CS289–Spring 2017 — Homework 2 Solutions

Shuhui Huang, SID 3032129712

Collaborators:

1. Conditional Probability

(a) (i) Let X=1{an archer hits her target}, Y=1{a gust of wind}.
$\therefore$ P(X=1|Y=1)=0.4, P(X=1|Y=0)=0.7, P(Y=1)=0.3.
$\therefore$ on a given shot there is a gust of wind and she hits her target:
P(X=1,Y=1)=$P(X=1|Y=1)\cdot P(Y=1)=0.3\cdot 0.4=0.12$

(ii) she hits the target with her first shot:
P(X=1)=P(X=1,Y=0)+P(X=1,Y=1)=$P(X=1|Y=1)\cdot P(Y=1)+P(X=1|Y=0)\cdot P(Y=0)=0.3\cdot 0.4+0.7\cdot 0.7=0.61$

(iii) she hits the target exactly once in two shots:
P(hits the target exactly once in two shots)=$P(X_1=1,X_2=0)+P(X_1=0,X_2=1)=2\cdot 0.39\cdot 0.61=0.4758$

(iv) there was no gust of wind on an occasion when she missed:
P(Y=0|X=0)=$\frac{P(X=0|Y=0)\cdot P(Y=0)}{P(X=0)}=\frac{0.3\cdot 0.7}{0.39}=\frac{7}{13}=0.53846$

(b) if $P(A|B,C)>P(A|B)$
$P(A|B,C)=\frac{P(A,B,C)}{P(B,C)}>P(A|B)$
$P(A|B,C^c)=\frac{P(A,B,C^c)}{P(B,C^c)}=\frac{P(A,B)-P(A,B,C)}{P(B)-P(B,C)}<\frac{P(A,B)-P(A|B)\cdot P(B,C)}{P(B)-P(B,C)}$
$=\frac{P(A|B)\cdot P(B)-P(A|B)\cdot P(B,C)}{P(B)-P(B,C)}=P(A|B)$
$\therefore P(A|B,C^c)<P(A|B)$

2.Positive Definiteness

(a) (1) first, prove (i) $\Longrightarrow$ (ii):
for a symmetric matrix A:
if $A \succeq 0$, $\forall x \in R^n, x^T Ax \geq 0$
$\because$ for any x, we have $x^T B^T ABX = (BX)^T A(BX)$
$\therefore$ according to definition, let $x_{new} = (Bx)$, we have $x_{new} \in R^n$, and $x_{new}^T Ax_{new} \geq 0$, which means $B^T AB \succeq 0$

Then, prove (ii) $\Longrightarrow$ (i):
if there exists invertible matrix B$\in R^{n*n}$, such that $B^T AB \succeq 0$
so for any x$\in R^n$, $x^T B^T ABx \geq 0$
$\therefore$ for $x_{new} = (Bx)$, when B is invertible matrix,$x \in R^n$, we have $x_{new} = (Bx) \in R^n$
$\because x_{new}^T Ax_{new} \geq 0$, for $x_new \in R^n$
$\therefore A \succeq 0$

(2)Then, to prove (i) $\Longrightarrow$ (iii):
$\because x^T Ax \geq 0$ for any x
$\therefore$ Consider v to be any eigenvector with Av = λv. Then $v^T Av = \lambda v^T v \geq 0$.
$\because$ $v^T v \geq 0$ (v is non-zero), we must have $\lambda \geq 0$.

(3)Then, to prove (iii) $\Longrightarrow$ (iv):
according to spectral theorem, we can obtain $A = H^T DH$ with orthogonal H and diagonal matrix D with eigenvalues on the diagonal.(An orthogonal matrix U satisfies, by definition, $H^T = H^{-1}$)
Let $E = D^{\frac{1}{2}} = E^T$ since all the eigenvalues of A are non-negative.
$\therefore$ we can get A=$H^T EEH = (EH)^T EH$
Let U=EH, so there exists U, satisfy A = $U^T U$

(4) Then, to prove (iv) $\Longrightarrow$ (i):
$\because$ there exists U such that $A = UU^T$
$\therefore$ for any x, we always have $x^T Ax = x^T U^T Ux = (Ux)^T Ux \geq 0$ (since Ux is non-zero)
$\therefore A \succeq 0$

