آرین شکر- تکلیف شماره هشت درس شناسایی آماری الگو

5.12. From the definitions of S_w and S_b in the textbook we have:

$$S_w = \sum_{i=1}^{M} P_i\, E_i[(x-\mu_i)(x-\mu_i)^T] \quad , \quad S_b = \sum_{i=1}^{M} P_i (\mu_i-\mu_0)(\mu_i-\mu_0)^T$$

$$S_m = E[(x-\mu_0)(x-\mu_0)^T]$$

where E_i is $E_{x \in \omega_i}$

$$\rightarrow S_w + S_b = \sum_{i=1}^{M} P_i \left[E_i[(x-\mu_i)(x-\mu_i)^T] + (\mu_i-\mu_0)(\mu_i-\mu_0)^T \right]$$

if we write $x-\mu_i$ as $x-\mu_0+\mu_0-\mu_i$ (so that it connects the expression to both S_b and S_m) we get:

$$S_w + S_b = \sum_{i=1}^{M} P_i \left[E_i[(x-\mu_0+\mu_0-\mu_i)(x-\mu_0+\mu_0-\mu_i)^T] + (\mu_i-\mu_0)(\mu_i-\mu_0)^T \right]$$

$$= \sum_{i=1}^{M} P_i \left[E_i\left[(x-\mu_0)(x-\mu_0)^T + 2(x-\mu_0)(\mu_0-\mu_i)^T + (\mu_0-\mu_i)(\mu_0-\mu_i)^T\right] + (\mu_i-\mu_0)(\mu_i-\mu_0)^T \right]$$

Expectation of this term is zero because $E(x-\mu_0)=0$ and $(x-\mu_0)$ and $(\mu_0-\mu_i)$ are independent

this term does not depend on x and is therefore a constant w.r.t. the expectation

(~~[illegible]~~ sum over all values of $(\mu_i-\mu_0)$ is 0)

$$\Rightarrow S_w + S_b = \sum_{i=1}^{M} P_i \left[E_i[(x-\mu_0)(x-\mu_0)^T] + \cancel{\left((\mu_0-\mu_i)(\mu_0-\mu_i)^T + (\mu_i-\mu_0)(\mu_i-\mu_0)^T\right)} \right]$$

$$= \sum_{i=1}^{M} P_i\, E_i[(x-\mu_0)(x-\mu_0)^T] = E[(x-\mu_0)(x-\mu_0)^T] = \underline{\underline{S_m}}$$

5.13. we have $\rho_{ij} = \dfrac{\sum_{n=1}^{N} x_{ni}\, x_{nj}}{\sqrt{\sum_{n=1}^{N} x_{ni}^2 \sum_{n=1}^{N} x_{nj}^2}}$, the cauchy-schwarz inequality states that for two vectors x and y we have:

$$|x^T y| \leq \|x\|\|y\|$$

we define $x = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ni} \end{bmatrix}$ and $y = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{Nj} \end{bmatrix} \rightarrow |x^T y| = \left| \sum_{n=1}^{N} x_{ni}\, x_{nj} \right| \leq \|x\|\|y\|$

$$= \sqrt{\sum_{n=1}^{N} x_{ni}^2}\sqrt{\sum_{n=1}^{N} x_{nj}^2}$$

$\xrightarrow[\text{sides by } \|x\|\|y\|]{\text{dividing both}}$

$$\frac{\left| \sum x_{ni}\, x_{nj} \right|}{\sqrt{\sum_{n=1}^{N} x_{ni}^2 \sum_{n=1}^{N} x_{nj}^2}} = \underline{\underline{|\rho_{ij}| \leq 1}}$$

5.14. we know that for two same-covariance distributions.

$$d_{ij} = (\mu_i-\mu_j)^T \Sigma^{-1} (\mu_i-\mu_j)$$ since there are only two classes:

$$d_{12} = (\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2)$$

We will now calculate $S_w^{-1} S_b$:

$$S_w = \sum_{i=1}^{2} P_i \Sigma_i = \frac{1}{2}\Sigma + \frac{1}{2}\Sigma = \Sigma \longrightarrow S_w^{-1} = \Sigma^{-1}$$

$$S_b = \sum_{i=1}^{2} P_i (\mu_i-\mu_0)(\mu_i-\mu_0)^T = \frac{1}{2}(\mu_1-\mu_0)(\mu_1-\mu_0)^T + \frac{1}{2}(\mu_2-\mu_0)(\mu_2-\mu_0)^T$$

using $\mu_0 = \sum_{i=1}^{M} P_i \mu_i \longrightarrow S_b \frac{1}{2}\left(\frac{1}{4}(\mu_1-\mu_2)(\mu_1-\mu_2)^T + \frac{1}{4}(\mu_2-\mu_1)(\mu_2-\mu_1)^T\right)$

