\sum_{i=1}^{n} \sum_{j=1}^{n} > 0$.

3. Derivatives and Norms

In the following questions, **show your work**, not just the final answer.

(a) Let $x, a \in \mathbb{R}^n$. Compute $\nabla_x(a^\top x)$.

Solution: $\nabla_x(a^\top x) = \begin{bmatrix} a_1 x_1 & a_2 x_2 & a_3 x_3 & \dots \end{bmatrix}^\top = a$

(b) Let $A \in \mathbb{R}^{n \times n}$, $x \in \mathbb{R}^n$. Compute $\nabla_x(x^\top A x)$.
How does the expression you derived simplify in the case that A is symmetric?

(Hint: to get a feeling for the problem, explicitly write down a 2×2 or 3×3 matrix A with components A_{11}, A_{12}, etc., explicitly expand $x^\top A x$ as a polynomial without matrix notation, calculate the gradient in the usual way, and put the result back into matrix form. Then generalize the result to the $n \times n$ case.)

Solution:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (A_{11}x_1 + A_{21}x_2)x_1 + (A_{12}x_1 + A_{22}x_2)x_2 = f$$

$$\nabla_x f = \begin{bmatrix} 2A_{11}x_1 + A_{21}x_2 + A_{12}x_2 \\ A_{21}x_1 + A_{12}x_1 + 2A_{22}x_2 \end{bmatrix} = (A + A^\top)x$$

$\nabla_x(x^\top A x) = (A + A^\top)x$, which can be simplified to $2Ax$ if A is symmetric.

(c) Let $A, X \in \mathbb{R}^{n \times n}$. Compute $\nabla_X(\text{trace}(A^\top X))$.

Solution:

$$\text{trace}\left(\begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots \\ A_{31} & A_{32} & A_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}^\top \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots \\ X_{21} & X_{22} & X_{23} & \cdots \\ X_{31} & X_{32} & X_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \right) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} X_{ij}$$

$$\nabla_X(\text{trace}(A^\top X)) = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & A_{22} & A_{23} & \cdots \\ A_{31} & A_{32} & A_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = A$$

(d) For a function $f : \mathbb{R}^d \to \mathbb{R}$ to be a norm, the distance metric $\delta(x, y) = f(x - y)$ must satisfy the triangle inequality. Is the function $f(x) = (\sqrt{|x_1|} + \sqrt{|x_2|})^2$ a norm for vectors $x \in \mathbb{R}^2$? Prove it or give a counterexample.

Solution: Consider $x_1 = [1 \quad 0]$, $x_2 = [-1 \quad 0]$
$f(x_1) = f(x_2) = \sqrt{1}^2 = 1$
$f(x_1 - x_2) = (\sqrt{1} + \sqrt{1})^2 = 4$
$4 > 1 + 1$
Therefore this function isn't a norm.

(e) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

Solution:
Suppose that x_i is the largest component of x. Then $\|x\|_\infty = x_i$. We also have
$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_i^2 + \cdots + x_n^2} \geq \sqrt{x_i^2} = x_i$.
$\therefore \|x\|_\infty \leq \|x\|_2$.

Suppose that x_i is the largest component of x. Then the largest possible value of $\|x\|_2$ is if every single component is equal to x_i.
$\|x\|_2 \leq \sqrt{n * x_i^2} = \sqrt{n}x_i$.
We also have $\|x\|_\infty = x_i \Rightarrow \sqrt{n}\|x\|_\infty = \sqrt{n}x_i$.
$\therefore \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.
$\therefore \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$

(f) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.
(Hint: The Cauchy–Schwarz inequality may come in handy.)

Solution:
$\|x\|_2^2 = \langle x, x\rangle = x_1^2 + x_2^2 + \cdots + x_n^2 = \sum_{i=1}^n x_i^2$
$\|x\|_1^2 = (x_1 + \cdots + x_n)(x_1 + \cdots + x_n) = (\sum_{i=1}^n x_i^2) + \sum_{i=1}^{n-1}\sum_{j=i+1}^n |x_i||x_j|$
$\therefore \|x\|_2^2 \leq \|x\|_1^2 \Rightarrow \|x\|_2 \leq \|x\|_1$

Let $\vec{1}$ denote a vector in $\mathbb{R}^n$ where $x_1 = x_2 = \cdots = x_n = 1$.
$\|x\|_1 = \sum_{i=1}^n x_i = \langle \vec{1}, x\rangle$
By Cauchy-Schwarz, $\langle \vec{1}, x\rangle \leq \|\vec{1}\|_2\|x\|_2 = \sqrt{n}\|x\|_2 \Rightarrow \|x\|_1 \leq \sqrt{n}\|x\|_2$

$\therefore \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.

