Notes

for slide 7

the means in projected space for the two classes.

$$\mu_1 = \frac{1}{N_1} \sum_{k \in C_1} z_k \qquad \mu_2 = \frac{1}{N_2} \sum_{k \in C_2} z_k$$

goal: find w that maximizes the differences of means.

$$\arg\max_w |\mu_2(w) - \mu_1(w)|, \quad \|w\| = 1$$

$$= \arg\max_w \left| \frac{1}{N_2} \sum_{k \in C_2} z_k(w) - \frac{1}{N_1} \sum_{k \in C_1} z_k(w) \right|$$

$$= \arg\max_w \left| \frac{1}{N_2} \sum_{k \in C_2} w^T \cdot x_k - \frac{1}{N_1} \sum_{k \in C_1} w^T x_k \right|$$

$$= \arg\max_w \left| w^T \cdot \left(\frac{1}{N_2} \sum_{k \in C_2} x_k - \frac{1}{N_1} \sum_{k \in C_1} x_k \right) \right|$$

$$= \arg\max_w \left| w^T \cdot (\mu_2 - \mu_1) \right|$$

conclusion: the best vector w is the one that aligns with $(\mu_2 - \mu_1)$, i.e.

$w = (\mu_2 - \mu_1) / \|\mu_2 - \mu_1\|$, $\quad w^T \cdot (\mu_2 - \mu_1) = \left(\frac{d}{\|d\|}\right)^T \cdot d = \frac{d^T \cdot d}{\|d\|} = \frac{\|d\|^2}{\|d\|} = \|d\|$

$d = \mu_2 - \mu_1$

Proof:

$d = \mu_2 - \mu_1$

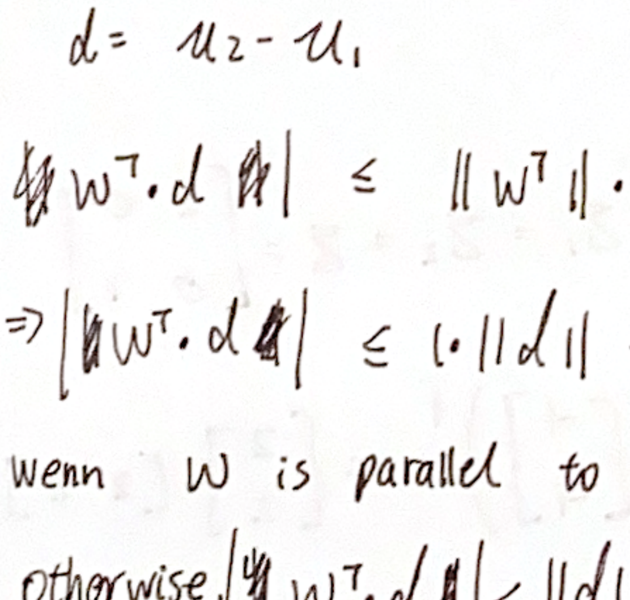
$|w^T \cdot d| \le \|w^T\| \cdot \|d\|, \quad \|w\| = 1$

$\Rightarrow |w^T \cdot d| \le 1 \cdot \|d\| = \|d\|$

wenn w is parallel to d, then the equation holds $|w^T \cdot d| = \|d\|$,

otherwise, $|w^T \cdot d| < \|d\|$

geometric meaning illustration



$|w_1 \cdot d| = |(-3, -1) \cdot (3, 1)| = |-9 - 1| = 10$

$|w_2 \cdot d| = |(1, -3) \cdot (3, 1)| = |3 - 3| = 0$

Slides 8

limitation of mean seperation

$w = \frac{u_2 - u_1}{\| u_2 - u_1 \|}$, w points to the direction connecting the two class centroids.

the class means are as far apart as possible after projection, but it not garantee good class seperability.

the reasons are :

1) significant class overlap in projected space
2) the method ignores class variances, the orientation of the clauds ; the shape of the distributions.

If we rotate the projection direction slightly (not using $u_2 - u_1$) ; the projected distributions became more seperated.

→ Motivation for Fischer Discriminant

slide 11

Fisher Discriminant

$$u_1 = \frac{1}{|C_1|}\sum_{k\in C_1} z_k \qquad u_2 = \frac{1}{|C_2|}\sum_{k\in C_2} z_k$$

$$S_1 = \sum_{k\in C_1}(z_k - u_1)^2 \qquad S_2 = \sum_{k\in C_2}(z_k - u_2)^2$$

$$J(W) = \underset{w}{\operatorname{argmax}} \frac{(u_2(W) - u_1(W))^2}{S_1(W) + S_2(W)}$$

$$= \frac{W^T S_B W}{W^T S_W W}$$

proof

where $S_B = (u_2 - u_1)(u_2 - u_1)^T$

$S_W = S_1 + S_2$

Why $(W^T V)^2 = W^T V V^T W$

Proof:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in R^{dx1}, \quad V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \in R^{dx1}$$

$$W^T \cdot V = [w_1 \ w_2 .. \ w_d] \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix}$$

$$= \sum_{i=1}^{d} w_i v_i = [w_1 v_1 \quad w_2 v_2 \quad \cdots \quad w_d v_d]$$