(b) (i) $\because x^T(A + \lambda I)x = x^T Ax + x^T \lambda Ix = x^T Ax + \lambda||x||^2 > 0$
$\therefore$ we have $A + \lambda I \succ 0$

(ii) $\because x^T(A - \gamma I)x = x^T Ax - x^T \gamma Ix = x^T Ax - \gamma||x||^2$
and there must exist $\epsilon > 0$ such that $x^T Ax - \epsilon > 0$
so let $\gamma = \frac{\epsilon}{||x||^2}$, we have $x^T(A - \gamma I)x = x^T Ax - x^T \gamma Ix = x^T Ax - \gamma||x||^2 > 0$
$\therefore$ There exists a $\gamma > 0$ such that $A - \gamma I \succ 0$

(iii) Let $e_1 = (1, 0, 0, \cdots, 0)$ and so on, where e_i is a vector of all zeros, except for a 1 in the ith place.
Since A is positive definite, then $x^T Ax > 0$ for any non-zero vector $x \in R^n$
so we always have $e_i Ae_i^T > 0$, which means $a_{ii} > 0$
So all the diagonal entries of A are positive.

(iv) Let $x^T = (1, 1, 1, 1, 1 \cdots 1)_n$
We have $x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} > 0$, since for any $x \in R^n$, $x^T Ax > 0$

3.Derivatives and Norms

(a) Let $X = (x_1, x_2, \cdots, x_n)^T$, and $a = (a_1, a_2, \cdots, a_n)^T$
Then $X^T a = \sum_{i=1}^{n} x_i a_i$
$\therefore \frac{\partial(X^T a)}{\partial X} = (\frac{\partial(X^T a)}{\partial x_1}, \frac{\partial(X^T a)}{\partial x_2}, \cdots, \frac{\partial(X^T a)}{\partial x_n})^T = (a_1, a_2, \cdots, a_n)^T = a$

(b) Let $A = [a_{ij}]_{n*n}$, so we have:
$x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j$
so for any k, we have :
$\frac{\partial x^T Ax}{\partial x_k} = \frac{\partial}{\partial x_k}(x_1 \sum_{j=1}^{n} a_{1j} x_j + x_2 \sum_{j=1}^{n} a_{2j} x_j + \cdots + x_k \sum_{j=1}^{n} a_{kj} x_j + \cdots + x_n \sum_{j=1}^{n} a_{nj} x_j)$
$= (x_1 a_{1k} + x_2 a_{2k} \cdots + x_n a_{nk}) + (\sum_{j=1}^{n} a_{kj} x_j)$
$= (\sum_{i=1}^{n} a_{ik} x_i) + (\sum_{j=1}^{n} a_{kj} x_j)$
$\therefore$ we have: $\frac{\partial(x^T Ax)}{\partial x} = (A + A^T)x$

(c) Let $A = [a_{ij}]_{n*n}$ and $X = [x_{ij}]_{n*n}$, so we have:
Trace(XA)=$\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} a_{ji}$
$\therefore$ we have:
$\frac{\partial Trace(XA)}{\partial x_{ij}} = \frac{\partial}{\partial x_{ij}}(\sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} a_{ji}) = a_{ji}$
$\therefore \frac{\partial Trace(XA)}{\partial X} = [a_j i]_{n*n} = A^T$

(d) f(x-y) should satisfy $f(x-y) \geq f(x) - f(y)$
Let $f(x) = (\sqrt{x_1} + \sqrt{x_2})^2$, $f(y) = (\sqrt{y_1} + \sqrt{y_2})^2$, $f(x-y) = (\sqrt{x_1 - y_1} + \sqrt{x_2 - y_2})^2$
$\therefore f(x) - f(y) = x_1 + x_2 - y_1 - y_2 + 2\sqrt{x_1 x_2} - 2\sqrt{y_1 y_2}$
$f(x-y) = x_1 + x_2 - y_1 - y_2 + 2\sqrt{(x_1 - y_1)(x_2 - y_2)}$
$\because (\sqrt{x_1 x_2} - \sqrt{y_1 y_2})^2 = x_1 x_2 + y_1 y_2 - 2\sqrt{(x_1 x_2 y_1 y_2)}$, and $(\sqrt{(x_1 - y_1)(x_2 - y_2)})^2 = x_1 x_2 + y_1 y_2 - x_1 y_2 - x_2 y_1$
$\because x_1 y_2 + x_2 y_1 \geq 2\sqrt{x_1 x_2 y_1 y_2}$
$\therefore f(x-y) \leq f(x) - f(y)$ so f(x) is not a norm. counterexample: $x_1 = 9, x_2 = 16, y_1 = 1, y_2 = 4$