4. Eigenvalues

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $A \succeq 0$.

(a) Prove that the largest eigenvalue of A is

$$\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x.$$

(Hint: Use the *Spectral Theorem for Symmetric Matrices* to reduce the problem to the diagonal case.)

Solution: By the Spectral Theorem for Symmetric Matrices, there is a diagonal matrix $D \in \mathbb{R}^{n \times n}$ and an orthogonal matrix $U \in \mathbb{R}^{n \times n}$ such that $A = UDU^\top$, where the diagonal entries of D are the eigenvalues of A and the columns of U are the corresponding eigenvectors. Let $U^\top x = y$. Since $U^\top$ is orthogonal, there is a one to one correspondence for vectors x and y, and $\|x\|_2 = 1 \Rightarrow \|y\|_2 = 1$. Thus, we can reduce $x^\top Ax$ to $x^\top UDU^\top x$ to $y^\top Dy$.

Let $\lambda_1 = D_{1,1}, \ldots, \lambda_n = D_{n,n}$. Let $y_1, \ldots, y_n$ be the entries of y. We are trying to maximize $y^\top Dy$ subject to the constraint $\|y\|_2 = 1$.
$y^\top Dy = y_1^2\lambda_1 + y_2^2\lambda_2 + \cdots + y_n^2\lambda_n$, and $y_1^2 + y_2^2 + \cdots + y_n^2 = 1$. Suppose $k = argmax_x(\lambda_x)$. Clearly we can maximize $y_1^2\lambda_1 + y_2^2\lambda_2 + \cdots + y_n^2\lambda_n$ by letting $y_k^2 = 1$ and $y_i^2 = 0$ for $i \neq k$. This means $y^\top Dy = \lambda_k$. However, λ_k is also the largest eigenvalue of D.

(b) Similarly, prove that the smallest eigenvalue of A is

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x.$$

Solution: Similar to part (a), we can reduce $x^\top Ax$ to the case with a diagonal matrix D with the expression $y^\top Dy$ with $\|y\|_2 = 1$.

Let $\lambda_1 = D_{1,1}, \ldots, \lambda_n = D_{n,n}$. Let $y_1, \ldots, y_n$ be the entries of y. We are trying to minimize $y^\top Dy$ subject to the constraint $\|y\|_2 = 1$.
$y^\top Dy = y_1^2\lambda_1 + y_2^2\lambda_2 + \cdots + y_n^2\lambda_n$, and $y_1^2 + y_2^2 + \cdots + y_n^2 = 1$. Suppose $k = argmin_x(\lambda_x)$. Clearly we can minimize $y_1^2\lambda_1 + y_2^2\lambda_2 + \cdots + y_n^2\lambda_n$ by letting $y_k^2 = 1$ and $y_i^2 = 0$ for $i \neq k$. This means $y^\top Dy = \lambda_k$. However, λ_k is also the smallest eigenvalue of D.

(c) Is either of the optimization problems described in parts (a) and (b) a convex program? Justify your answer.

Solution: No. For a function to be a convex set, $S \subset \mathbb{R}^n$ if and only if
$\forall x, y \in S, \forall t \in [0,1], tx + (1-t)y \in S$. Consider the vectors $x = [1, 0, \ldots, 0]$ and $y = [-1, 0, \ldots, 0]$, and $t = \frac{1}{2}$.
$tx + (1-t)y = \frac{1}{2}[1, 0, \ldots, 0] + \frac{1}{2}[-1, 0, \ldots, 0] = [0, 0, \ldots, 0]$, which clearly doesn't have a 2-norm of 1. Therefore, the optimization problems described above are not convex.

(d) Show that if λ is an eigenvalue of A then λ^2 is an eigenvalue of A^2, and deduce that

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

Solution: Suppose that λ is an eigenvalue of A and x is the corresponding eigenvector. Then $Ax = \lambda x$. Suppose we want to solve for A^2x.
$A^2x = AAx = A(Ax) = A(\lambda x) = \lambda(Ax) = \lambda(\lambda x) = \lambda^2 x$
Therefore, λ^2 is an eigenvalue of A^2. It follows that for every eigenvalue λ of A, λ^2 is an eigenvalue of A^2. Also, since $A \succeq 0$, all the eigenvalues of A are nonnegative, which means $\lambda_1 < \lambda_2 \Rightarrow \lambda_1^2 < \lambda_2^2$. It clearly follows that if $\lambda_{\max}(A) = k$, then $\lambda_{\max}(A^2) = k^2 = \lambda_{\max}(A)^2$, and vice versa for the minimum eigenvalue.