$$= \frac{1}{4}(\mu_1-\mu_2)(\mu_1-\mu_2)^T$$

$$\rightarrow S_w^{-1} S_b = \frac{1}{4}\Sigma^{-1}(\mu_1-\mu_2)(\mu_1-\mu_2)^T \rightarrow trace\{S_w^{-1} S_b\} = \frac{1}{4} trace\{\underbrace{\Sigma^{-1}}_{A}\underbrace{(\mu_1-\mu_2)}_{B}\underbrace{(\mu_1-\mu_2)^T}_{C}\}$$

$$= \frac{1}{4} trace\{\underbrace{(\mu_1-\mu_2)^T}_{C}\underbrace{\Sigma^{-1}}_{A}\underbrace{(\mu_1-\mu_2)^{\cancel{T}}}_{B}\}$$

trace for the product of three matrices is invariant under cyclic permutations

on the other hand because d_{12} is a scalar we have

$$trace\{(\mu_1-\mu_2)^T \Sigma^{-1} (\mu_1-\mu_2)\}$$

$$= (\mu_1-\mu_2)\Sigma^{-1}(\mu_1-\mu_2) = d_{12}$$

$$trace\{S_w^{-1} S_b\} = \underline{\frac{1}{4} d_{12}}$$

5.17. From the matrix cookbook we know the following:

$$\partial/\partial X \; trace\{XA\} = \partial/\partial X \; trace\{AX\} = A^T \qquad (eq\ 100)$$

$$\partial/\partial X \; trace\{X^T B X\} = (B+B^T)X \qquad (eq\ 108)$$

(where C is symmetric)

$$\partial/\partial X \; trace\{(X^T C X)^{-1} A\} = -(CX(X^T C X)^{-1})(A+A^T)(X^T C X)^{-1} \qquad (eq\ 125)$$

the task is to prove (eq 126) using these presuppositions.

$$\frac{\partial}{\partial A} \text{trace}\{(A^T S_1 A)^{-1}(A^T S_2 A)\} = \frac{\partial\, \text{trace}}{\partial A}\{(A^T S_1 A)^{-1} C_1\} + \frac{\partial\, \text{trace}}{\partial A}\{C_2 (A^T S_2 A)\}$$

where C_1 and C_2 are constants w.r.t. their derivatives

and $C_1 = A^T S_2 A$, $C_2 = (A^T S_1 A)^{-1}$

→ we use eq. 125 an eq. 117 from the matrix cookbook:

$$\frac{\partial}{\partial A} \text{trace}\{(A^T S_1 A)^{-1} C_1\} = -(S_1 A (A^T S_1 A)^{-1})(C_1 + C_1^T)(A^T S_1 A)^{-1}$$

Substituting C_1 and the fact that $C_1 + C_1^T = 2C_1$

$$\rightarrow \frac{\partial}{\partial A} \text{trace}\{(A^T S_1 A)^{-1} C_1\} = -2S_1 A (A^T S_1 A)^{-1} (A^T S_2 A)(A^T S_1 A)^{-1}$$

$$\frac{\partial}{\partial A} \text{trace}\{C_2 (A^T S_2 A)\} = \frac{\partial}{\partial A} \text{trace}\{(A^T S_2 A) C_2\} \xrightarrow{\text{eq } 117}$$

$$= S_2 A C_2 + S_2^T A C_2^T \xrightarrow[C_2 = C_2^T]{S_2 = S_2^T} 2 S_2 A C_2 = 2 S_2 A (A^T S_1 A)^{-1}$$

$$\rightarrow \frac{\partial}{\partial A} \text{trace}\{(A^T S_1 A)^{-1}(A^T S_2 A)\} = -2S_1 A (A^T S_1 A)^{-1} (A^T S_2 A)(A^T S_1 A)^{-1} + 2S_2 A (A^T S_1 A)^{-1}$$

5.19. As the hint suggests we will define $M-1$ functions $g_i(x) := f_i(x) - f_{i+1}(x)$ $\forall 1 \leq i \leq M-1$

these $M-1$ $g(\cdot)$ functions are enough for classification.

We prove that these $M-1$ $g(\cdot)$ functions are enough for classification based on the discriminant functions $f_i(\cdot)$s

