CS 189 HW01

Hanze Yao
SID: 3033093286

1. Properties of Gaussians

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

$$E(e^{\lambda x}) = \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x+\lambda\sigma^2)^2 - \lambda^2\sigma^4}{2\sigma^2}} dx$$

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

$$= e^{\frac{\lambda^2\sigma^2}{2}} = e^{\frac{\sigma^2\lambda^2}{2}}$$

(b) Use Markov inequality

$P(X \geq a) \leq \frac{E(X)}{a}$ (when X is nonnegative and $a > 0$)

$P(X \geq t) \leq \frac{E(X)}{t}$

$$P(X \geq t) = P\left(e^{\frac{t}{\sigma^2}X} \geq e^{\frac{t}{\sigma^2}x \cdot t}\right) \leq \frac{E\left(e^{\frac{t}{\sigma^2}X}\right)}{e^{\frac{t}{\sigma^2}x \cdot t}} = \frac{e^{\frac{\left(\frac{t}{\sigma^2}\right)^2 \times \sigma^2}{2}}}{e^{\frac{t^2}{\sigma^2}}} = \frac{e^{\frac{t^2}{2\sigma^2}}}{e^{\frac{t^2}{\sigma^2}}} = e^{-\frac{t^2}{2\sigma^2}}$$

This is because $\frac{t}{\sigma^2} > 0$, then $e^{\frac{t}{\sigma^2}x}$ is an increasing function.

$\because X \sim N(0, \sigma^2) \quad \therefore f(x) = f(-x)$

$\therefore P(X \geq t) = P(X \leq -t)$

$\therefore P(X \leq -t \text{ or } X \geq t) \leq 2P(x \geq t) = 2e^{-\frac{t^2}{2\sigma^2}}$

$$P(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

(c) $\because X_1, \cdots, X_n \overset{iid}{\sim} N(0, \sigma^2) \quad \therefore \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(0, \frac{n\sigma^2}{n^2}\right)$

$N\left(0, \frac{\sigma^2}{n}\right)$

$$P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\times\frac{\sigma^2}{n}}\right) = \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

as $n \to \infty \quad P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) \to 0$

(d) First, need to prove that U_X and V_X are uncorrelated.

Need to prove that $E(U_X V_X) = E(U_X)\,E(V_X)$

$U_X = \langle U, X \rangle = U^T X \sim N(0, \|U\|_2^2)$

$V_X = \langle V, X \rangle = V^T X \sim N(0, \|V\|_2^2)$

$E(U_X V_X) = E\left((U_1X_1 + U_2X_2 + \cdots + U_dX_d)(V_1X_1 + V_2X_2 + \cdots + V_dX_d)\right)$

$$= E\left(u_1 v_1 x_1^2 + u_2 v_2 x_2^2 + \cdots + u_d v_d x_d^2 + \sum_{i \neq j} u_i v_j x_i x_j\right)$$

$\because X_1 \cdots X_d$ iid $N(0, 1)$

$\therefore E(X_1^2) = E(X_2^2) = \cdots = E(X_d)^2 = Var(X) - (E(X))^2 = 1 - 0^2 = 1$

$E\left(\sum_{i \neq j} u_i v_j x_i x_j\right) = \sum_{i \neq j} u_i v_j E(X_i) E(X_j) = 0$

$\therefore E(u_x v_x) = u_1 v_1 + u_2 v_2 + \cdots + u_d v_d = \langle u, v \rangle$

$\because u \perp v \quad \therefore E(u_x v_x) = 0$

$\because E(u_x) = E(v_x) = 0$

$\therefore E(u_x v_x) = E(u_x) E(v_x)$

$\therefore u_x$ and v_x are uncorrelated

$\because$ Jointly normal random variables are independent iff they are uncorrelated

$\therefore u_x$ and v_x are independent.