(e) From parts (a), (b), and (d), show that for any vector $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$,

$$\lambda_{\min}(A) \leq \|Ax\|_2 \leq \lambda_{\max}(A).$$

Solution:
$\|Ax\|_2^2 = \langle Ax, Ax\rangle = x^\top A^\top Ax = x^\top A^2 x$ since A is symmetric.
From parts (a) and (b), we can conclude $\lambda_{\min}(A^2) \leq x^\top A^2 x \leq \lambda_{\max}(A^2)$.
From part (d), we can conclude $\lambda_{\min}(A)^2 \leq \|Ax\|_2^2 \leq \lambda_{\max}(A)^2$.
Taking the square root of the terms tells us $\lambda_{\min}(A) \leq \|Ax\|_2 \leq \lambda_{\max}(A)$.

(f) From part (e), deduce that for any vector $x \in \mathbb{R}^n$,

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

Solution:
Suppose that $\|x\|_2 = c$. Let $y = \frac{1}{c}x$. Then $\|y\|_2 = 1$ and we have
$\|Ax\|_2 = \|cAy\|_2 = c\|Ay\|_2$.
From part (e), we know $\lambda_{\min}(A) \leq \|Ay\|_2 \leq \lambda_{\max}(A)$. This implies that
$c\lambda_{\min}(A) \leq c\|Ay\|_2 \leq c\lambda_{\max}(A)$. Since $c = \|x\|_2$ and $\|Ax\|_2 = c\|Ay\|_2$, this implies that
$\lambda_{\min}(A)\|x\|_2 \leq \|Ax\|_2 \leq \lambda_{\max}(A)\|x\|_2$.

5. Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top A x - b^\top x$, where A is a symmetric matrix with $0 < \lambda_{\min}(A)$ and $\lambda_{\max}(A) < 1$.

(a) Using the first order optimality conditions, derive a closed-form solution for the minimum possible value of x, which we denote x^*.

Solution: We set the gradient of the objective function equal to 0 and solve for x. A is symmetric and invertible since all eigenvalues are nonzero.

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

(b) Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^*. Write down the update rule for gradient descent with a step size of 1.

Solution:

$$x^{(k)} = x^{(k-1)} - (Ax^{(k-1)} - b) = x^{(k-1)} - Ax^{(k-1)} + b$$

(c) Show that the iterates $x^{(k)}$ satisfy the recursion

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

Solution:

$$x^{(k)} = x^{(k-1)} - Ax^{(k-1)} + b$$
$$x^{(k)} - x^* = x^{(k-1)} - Ax^{(k-1)} + b - x^*$$
$$x^{(k)} - x^* = x^{(k-1)} - Ax^{(k-1)} - x^* + Ax^*$$
$$x^{(k)} - x^* = (I - A)x^{(k-1)} + (I - A)(-x^*)$$
$$x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$$

(d) Show that for some $0 < \rho < 1$,

$$\|x^{(k)} - x^*\|_2 \leq \rho \|x^{(k-1)} - x^*\|_2.$$

Solution: Using the Spectral Theorem for Symmetric Matrices, we have $A = UDU^\top$, where U is orthogonal and D is diagonal and has the eigenvalues of A. Also, since U is orthogonal, $UU^\top = I$, and $UIU^\top = I$. Consider the matrix $(I - A)$.
$(I - A) = UIU^\top - UDU^\top = U(I - D)U^\top$. Since the eigenvalues of A are all between 0 and 1, all the diagonal entries of D are between 0 and 1 and it's clear that $U(I - D)U^\top$ has eigenvalues strictly between 0 and 1, which implies $0 < \lambda_{\min}(A) \leq \lambda_{\max}(A) < 1$. From 4(f), we know $\lambda_{\min}(I - A)\|x^{(k-1)} - x^*\|_2 \leq \|(I - A)(x^{(k-1)} - x^*)\|_2 \leq \lambda_{\max}(I - A)\|x^{(k-1)} - x^*\|_2$. Using what we know about the eigenvalues of $(I - A)$, $0 < \|(I - A)(x^{(k-1)} - x^*)\|_2 < \|x^{(k-1)} - x^*\|_2$. From 5(c), it's clear that $\|x^{(k)} - x^*\|_2 = \|(I - A)(x^{(k-1)} - x^*\|_2$. From this, we can conclude $0 < \|x^{(k)} - x^*\|_2 < \|x^{(k-1)} - x^*\|_2 \Rightarrow \|x^{(k)} - x^*\|_2 \leq \rho\|x^{(k-1)} - x^*\|_2$ for some $0 < \rho < 1$.