In order to make a classification based on the discriminant functions $f_i(\cdot)$s we assign x to class(i) if $i = \underset{j}{\text{argmax}}\, f_j(x)$ or $f_i(x)$ is the maximum among all possible is. This means $M-1$ equalities must hold:

$$\forall_{1 \leq j < i} \quad f_i(x) > f_j(x)$$

$$\forall_{i+1 \leq k \leq M} \quad f_i(x) > f_k(x)$$

we can construct these inequalities using $g_i(\cdot)$ as follows:

for two arbitrary indices $1 \le i, j \le M$, $i \neq j$

we want to prove $f_i(x) > f_j(x)$ or $f_i(x) - f_j(x) > 0$

if $j < i$:

$$f_i(x) - f_j(x) > 0 \rightarrow f_j(x) - f_i(x) < 0$$

$$\rightarrow \cancel{g_j(x)} \; (f_j(x) - f_{j+1}(x)) + (f_{j+1}(x) - f_{j+2}(x))$$

$$+ \dots + (f_{i-1}(x) + f_i(x))$$

$$= \sum_{k=j}^{i-1} g_k(x) < 0$$

if $j > i$:

$$f_i(x) - f_j(x) > 0$$

$$\rightarrow (f_i(x) - f_{i+1}(x)) + (f_{i+1}(x) - f_{i+2}(x))$$

$$+ \dots + (f_{j-1}(x) + f_j(x)) = \sum_{k=i}^{j-1} g_k(x) > 0$$

$\longrightarrow$ $M-1$ functions $(g_i(\cdot))$ are enough for classification.

5.20. $S_b = \sum_{i=1}^{2} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T$. Substituting $\mu_0 = \sum_{i=1}^{2} P_i \mu_i$ and expanding we have:

$$S_b = P_1(P_2(\mu_1 - \mu_2)P_2(\mu_1 - \mu_2)^T) + P_2(P_1(\mu_1 - \mu_2)P_1(\mu_1 - \mu_2))^T$$

$= P_1 P_2 (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \rightarrow S_b$ is a scalar therefore it is of rank $\underline{1}$

$\rightarrow S_w^{-1} S_b$ has exactly one non-zero eigenvalue

$\rightarrow$ Sum of the eigen values of $S_w^{-1} S_b$ is equal to its only non-zero eigenvalue

↳ trace

$$\rightarrow \lambda = \text{trace}\{P_1 P_2 \underbrace{S_w^{-1}}_{l \times l} \underbrace{(\mu_1 - \mu_2)}_{l \times 1} \underbrace{(\mu_1 - \mu_2)^T}_{1 \times l}\} = P_1 P_2 (\mu_1 - \mu_2)^T S_w^{-1} (\mu_1 - \mu_2)$$

produces a scalar which is equal to the trace if $S_w^{-1}(\mu_1-\mu_2)$ is an eigenvector we must have:

$$\underbrace{P_1 P_2 (\mu_1-\mu_2)^T S_w^{-1} (\mu_1-\mu_2)}_{\lambda} \left(S_w^{-1}(\mu_1-\mu_2)\right) = (S_w^{-1} S_b)\left(S_w^{-1}(\mu_1-\mu_2)\right)$$

$$P_1 P_2 S_w^{-1} (\mu_1-\mu_2)(\mu_1-\mu_2)^T \left(S_w^{-1}(\mu_1-\mu_2)\right)$$

$$= P_1 P_2 (\mu_1-\mu_2)^T S_w^{-1}(\mu_1-\mu_2)\, S_w^{-1}(\mu_1-\mu_2) = \underline{\lambda S_w^{-1}(\mu_1-\mu_2)}$$

5.21. let A diagonalize both Σ_1 and Σ_2: $A^T\Sigma_1 A = I \rightarrow \Sigma_1 = A^{-T}A^{-1}$

$A^T \Sigma_2 A = D \rightarrow \Sigma_2 = A^{-T} D A^{-1}$

$$\rightarrow \Sigma_1^{-1}\Sigma_2 = (A^{-T}A^{-1})^{-1}(A^{-T}DA^{-1}) = (AA^T)(A^{-T}DA^{-1}) = ADA^{-1}$$

let $\lambda_i \neq \lambda_j$ be two distinct eigenvalues of $\Sigma_1^{-1}\Sigma_2$:

$$ADA^{-1}v_i = \lambda_i v_i \rightarrow D\underbrace{A^{-1}v_i}_{v'_i} = \lambda_i \underbrace{A^{-1}v_i}_{v'_i} \rightarrow \begin{cases} Dv'_i = \lambda_i v'_i \\ Dv'_j = \lambda_j v'_j \end{cases}$$

$$ADA^{-1}v_j = \lambda_j v_j \rightarrow D\underbrace{A^{-1}v_j}_{v'_j} = \lambda_j \underbrace{A^{-1}v_j}_{v'_j}$$

$\rightarrow$ v'_i and v'_j are eigenvectors of D. Since D is a diagonal matrix it is also symmetrical and the eigenvectors of a symmetrical matrix are orthogonal $\longrightarrow$ $v_i'^T v'_j = 0$ (or δ_{ij} if $i=j$ is allowed)

$$\Rightarrow (A^{-1}v_i)^T (A^{-1}v_j) = v_i^T \underbrace{A^{-T}A^{-1}}_{\Sigma_1} v_j = v_i^T \Sigma_1 v_j = 0$$

$\underline{(\text{or } \delta_{ij})}$