2. Identities with expectation

(a) $f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$

$E(X^k) = \int_0^\infty x^k \lambda e^{-\lambda x} dx = -x^k e^{-\lambda x}\Big|_0^\infty + \int_0^\infty k x^{k-1} e^{-\lambda x} dx$

$= 0 - \frac{1}{\lambda} e^{-\lambda x} k x^{k-1}\Big|_0^\infty$

$= \underbrace{-\frac{1}{\lambda} e^{-\lambda x} k x^{k-1}\Big|_0^\infty}_{=0} + \int_0^\infty \frac{1}{\lambda} e^{-\lambda x} k(k-1) x^{k-2} dx$

$= \int_0^\infty \frac{1}{\lambda} e^{-\lambda x} k(k-1) x^{k-2} dx$

When this integration by parts process is over,

$E(X^k) = \int_0^\infty \frac{1}{\lambda^{k-1}} e^{-\lambda x} k! \, dx$

$= \frac{k!}{\lambda^{k-1}} \int_0^\infty e^{-\lambda x} dx$

$= \frac{k!}{\lambda^{k-1}} \times \left(-\frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty \right) = \frac{k!}{\lambda^k}$

(b) $X = \int_0^X dt = \int_0^\infty 1\{x \ge t\} dt$

$E(X) = E\left[\int_0^\infty 1\{x \ge t\} dt\right] = \int_0^\infty E[1\{x \ge t\}] dt$

$= \int_0^\infty P(x \ge t) dt$

(c) $P(X > 0) = E(1\{x > 0\})$

Use Cauchy-Schwartz inequality

$(E(\underbrace{X \times 1\{x>0\}}))^2 \le E(X^2) \, E(\underbrace{(1\{x>0\})^2})$

$\because X \ge 0$

$\therefore X1\{x>0\} = X, \; x > 0$

$X1\{x>0\} = 0, \; x = 0$

$\therefore X1\{x>0\} = X$

$\because (1\{x>0\})^2$

$\because 1\{x>0\}$ can only be 0 or 1

$\therefore E((1\{x>0\})^2)$ is the same as $E(1\{x>0\})$

$\therefore (E(X))^2 \le E(X^2) E(1\{x>0\}) \Rightarrow [E(X)]^2 \le E(X^2) P(X>0) \Rightarrow P(X>0) \ge \frac{[E(X)]^2}{E(X^2)}$

3. Gradients and Norms

(a) $\|x\|_2 = \sqrt{\sum_{j=1}^{d} |x_j|^2}$ $\quad \|x\|_2^2 = \sum_{j=1}^{d} |x_j|^2$

$\|x\|_1 = \sum_{j=1}^{d} |x_j|$ $\quad \|x\|_1^2 = \sum_{j=1}^{d} |x_j|^2 + 2\sum_{i \neq j} |x_i||x_j| \geq \|x\|_2^2$

$\therefore \|x\|_2 \geq 0$ and $\|x\|_1 \geq 0$ $\therefore$ Can prove $\|x\|_2 \leq \|x\|_1$ by proving $\|x\|_2^2 \leq \|x\|_1^2$

Now use Cauchy-Schwarz inequality

$\|x\|_1 = \langle |x|, \underbrace{1}_{\text{is a } \begin{pmatrix} \vdots \\ \vdots \end{pmatrix}_{d \times 1} \text{ vector}} \rangle \leq \sqrt{\langle |x|, |x| \rangle}\sqrt{\langle 1, 1 \rangle}$

$= \|x\|_2 \times \sqrt{d} = \sqrt{d}\|x\|_2$

(b) (i) $\alpha = \sum_{j=1}^{d} y_j \ln \beta_j$

$$\frac{\partial \alpha}{\partial \beta_j} = \left(\frac{\partial \alpha}{\partial \beta_1} \quad \frac{\partial \alpha}{\partial \beta_2} \quad \cdots \quad \frac{\partial \alpha}{\partial \beta_d} \right) = \left(\frac{y_1}{\beta_1} \quad \frac{y_2}{\beta_2} \quad \cdots \quad \frac{y_d}{\beta_d} \right)$$