(e) Let $x^{(0)} \in \mathbb{R}^n$ be a starting value for our gradient descent iterations. If we want our solution $x^{(k)}$ to be $\epsilon > 0$ close to x^*, i.e. $\|x^{(k)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2$, and ϵ. Note that $0 < \rho < 1$, so $\log \rho < 0$.

Solution: Every iteration of gradient descent, our distance from x^* is multiplied by ρ. Thus, we want to solve for the value of k that satisfies the following inequality:

$$\|x^{(0)} - x^*\|\rho^k \leq \epsilon$$

$$\rho^k \leq \frac{\epsilon}{\|x^{(0)} - x^*\|}$$

$$k \log(\rho) \leq log\left(\frac{\epsilon}{\|x^{(0)} - x^*\|}\right)$$

$$k \geq \log_\rho \left(\frac{\epsilon}{\|x^{(0)} - x^*\|}\right)$$

(f) Observe that the running time of each iteration of gradient descent is dominated by a matrix-vector product. What is the overall running time of gradient descent to achieve a solution $x^{(k)}$ which is ϵ-close to x^*? Give your answer in terms of $\rho, \|x^{(0)} - x^*\|_2, \epsilon$, and n.

Solution: The runtime for the matrix-vector product is $O(n^2)$. As we calculated in the last part, we need to perform $\log_\rho(\frac{\epsilon}{\|x^{(0)} - x^*\|})$ iterations. Thus, the runtime is

$$O(n^2 \log_\rho \left(\frac{\epsilon}{\|x^{(0)} - x^*\|}\right))$$

6. Classification

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional "doubt" category labeled $c+1$. Let $f : \mathbb{R}^d \to \{1, \ldots, c+1\}$ be a decision rule. Define the loss function

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

where $\lambda_r \geq 0$ is the loss incurred for choosing doubt and $\lambda_s \geq 0$ is the loss incurred for making a misclassification. Hence the risk of classifying a new data point x as class $i \in \{1, 2, \ldots, c+1\}$ is

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

(a) Show that the following policy obtains the minimum risk. (1) Choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and $P(Y = i|x) \geq 1 - \lambda_r/\lambda_s$; (2) choose doubt otherwise.

Solution: Obviously we want to pick the class that x is most likely to fall under. If we pick class i, and $P(Y = i|x) \geq P(Y = j)$ for all j, then we should pick class i, because x is at least as likely to be categorized under class i as it will under any other individual class. However, it is possible that it would be better to classify x under the doubt category if the expected risk for classifying x under i is too high. $E[R(f(x) = i|x)] = (1 - P(Y = i|x))\lambda_s$, and $E[R(f(x) = c+1|x) = \lambda_r$. Thus, we should pick class i if $(1 - P(Y = i|x))\lambda_s \leq \lambda_r$, and doubt otherwise.

$$(1 - P(Y = i|x))\lambda_s \leq \lambda_r$$

$$\Rightarrow 1 - P(Y = i|x) \leq \frac{\lambda_r}{\lambda_s}$$

$$\Rightarrow P(Y = i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Thus, the proposed policy obtains the minimum risk.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Explain why this is consistent with what one would expect intuitively.

Solution: If $\lambda_r = 0$, then we will only classify x under class i if there is a 100% chance that it belongs under class i. This makes sense intuitively, because there is no risk for not classifying x, so if we ever have any doubt we might as well not classify that point. If $\lambda_r > \lambda_s$, then we will classify every single point. This makes sense because if there is a higher risk associated with saying that we have doubt about where the point goes than guessing wrong, then we might as well classify every point even if we have no clue where it belongs.