$$(W^T V)^2 = \left(\sum_{i=1}^{d} w_i v_i\right)^2$$

$$V \cdot V^T = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix} \cdot [v_1 \ v_2 \ldots v_d] = \begin{bmatrix} v_1v_1 & v_1v_2 & \ldots & v_1v_d \\ v_2v_1 & v_2v_2 & \ldots & v_2v_d \\ \vdots & & & \\ v_dv_1 & v_dv_2 & \ldots & v_dv_d \end{bmatrix}$$

$$W^T \cdot (V V^T) W = (W^T V) \cdot (V^T W)$$

$$V^T \cdot W = [v_1 \ v_2 .. \ v_d] \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = [v_1 w_1 \quad v_2 w_2 \ldots \quad v_d w_d]$$

$$\Rightarrow W^T (V V^T) W = (W^T V)^2$$

Herleitung von S_j

$$S_j(W) = \sum_{k\in C_j} (z_k - u_j)^2$$

$$= \sum_{k\in C_j} (W^T x_k - W^T u_j)^2$$

$$= W^T \underbrace{\sum_{k\in C_j} (x_k - u_j)(x_k - u_j)^T}_{S_j} W = W^T S_j W$$

S_j is a scatter matrix / unnormalized covariance matrix for the data of class j.

等式后转换:

$$u_j = \frac{1}{|C_1|}\sum_{k\in C_1} z_k = \frac{1}{|C_1|}\sum_{k\in C_1} W^T x_k = W^T u_j$$

$$u_2(W) - u_1(W) = W^T(u_2 - u_1)$$

$$(u_2(W) - u_1(W))^2 = (W^T(u_2 - u_1))^2$$

$$= W^T (u_2 - u_1)(u_2 - u_1)^T W$$

← * reason see next page. detail explanation about how to calculate.

$$S_j(W) = W^T S_j W$$

$$S_1(W) = W^T S_1 W, \quad S_2(W) = W^T S_2 W$$

$$S_1(W) + S_2(W) = W^T (S_1 + S_2) W = W^T S_W W$$

slide 13

$$\mathcal{L}(w,\lambda) = w^T S_B w + \lambda(1 - w^T S_w w)$$

explain why $\frac{\partial}{\partial w}(w^T A w) = 2Aw$.

where A is symmetric , $A = A^T$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in R^{d \times 1} \quad , A = (a_{ij})$$

$$w^T A w = \sum_{i=1}^{d} \sum_{j=1}^{d} w_i a_{ij} w_j \quad , \text{because}$$

$$[w_1\ w_2\ \dots\ w_d] \cdot \begin{bmatrix} a_{11} & a_{12} \cdots & a_{1d} \\ a_{21} & a_{22} \cdots & a_{2d} \\ \vdots & & \\ a_{d1} & \cdots\cdots & a_{dd} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$= \begin{bmatrix} \overset{1}{0} & \overset{2}{0} \dots & \overset{d}{0} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \text{a scalar}$$

$$w^T A w = \sum_{i}^{d} \sum_{j=1}^{d} w_i a_{ij} w_j \quad , \text{observe } w_k$$

case 1 : $i = k$

$\Rightarrow$ $w_k a_{kj} w_j$

$$\frac{2}{2w_k}(w_k a_{kj} w_j) = a_{kj} w_j$$

sum over all j $\quad \sum_j a_{kj} w_j$

case 2 : $j = k$

$\Rightarrow$ $w_i a_{ik} w_k$ $\quad \frac{2}{2w_k}(w_i a_{ik} w_k) = a_{ik} w_k$

sum over all i $\quad \sum_i a_{ik} w_k$

A is symmetric $\Rightarrow$ $a_{ij} = a_{ji}$

$$\Rightarrow \frac{\partial}{\partial w}(w^T A w) = 2Aw$$

slide 14

explain how from

$$S_B w = \lambda S_w w$$

~~$S_w^{-1} S_B w = \lambda w$~~ to $(S_w^{-1} S_B) w = \lambda w$,

where S_w is invertable.

$$S_B w = \lambda S_w w$$

$$S_w^{-1} S_B w = S_w^{-1} \lambda S_w w$$

$$= \lambda S_w^{-1} S_w w$$

$$= \lambda I w$$

$$= \lambda w$$

Slide 25

$\nabla_w \xi_k(w,b)$ 是如何计算出来的

$z_k = w^T x_k + b$, $y_k = \text{sign}(z_k)$

$$t_k = \begin{cases} 1 & \text{, the true class of } x_k \text{ is } w_1 \\ -1 & \text{, } \text{-}\text{-}\text{-}\text{-}\text{-}\text{-} \text{ } w_2 \end{cases}$$

$$\xi(w,b) = \frac{1}{N} \sum_{k=1}^{l} \underbrace{\max(0, -z_k t_k)}_{\xi_k(w,b)}$$

$$-z_k t_k = -w^T x_k t_k - b t_k$$

$$\frac{\partial \xi_k(w,b)}{\partial w} = \begin{cases} -x_k t_k & , y_k \neq t_k \\ 0 & , y_k = t_k \end{cases}$$