(ii)
$$\begin{pmatrix} \frac{\partial \beta_1}{\partial \gamma_1} & \frac{\partial \beta_1}{\partial \gamma_2} & \cdots & \frac{\partial \beta_1}{\partial \gamma_d} \\ \frac{\partial \beta_2}{\partial \gamma_1} & \frac{\partial \beta_2}{\partial \gamma_2} & \cdots & \frac{\partial \beta_2}{\partial \gamma_d} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \beta_d}{\partial \gamma_1} & \frac{\partial \beta_d}{\partial \gamma_2} & \cdots & \frac{\partial \beta_d}{\partial \gamma_d} \end{pmatrix} = \begin{pmatrix} \cosh(\gamma_1) & & & 0 \\ & \cosh(\gamma_2) & & \\ & & \ddots & \\ 0 & & & \cosh(\gamma_d) \end{pmatrix}$$

This is a diagonal matrix

(iii) Let $A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & & \vdots \\ A_{d1} & A_{d2} & \cdots & A_{dd} \end{pmatrix}$ $\quad \rho = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_m \end{pmatrix}$ $\quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix}$

$$\gamma = \begin{pmatrix} \sum_{j=1}^{m} A_{1j}\rho_j + b_1 \\ \sum_{j=1}^{m} A_{2j}\rho_j + b_2 \\ \vdots \\ \sum_{j=1}^{m} A_{dj}\rho_j + b_d \end{pmatrix} \therefore \begin{pmatrix} \frac{\partial \gamma_1}{\partial \rho_1} & \frac{\partial \gamma_1}{\partial \rho_2} & \cdots & \frac{\partial \gamma_1}{\partial \rho_m} \\ \frac{\partial \gamma_2}{\partial \rho_1} & \frac{\partial \gamma_2}{\partial \rho_2} & \cdots & \frac{\partial \gamma_2}{\partial \rho_m} \\ \vdots & \vdots & & \vdots \\ \frac{\partial \gamma_d}{\partial \rho_1} & \frac{\partial \gamma_d}{\partial \rho_2} & \cdots & \frac{\partial \gamma_d}{\partial \rho_m} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & & \vdots \\ A_{d1} & A_{d2} & \cdots & A_{dm} \end{pmatrix}$$

(c) $X, A \in R^{d \times d}$

$\nabla_X Tr(A^T X)$

$f(X) = Tr(A^T X)$

$f(X+\Delta) = Tr(A^T(X+\Delta))$

$= Tr(A^T X + A^T \Delta) = Tr(A^T X) + Tr(A^T \Delta)$

$= f(X) + Tr(A^T \Delta)$

$\therefore f(X+\Delta) = f(X) + Tr\left(\frac{\partial f}{\partial X} \Delta\right)$

$\therefore \frac{\partial f}{\partial X} = A^T \quad \therefore \nabla_X Tr(A^T X) = \left(\frac{\partial f}{\partial X}\right)^T = (A^T)^T = A$

(d) $Xw = y$

$\|y - Xw\|_2^2 = (y - Xw)^T (y - Xw)$

$= (y^T - w^T X^T)(y - Xw)$

$= y^T y - y^T X w - w^T X^T y + w^T X^T X w$

$= y^T y - 2 y^T X w + w^T X^T X w$

$\nabla_w \|y - Xw\|_2^2 = (-2 y^T X)^T + (X^T X + X^T X) w$

$= 2 X^T X w - 2 X^T y$

$\therefore w^* = \text{argmin}_w \|y - Xw\|_2^2$

$\therefore \nabla_w \|y - X w^*\|_2^2 = 0 \Rightarrow 2 X^T X w^* - 2 X^T y = 0$

$2 X^T X w^* = 2 X^T y \qquad X^T X w^* = X^T y$

$\because X$ is full rank $\therefore X^T X$ is full rank as well

$\therefore w^* = (X^T X)^{-1} X^T y$

4. Linear Algebra Review

(a) First: prove (i) $\Rightarrow$ (ii): $X^TAX = X^T(U\Lambda U^T)X = (X^TU)\Lambda(U^TX)$

Let $Z = U^TX$ $\quad = Z^T\Lambda Z = \sum_{i=1}^{d}\lambda_i Z_i^2 \geq 0$

Then all λ_i must be non negative. Which is: All eigenvalues of A are non-negative.