(e) $\because ||X||_\infty = max\{|x_1|, |x_2|, \cdots, |x_n|\}$, and $||X||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \geq \sqrt{max\{|x_1|, \cdots, |x_n|\}^2}$
$\therefore ||x||_\infty \leq ||x||_2$
and because $||x||_2 \leq \sqrt{n \cdot max\{|x_1|, \cdots, |x_n|\}^2} = \sqrt{n}||x||_\infty$
$\therefore ||x||_\infty \leq ||x||_2 \leq \sqrt{n}||x||_\infty$

(f) $\because ||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \leq \sqrt{(|x_1| + |x_2| + \cdots + |x_n|)^2}$
$= |x_1| + |x_2| + \cdots + |x_n| = ||x||_1 \therefore ||x_2|| \leq ||x_1||$
and $\because ||x||_1 = (|x_1|, |x_2|, \cdots, |x_n|) \cdot (1, 1, 1, \cdots, 1)^T \leq \sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{n}$(according to Cauchy-Schwarz inequality)
$\therefore ||x||_2 \leq ||x||_1 \leq \sqrt{n}||x||_2$

4.Eigenvalues

(a) $\because$ A is a symmetric matrix with A $\succeq$ 0, and use the spectral theorem for symmetric matrix, we can get $A = H^T DH$ with orthogonal H and diagonal matrix D with eigenvalues on the diagonal.(An orthogonal matrix U satisfies, by definition, $H^T = H^{-1}$)
$\therefore$ $||H||_2 = 1$, and the largest eigenvalue should be the largest element of D, so $\lambda_{max}(A) = max_{||x||_2=1} x^T Ax$

(b) Similarly, $\because$ A is a symmetric matrix with A $\succeq$ 0, and use the spectral theorem for symmetric matrix, we can get $A = H^T DH$ with orthogonal H and diagonal matrix D with eigenvalues on the diagonal.(An orthogonal matrix U satisfies, by definition, $H^T = H^{-1}$)
$\therefore$ $||H||_2 = 1$, and the smallest eigenvalue of A should be the smallest element of D, so $\lambda_{min}(A) = min_{||x||_2=1} x^T Ax$

(c) Yes, they are convex. because $\frac{\partial^2 x^T Ax}{\partial x^2} \geq 0$, so they are convex program.

(d) if λ is an eigenvalue of A, so there is a vector v such that $\lambda v = Av$
$\therefore A^2 v = \lambda Av = \lambda^2 v$, which means λ^2 is an eigenvalue of A^2
since we have already proved in problem 2 that λ is non-negative,
so we can deduce that $\lambda_{max}(A^2) = \lambda_{max}(A)^2$ and $\lambda_{min}(A^2) = \lambda_{min}(A)^2$

(e) $\lambda_{min}(A) = min_{||x||_2=1} x^T Ax \leq ||Ax||_2 = \sqrt{\sum \lambda^2 x^2} \leq \lambda_{max}(A) = max_{||x||_2=1} x^T Ax$

(f) $\because ||x||_2 = 1$
$\therefore \lambda_{min}(A) = \lambda_{min}(A)||x||_2 \leq ||Ax||_2 \leq \lambda_{max}(A)||x||_2 = \lambda_{max}(A)$

5.Gradient Descent

(a) Let $\frac{\partial(\frac{1}{2}x^TAx-b^Tx)}{\partial x} = Ax - b = 0$
$\therefore x^* = A^{-1}b$

(b) w :arbitrary nonzero starting point (good choice is any $A^{-1}b$)
$R(w) > 0$
V : set of indices i for which $(A_iw - b_i) \cdot w < 0$
w: w+$\sum_{i \in V}(A_iw - b_i)$ return w

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

(d) $||x^{(k)} - x^*||_2 = ||(I - A)(x^{(k-1)} - x^*)||_2$
$\because$ A is a symmetric matrix with $0 < \lambda_min(A)$ and $\lambda_{max}(A) < 1$
$\therefore ||x^{(k)} - x^*||_2 = ||(I - A)(x^{(k-1)} - x^*)||_2 \leq \rho||(x^{(k-1)} - x^*)||_2$

(e) according to do d ,$||x^{(0)} - x^*||_2 * \rho^k = ||x^{(k)} - x^*||_2 \leq \epsilon$
$\therefore k \geq log_\rho \frac{\epsilon}{||x^{(0)} - x^*||_2}$

(f) the running time should be $T(n) = n \cdot log_\rho \frac{\epsilon}{||x^{(0)} - x^*||_2}$

6.