7. Gaussian Classification

Let $P(x|\omega_i) \sim N(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with classes ω_1 and ω_2, $P(\omega_1) = P(\omega_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

Solution: For a normal distribution,

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\mu)^2/(2\sigma^2)}$$

To find the decision boundary, we want to find the point where x is equally likely to be in class ω_1 and ω_2. In essence, we want

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

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

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

$$x - \mu_1 = \mu_2 - x$$

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

$$x = \frac{\mu_1 + \mu_2}{2}$$

(b) The Bayes error is the probability of misclassification,

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

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 $a = \frac{\mu_2 - \mu_1}{2\sigma}$.

Solution: σ is the same for both distributions, and $P(\omega_1) = P(\omega_2) = \frac{1}{2}$, so by symmetry:

$$P((\text{misclassified as } \omega_1)|\omega_2)P(\omega_2) = P((\text{misclassified as } \omega_2)|\omega_1)P(\omega_1) \Rightarrow$$

$$P_e = 2P((\text{misclassified as } \omega_2)|\omega_1)P(\omega_1) = P((\text{misclassified as } \omega_2|\omega_1)$$

$$= \int_{\frac{\mu_1+\mu_2}{2}}^\infty \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_1)^2/(2\sigma^2)} dx$$

$$\text{Let } z = \frac{1}{\sigma}(x-\mu_1) \Rightarrow dz = \frac{1}{\sigma} dx$$

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

8. Maximum Likelihood Estimation

Let X be a discrete random variable which takes values in $\{1, 2, 3\}$ with probabilities $P(X = 1) = p_1, P(X = 2) = p_2$, and $P(X = 3) = p_3$, where $p_1 + p_2 + p_3 = 1$. Show how to use the method of maximum likelihood to estimate p_1, p_2, and p_3 from n observations of $X : x_1, \ldots, x_n$. Express your answer in terms of the counts

$$k_1 = \sum_{i=1}^{n} \mathbb{1}(x_i = 1), k_2 = \sum_{i=1}^{n} \mathbb{1}(x_i = 2), \text{ and } k_3 = \sum_{i=1}^{n} \mathbb{1}(x_i = 3),$$

where

$$\mathbb{1}(x = a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a. \end{cases}$$

We want to maximize the function $L(p) = p_1^{k_1} p_2^{k_2} p_3^{k_3} = p_1^{k_1} p_2^{k_2} (1 - p_1 - p_2)^{k_3}$. This is equivalent to maximizing $\ln(L(p_1, p_2)) = k_1 \ln(p_1) + k_2 \ln(p_2) + k_3 \ln(1 - p_1 - p_2)$.

$$H(\ln(L(p_1, p_2))) = \begin{bmatrix} -\frac{k_1}{p_1^2} - \frac{k_3}{(1-p_1-p_2)^2} & -\frac{k_3}{(1-p_1-p_2)^2} \\ -\frac{k_3}{(1-p_1-p_2)^2} & -\frac{k_2}{p_2^2} - \frac{k_3}{(1-p_1-p_2)^2} \end{bmatrix}$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} H \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -x_1\left(\frac{x_1 k_1}{p_1^2} + \frac{(x_1 + x_2)k_3}{(1 - p_1 - p_2)^2}\right. - x_2\left(\frac{x_2 k_2}{p_2^2} + \frac{(x_1 + x_2)k_3}{(1 - p_1 - p_2)^2}\right.$$

$$= -\frac{x_1^2 k_1}{p_1^2} - \frac{x_2^2 k_2}{p_2^2} - \frac{(x_1 + x_2)^2 k_3}{(1 - p_1 - p_2)^2}$$

From this it's easy to deuce that the hessian of $\ln(L(p_1, p_2))$ is negative definite, which means the function is concave. This means if we find the point where the gradient is 0, it is guaranteed to be the maximum. We set the gradient equal to 0.

$$\begin{bmatrix} \frac{k_1}{p_1} - \frac{k_3}{1-p_1-p_2} \\ \frac{k_2}{p_2} - \frac{k_3}{1-p_1-p_2} \end{bmatrix} = \begin{bmatrix} \frac{k_1}{p_1} - \frac{k_3}{p_3} \\ \frac{k_2}{p_2} - \frac{k_3}{p_3} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\frac{k_1}{p_1} = \frac{k_2}{p_2} = \frac{k_3}{p_3}, \quad p_1 + p_2 + p_3 = 1$$

From here, it's clear that $p_1 = \frac{k_1}{k_1+k_2+k_3}$, $p_2 = \frac{k_2}{k_1+k_2+k_3}$, $p_3 = \frac{k_3}{k_1+k_2+k_3}$.