CS189/289A – Spring 2017 — Homework 2

Yicheng Chen, SID 26943685

1. Conditional Probability

(a) i) P(Wind and hits) $= 0.3 \times 0.4 = 0.12$

ii) P(hits with first shot) $= 0.3 \times 0.4 + (1 - 0.3) \times 0.7 = 0.61$

iii) P(hits once in two shots) $= 0.61 \times (1 - 0.61) \times 2 = 0.4758$

iv) P(no wind when missed) $= \frac{0.7*0.3}{0.7*0.3+0.3*0.6} = \frac{21}{39} = \frac{3}{13}$

(b) We know that

$$P(ABC) + P(ABC^C) = P(AB) = P(AB)P(C|B) + P(AB)P(C^C|B)$$

Because

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

$$P(ABC) > P(AB)P(BC)/P(B) = P(AB)P(C|B)$$

Therefore

$$P(ABC^C) < P(AB)P(C^C|B)$$

Which is equivalent to

$$\frac{P(ABC^C)}{P(BC^C)} < \frac{P(AB)}{P(B)}$$

Or,

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

2. Positive Definiteness

(a) (i)→(ii): if $A \succeq 0$, then $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top B^\top ABx = (Bx)^\top A(Bx) \geq 0$. Therefore $B^\top AB \succeq 0$.

(ii)→(i): if $B^\top AB \succeq 0$, because B is invertible, then $\forall x \in \mathbb{R}^n - \{0\}$, there is a y that $x = By$. Then $x^\top Ax = (By)^\top A(By) \geq 0$. Hence $A \succeq 0$.

(i)→(iii): Suppose there exist an eigenvalue λ_i that is negative, then let x be the eigen vector corresponding to that $lambda_i$, then $x^\top Ax < 0$. This contradicts the assumption that $A \succeq 0$. Therefore all the eigenvalues or A are nonnegative.

(iii)→(iv): According to the spectral theorem, $A = QBQ^\top$, where B is a diagonal matrix with entries being the eigenvalues of A. If all the eigenvalues of A are nonegative, then there exists C that $B = C^\top C$. Therefore $A = QC^\top CQ^\top$. Let $U = QC^\top$, so $A = UU^\top$.

(iv)→(i): If there is a matrix U such that $A = UU^\top$, then $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top Ax = x^\top UU^\top x = ||(U^\top x)||^2 \geq 0$, so $A \succeq 0$.

(b) i) Obviously, $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top (A_\lambda I)x = x^\top Ax + \lambda||x||^2 > 0$, therefore $A + \lambda I \succeq 0$.

ii) Using the spectral theorem, $A = QBQ^\top$, where B is diagonal matrix with all the diagonal entries greater than 0. Then $A - \gamma I = QBQ^\top - \gamma I = Q(B - \gamma I)Q^\top$. Suppose the smallest eigenvalue of A is a, let $\gamma < a$, then $B' = B - \gamma I$ is still a all-possitive diagonal matrix. Therefore $A' = QB'Q^\top$ is also positive definite.

iii) Suppose $A_{jj} <= 0$, then let $x = (0, 0, ...1, ..., 0)^\top$ be a vector with the jth entry be 1. Then $x^\top Ax = A_{jj} \leq 0$, which contradicts to the assumption that $A \succ 0$. Therefore $A_{jj} > 0$ for all j.

iv) Similar to (iii), let $x = (1, 1, 1, ..., 1)^\top$ be a all-one vector, then this requires that $\sum_i \sum_j A_{ij} > 0$.

3. Derivatives and Norms

(a)
$$\nabla_x(a^\top x) = \nabla_x(a_1x_1 + a_2x_2 + ... + a_nx_n) = (a_1, a_2, ..., a_n)^\top = a$$

(b) If A is 2-by-2 matrix, we can get the following result:
$$\nabla_x(x^\top Ax) = (2A_{11}x_1 + (A_{12} + A_{21})x_2, 2A_{22}x_2 + (A_{12} + A_{21})x_1)$$

This is equivalent to the matrix form $\nabla_x(x^\top Ax) = (A + A^\top)x$. If the matrix A is symmetric, then $\nabla_x(x^\top Ax) = 2Ax$.

(c)
$$\nabla_X(\text{trace}(A^\top X)) = \nabla_X(\sum_i \sum_j A_{ij}X_{ij}) = A$$

(d) Let $x = (1, 1), y = (-1, -1)$, then $\delta(x, y) = f(x, y) = (\sqrt{1+1} + \sqrt{1+1})^2 = 8$. But $f(x) = f(y) = 2$. This does not satisfy the triangle inequality that $\delta(x, y) \leq f(x) + f(y)$, so this function $f(x)$ is not a norm.

(e) Let $||x||_\infty = \max x_i = x_M$, then
$$||x||_2 = \sqrt{\sum_i x_i^2} \geq \sqrt{x_M^2} = ||x||_\infty$$

And
$$||x||_2 = \sqrt{\sum_i x_i^2} \leq \sqrt{nx_M^2} = \sqrt{n}||x||_\infty$$

(f)
$$||x||_2^2 = \sum x_i^2 \leq (|x_1| + |x_2| + ... + |x_n|)^2 = ||x||_1^2$$

Therefore
$$||x||_2 \leq ||x||_1$$

Using the Cauchy-Schwarz inequality $| < a, b > | \leq ||a||_2||b||_2$,
$$| < x, \vec{1} > | = ||x||_1 \leq ||x||_2||\vec{1}||_2 = \sqrt{n}||x||_2$$

4. Eigenvalues

(a) According to the spectral theorem, $A = Q^\top \Lambda Q$, where Λ is a diagonal matrix whose diagonal entries are eigenvalues of A. Let $y = Qx$, because Q is an orthogonal matrix, $||y||_2 = 1$. then

$$\max_{||x||_2=1} x^\top Ax = \max_{||y||_2=1} y^\top \Lambda y = \max_{||y||_2=1} \sum \lambda_i y_i^2 = \lambda_{max}(A)$$

(b) Similar to (a),

$$\min_{||x||_2=1} x^\top Ax = \min_{||y||_2=1} y^\top \Lambda y = \min_{||y||_2=1} \sum \lambda_i y_i^2 = \lambda_{min}(A)$$

(c) The optimization problem in (a) is equivalent to:

$$\max_{z_1, z_2, \ldots z_n} \sum_i \lambda_i z_i$$

$$\text{s.t. } \sum_i z_i = 1$$

Which is a linear program, so it is convex. The problem in (b) is similar to (a), except that it is minimzation problem instead of maximization.