(a) Define $\lambda_{ij} l(f(x) = i, y = j)$ The risk of classifying a new data point as class i is:
$R(\alpha_i|x) = \sum_j \lambda_{ij} P(\omega_j|x) = \lambda_s(1 - P(\omega_i|x))$
and the risk of classifying the new data point as doubt is:
$R(\alpha_{c+1}|x) = \lambda_r \sum_j P(\omega_j|x) = \lambda_r$
For choosing doubt to be better than choosing any of the classes, the ratio of the risks must satisfy:
$\frac{R(\alpha_{c+1}|x)}{R(\alpha_i|x)} = \frac{\lambda_r}{\lambda_s(1-P(\omega_i|x))} < 1$
$\therefore P(\omega_i|x) < 1 - \frac{\lambda_r}{\lambda_s}$
$\therefore$ any particular i for which $P(\omega_i|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$ should not be assigned doubt

(b) If $\lambda_r = 0$, then doubt will always be assigned, since for all i, $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s = 1$ is not satisfied unless $P(\omega_i|x) = 1$.
If $\lambda_r > \lambda_s$, then doubt will never be assigned, since for all i, $P(\omega_i|x) \geq 0 > 1 - \lambda_r/\lambda_s$.

7.Gaussian Classification

(a) $\because P(\omega_1|x) = P(\omega_2|x) \rightarrow P(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2) \rightarrow P(x|\omega_1) = P(x|\omega_2) \rightarrow N(\mu_1, \sigma^2) = N(\mu_2, \sigma^2) \rightarrow (x-\mu_1)^2 = (x-\mu_2)^2 \rightarrow x = \frac{\mu_1+\mu_2}{2}$
$\therefore$ The decision rule is to select ω_1 if $x < \frac{\mu_1+\mu_2}{2}$ and otherwise ω_2.

(b) $\because P_e = \frac{1}{2}\int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} N(\mu_2, \sigma^2) + \frac{1}{2}\int_{\frac{\mu_1+\mu_2}{2}}^{\infty} N(\mu_1, \sigma^2)du = 1 - \phi(\frac{\mu_2-\mu_1}{2\sigma})$, where ϕ $N(0,1)$
$\therefore P_e = 1 - \phi(\frac{\mu_2-\mu_1}{2\sigma}) = \int_a^\infty \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz$, where $a = \frac{\mu_2-\mu_1}{2\sigma}$

8.Maximum Likelihood Estimation

$P(x_1, x_2, \cdots, x_n) = \frac{n!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3}$
$\therefore l(x_1, x_2, \cdots, x_n) = log(P(x_1, x_2, \cdots, x_n)) = constant + k_1 log(p_1) + k_2 log(p_2) + k_3 log(p_3)$
$= constant + k_1 log(p_1) + k_2 log(p_2) + (n - k_1 - k_2) log(1 - p_1 - p_2)$
since $k_1 + k_2 + k_3 = n,\ p_1 + p_2 + p_3 = 1$
$\therefore$ let $\frac{\partial l(x_1, x_2, \cdots, x_n)}{\partial p_1} = \frac{k_1}{p_1} + \frac{k_1 + k_2 - n}{1 - p_1 - p_2} = 0,\ \frac{\partial l(x_1, x_2, \cdots, x_n)}{\partial p_2} = \frac{k_2}{p_2} + \frac{k_1 + k_2 - n}{1 - p_1 - p_2} = 0$
$\therefore p_1 = \frac{k_1(1 - p_2)}{n - k_2},\ p_2 = \frac{k_2(1 - p_1)}{n - k_1}$
$\therefore$ we get $\hat{P}_1 = \frac{k_1}{n}, \hat{P}_2 = \frac{k_2}{n}, \hat{P}_3 = \frac{k_3}{n}$, which is the MLE of p_1, p_2, p_3