Second: prove (ii) $\Rightarrow$ (iii): If all A's eigenvalues are non negative. Then $A = V\Lambda V^T = V\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}V^T$

$\because \Lambda = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$ and $\lambda_1, \lambda_2, \cdots, \lambda_d \geq 0$

$\therefore \Lambda^{\frac{1}{2}} = \text{diag}(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \cdots, \lambda_d^{\frac{1}{2}})$

$\because \Lambda^{\frac{1}{2}}$ is a diagonal matrix $\therefore (\Lambda^{\frac{1}{2}})^T = \Lambda^{\frac{1}{2}}$

$\therefore A = (V(\Lambda^{\frac{1}{2}})^T)(\Lambda^{\frac{1}{2}}V^T) = (V(\Lambda^{\frac{1}{2}})^T)(V(\Lambda^{\frac{1}{2}})^T)^T$

$\therefore U = V(\Lambda^{\frac{1}{2}})^T$

Third: prove (iii) $\Rightarrow$ (i): According to eigen decomposition, $A = V\Lambda V^T$

$\therefore A = UU^T = V(\Lambda^{\frac{1}{2}})^T(V(\Lambda^{\frac{1}{2}})^T)^T = V(\Lambda^{\frac{1}{2}})^T\Lambda^{\frac{1}{2}}V$

$\therefore V(\Lambda^{\frac{1}{2}})^T\Lambda^{\frac{1}{2}}V = A = V\Lambda V^T \therefore (\Lambda^{\frac{1}{2}})^T = \Lambda^{\frac{1}{2}}$ $\because \Lambda$ is a diagonal matrix with A's eigenvalues on the diagonal $\therefore \Lambda^{\frac{1}{2}}$ must be $\text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_d})$ $\therefore \lambda_1, \lambda_2 \cdots \lambda_d \geq 0$ $\therefore \sum_{i=1}^{d}\lambda_i X_i^2 \geq 0$ which is $X^TAX \geq 0$ for all $X \in R^d$

(b)

(i) $\because A$ and B are PSD

$\therefore A$ and B's eigenvalues λ_{iA} and λ_{iB} are all non-negative

$\therefore 2A+3B$'s eigenvalues $= 2\lambda_{iA} + 3\lambda_{iB}$

$\therefore 2A+3B$'s eigenvalues are all non-negative as well

$\therefore 2A+3B$ is PSD

(ii) According to a(iii), There exists a matrix $U \in R^{d\times d}$, $A = UU^T$

Assume $U = \begin{pmatrix} -u_1- \\ -u_2- \\ \vdots \\ -u_d- \end{pmatrix}$ $\therefore A = \begin{pmatrix} -u_1- \\ -u_2- \\ \vdots \\ -u_d- \end{pmatrix}\begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_d \\ | & | & & | \end{pmatrix}$

Then diagonal entries of A are $u_1^Tu_1$, $u_2^Tu_2$, $\cdots$, $u_d^Tu_d$ which are $\|u_1\|_2^2, \|u_2\|_2^2, \cdots, \|u_d\|_2^2$.

Obviously they are non-negative

(iii) According to a(i), for all $X \in R^d$ $X^TAX \geq 0$

If $= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{d\times 1}$ Then $X^TAX = (1 \cdots 1)A\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \geq 0$

$\sum_{i=1}^{d}\sum_{j=1}^{d}A_{ij} \geq 0$

Because $Tr(AB) = Tr(BA)$

(iv) $Tr(AB) = Tr(A B^{\frac{1}{2}} B^{\frac{1}{2}}) = Tr(B^{\frac{1}{2}} A B^{\frac{1}{2}})$

$= Tr(B^{\frac{1}{2}} A^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{2}})$

$\because$ A, B are PSD matrix $\therefore$ A, B are symmetric

$\therefore A^{\frac{1}{2}}, B^{\frac{1}{2}}$ are symmetric

$\therefore Tr(AB) = Tr(B^{\frac{1}{2}} A^{\frac{1}{2}} A^{\frac{1}{2}} B^{\frac{1}{2}}) = Tr(B^{\frac{1}{2}} A^{\frac{1}{2}} (B^{\frac{1}{2}} A^{\frac{1}{2}})^T)$