(d) If $Ax = \lambda x$, then

$$A^2 x = A(Ax) = A(\lambda x) = \lambda Ax = \lambda \times \lambda x = \lambda^2 x$$

Therefore λ^2 is an eigenvalue of A^2.
Because $A \succeq 0$, all $\lambda > 0$. Therefore for all eigenvalues λ of A, λ^2 are eigenvalues of A^2. Among them, the largest one is λ_{max}^2 and the smallest one is λ_{min}^2, therefore

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

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

(e)

$$\lambda_{min}(A) = \sqrt{\lambda_{min}(A^2)} \leq \sqrt{x^\top A^\top Ax} = ||Ax||_2 \leq \sqrt{\lambda_{max}(A^2)} = \lambda_{max}(A)$$

(f) For any vector x, there is a vector $y = \frac{x}{||x||_2}$ which satisfies $||y||_2 = 1$. Then according to (e), there is

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

Therefore,

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

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

5. Gradient Descent

(a) The first-order optimality condition is:

$$Ax^* = b$$

Therefore the closed-form solution is:

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

(b) $x^{(k+1)} = x^{(k)} - \nabla f(x^{(k)})$

(c) $x^{(k)} - x^* = x^{(k-1)} - \nabla f(x^{(k-1)}) - x^* = x^{(k-1)} - x^* - (Ax^{(k-1)} - Ax^*) = (I - A)(x^{(k-1)} - x^*)$

(d) Because the eigenvalues of A are all between 0 and 1, then the eigenvalues of matrix $B = I - A$ is also between 0 and 1. From (c), we got $x^{(k)} - x^* = (I - A)(x^{(k-1)} - x^*)$, so

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

. Let $\rho = \lambda_{max}(B)$ then there is

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

.

(e) From (d), there is:

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

To achieve $||x^{(k)} - x^*||_2 \leq \epsilon$, there should be

$$\rho^k||x^{(0)} - x^*||_2 \leq \epsilon$$

Therefore,

$$k \geq \frac{\log \frac{\epsilon}{||x^{(0)} - x^*||_2}}{\log \rho}$$

(f) The running time of each iteration is $O(n^2)$, and there are $\frac{\log \frac{\epsilon}{||x^{(0)} - x^*||_2}}{\log \rho}$ iteration. So the total running time is $O(\frac{\log \frac{\epsilon}{||x^{(0)} - x^*||_2}}{\log \rho} n^2)$.

6. Classification

(a) According to the definition of the loss function, if we choose $c+1$, $R(f(x)=i|x)=\lambda_r$. If we don't choose doubt, then $R(f(x)=i|x)=\lambda_s(1-P(Y=k|x))$. To minimize the risk, obviously when $\lambda_r < \lambda_s(1-P(Y=k|x))$ for all k, we should choose $c+1$. Otherwise when $\lambda_r \geq \lambda_s(1-P(Y=k|x))$ for some k, we should choose the k such that $P(Y=k|x)$ is maximized. Therefore we got the policy to obtain the minimum risk.

(b) If $\lambda_r = 0$, then we always classify the input as "doubt", because this will have zero risk.
If $\lambda_r > \lambda_s$, the first policy is always adopted, so we never choose "doubt". This is consistent with our intuitive because when $\lambda_r > \lambda_s$, no matter which class we decide, we will have a lower risk than "doubt", so we never "doubt".

7. Gaussian Classification

(a) According to the Bayes optimal decision theory, the boundary is the point where the two Gaussian distribution intersects, which is $\frac{\mu_1+\mu_2}{2}$. And the desicion rule is that when $x \leq \frac{\mu_1+\mu_2}{2}$, $f(x) = \omega_1$, otherwise $f(x) = \omega_2$.

(b) Because the two distributions are symmetric with respect to the decision boundary, so

$$P_e = 2 \times 0.5 \times \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx$$

Let $z = \frac{x-\mu_1}{\sigma}$, then

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

8. Maximum Likelihood Estimation

The likelihood function is:

$$L(\vec{p}) = P(X_1 = x_1, X_2 = x_2, ., ..X_n = x_n) = p_1^{k_1} p_2^{k_2} p_3^{k_3}$$

Where $k_j = \sum_{i=1}^{n} \mathbb{I}(x_i = j)$. To maximize the likelihood, take the logarithm of the likelihood and then the derivative of the log likelihood should be zero:

$$L'(\vec{p}) = k_1 \log p_1 + k_2 \log p_2 + k_3 \log(1 - p_1 - p_2)$$

$$\frac{\partial L'(\vec{p})}{\partial p_1} = \frac{k_1}{p_1} - \frac{n - k_1 - k_2}{1 - p_1 - p_2} = 0$$

$$\frac{\partial L'(\vec{p})}{\partial p_2} = \frac{k_1}{p_2} - \frac{n - k_1 - k_2}{1 - p_1 - p_2} = 0$$

Solve the above two equation,

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