According to a(iii) $B^{\frac{1}{2}} A^{\frac{1}{2}} (B^{\frac{1}{2}} A^{\frac{1}{2}})^T$ is a PSD matrix

$\therefore Tr(AB) = Tr(B^{\frac{1}{2}} A^{\frac{1}{2}} (B^{\frac{1}{2}} A^{\frac{1}{2}})^T) =$ sum (eigenvalues of $B^{\frac{1}{2}} A^{\frac{1}{2}} (B^{\frac{1}{2}} A^{\frac{1}{2}})^T$)

≥ 0

(v) According to (iv) $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is a PSD matrix,

$\therefore B^{\frac{1}{2}} A B^{\frac{1}{2}}$ is a symmetric matrix, so it is diagonalizable

If $Tr(AB) = 0$, then $Tr(B^{\frac{1}{2}} A B^{\frac{1}{2}}) = 0$

$\because$ All eigenvalues of $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ are non-negative

$\therefore$ All eigenvalues of $B^{\frac{1}{2}} A B^{\frac{1}{2}}$ are 0.

$\therefore B^{\frac{1}{2}} A B^{\frac{1}{2}} = P \begin{pmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{pmatrix} P^{-1}$

↳ a zero matrix.

$\therefore AB = B^{-\frac{1}{2}} (B^{\frac{1}{2}} A B^{\frac{1}{2}}) B^{\frac{1}{2}} = B^{-\frac{1}{2}} P \begin{pmatrix} 0 & & \\ & \ddots & \\ & & 0 \end{pmatrix} P B^{\frac{1}{2}} = 0$

↳ a zero matrix

When $AB = 0$, of course $Tr(AB) = 0$

$\therefore$ If A and B are PSD, then $Tr(AB) = 0$ if and only if $AB = 0$

(c) According to the definition of eigenvalue

$\lambda_A X = AX$ $\because \lambda_A$ is a scalar

$\therefore X \lambda_A = AX$

$X^T X \lambda_A = X^T A X$

If $\|X\|_2 = 1$ Then $X^T X = 1$

$\therefore \lambda_A = X^T A X$

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

5. Covariance Practice

$C = E[(z-\mu)(z-\mu)^T]$

For $X \in \forall R^d$

$E(\|Z-\mu)^T X\|_2^2) = E([(z-\mu)^T X]^T[(z-\mu)^T X]) \geq 0$

$E(X^T(z-\mu)(z-\mu)^T X] \geq 0$

$X^T[E(z-\mu)(z-\mu)^T]X \geq 0$

$X^T C X \geq 0$ for all $X \in R^d$

$\therefore$ C is PSD $\Rightarrow$ The covariance matrix is always positive semi-definite.

6. A simple Classification Approach.

(b) The residual error is : $\|Xw - y\|_2^2 = 422.75085$

The first 20 entries of w are as follows:

-0.33027663, 0.39177105, 0.14818856, -0.16060987
0.10327045, -0.01870318, -0.12769328, 0.00945555
-0.01714958, -0.00567482, -0.00068246, -0.01126736
-0.00570973, 0.00458835, 0.01787667, -0.03010068
0.01003883, -0.06487594, -0.02304357, -0.02630404

(c) Training set percentage of correctly classified: 0.9976

Test set percentage of correctly classified: 0.8881

(d) The performance evaluated on a separate test set is to avoid over-fitting problem.

Performance is similar in our case is because there are only two classes and we want to separate them using linear regression. The handwritten numbers 0 and 1 are very easy to classify.

(e) With 0/1 target

Training set percentage of correctly classified: 0.9897

Test set - - - - - - - - - - : 0.9915

After adding bias term

With -1/1 target

Training = 0.9941

Test = 0.9962

With 0/1 target

Training = 0.9941

Test = 0.9962

After adding this bias term means the linear model has an interception.

For 0/1 target, the hyperplane actually needs an intercepion so after adding the bias term its accuracy increases.

For -1/1 target, the hyperplane does not need an intercepion. so after adding the bias term its accuracy decreases a little bit.

After adding the bias term, the result of the two kinds of target become the same.