Σ symmetric positive-definite. (e.g. covariance matrix).

eigen-decomposition:

$$\Sigma = V D V^T.$$

$$D = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix}.$$ $\lambda_{(1)} \geq \lambda_{(2)}, \cdots \geq \lambda_{(n)} \geq 0$ eigen-values.

$$V = \begin{pmatrix} v_1^T & v_2^T & \cdots & v_n^T \end{pmatrix}.$$ $v_1 \cdot v_2 \cdots v_n$ are eigen-vectors.

Properties:

(1) $v_i \perp v_j$ $\quad (\langle v_i \cdot v_j \rangle = 0).$

(2). $\Sigma^{\frac{1}{2}} = V D^{\frac{1}{2}} V^T$ $\quad \left(\because \Sigma^{\frac{1}{2}} \cdot \Sigma^{\frac{1}{2}} = V D^{\frac{1}{2}} \underbrace{V^T V}_{= I} D^{\frac{1}{2}} V^T = V D V^T = \Sigma \right).$

(3). Suppose $cov(X) = \Sigma$;

then $Var(\alpha X) = \alpha \Sigma \alpha^T.$

Consider $\max\limits_{\alpha} \; Var(\alpha X) = \alpha \Sigma \alpha^T$ subject to $|\alpha| = 1$,

the solution is $\alpha^* = v_1$ and $Var(\alpha^* X) = v_1 \cdot V D V^T v_1^T$

$$= \lambda_1$$

$$f(X|Y=k) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(X-\mu_k)^T \Sigma_k^{-1} (X-\mu_k)}.$$

Linear discriminant analysis (LDA): $\Sigma_k = \Sigma$, $\forall k$.

$$C(X) = \arg\max_{k} \left(-\log\left((2\pi)^{p/2} |\Sigma|^{1/2} \right) - \frac{1}{2}(X-\mu_k)^T \Sigma^{-1} (X-\mu_k) + \log \pi_k \right).$$

$$= \arg\max_{k} \left(-\frac{1}{2}(X-\mu_k)^T \Sigma^{-1} (X-\mu_k) + \log(\pi_k) \right).$$

note: $\frac{1}{2}(X-\mu_k)^T \Sigma^{-1} (X-\mu_k) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \frac{1}{2} x^T \Sigma^{-1} x$

$$= \arg\max_{k} \left(x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) \right).$$

$$= \arg\min_{k} \left(\mu_k^T \Sigma^{-1} \mu_k - 2 x^T \Sigma^{-1} \mu_k \right).$$

uniform prior: $\pi_k = \pi$, $\forall k$

Decision boundary between class l and m:

$$(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) - 2 X^T \Sigma^{-1} (\mu_k - \mu_l) = 0.$$

The decision boundaries are hyper-planes.

Computation (LDA): (1) estimate $\hat{\Sigma}$

(2). $\hat{\Sigma} = U D U^T$ (eigen-decomposition).

D: diagonal matrix with eigen-values.

(3). Sphere the data $D^{-\frac{1}{2}} U^T X \rightarrow X^*$

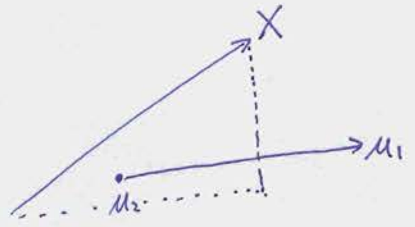
$D^{-\frac{1}{2}} U^T \mu_k \rightarrow \mu_k^*$

(4). Classify x^* to the closest centroid μ_k^*

LDA as a dimension reduction method.

- decision boundary: $-\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + x^T\Sigma^{-1}(\mu_1-\mu_2) = 0$.
- only the projection of X on the direction. $\Sigma^{-1}(\mu_1-\mu_2)$ matters.

Suppose. $\Sigma = \sigma^2 I$ is sphered. Only the projection of X onto. $\mu_1-\mu_2$ is needed.



$K=2 \Rightarrow 1D.$ $\quad H_1$

$K=3 \Rightarrow 2D.$ $\quad H_2$

$K=4 \Rightarrow 3D.$ $\quad H_3$.

$\vdots$

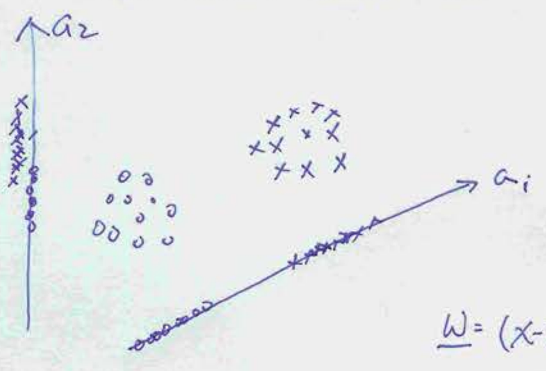
Fisher's Optimization Criteria

— When $K>3$, we might want to. find a subspace $H_L \subseteq H_{K-1}$ optimal for LDA in some sense.

Idea:.

$$\max_a \frac{a^T B a}{a^T W a}.$$

B: between-class variance
W: within-class variance.
in original spa



$\underline{W} = (x-\bar{x})^T(x-\bar{x})$.

$(ax-\overline{ax})^T(ax-\overline{ax})$

$= a^T(x-\bar{x})(x-\bar{x})a$.

within class variance in projected space. $= \underline{a^T W a}$.

$W = V D V^T. \Rightarrow W^{\frac{1}{2}} = D^{\frac{1}{2}} V^T$

Define $b = W^{\frac{1}{2}} a$.

$$\Rightarrow \max_b \frac{b^T (W^{-\frac{1}{2}})^T B W^{-\frac{1}{2}} b}{b^T b}.$$

Define $B^* = (W^{-\frac{1}{2}})^T B W^{-\frac{1}{2}}$.

Eigen-decompose. $B^* = V^* D^* V^{*T}$

We know maximization is achieved by.

$b_1 = v_1^*$, the first eigen vector of B^*.

Similarly, one can find the next direction. $b_2 = v_2^*$.

s.t. $b_2 \perp b_1$ and. maximizes $b_2^T B^* b_2 / b_2^T b_2$.

Computation of Fisher Optimization.

— find centroids of all classes and calculate.

between-class covariance matrix $\hat{B}$.

within-class covariance matrix $\hat{W}$.

— $\hat{W} = \hat{V} \hat{D} \hat{V}^T$

— $\hat{B}^* = (\hat{W}^{-\frac{1}{2}})^T \hat{B} \hat{W}^{-\frac{1}{2}}$

— $\widehat{B^*} = \widehat{V^*} \widehat{D^*} \widehat{V^*}^T$

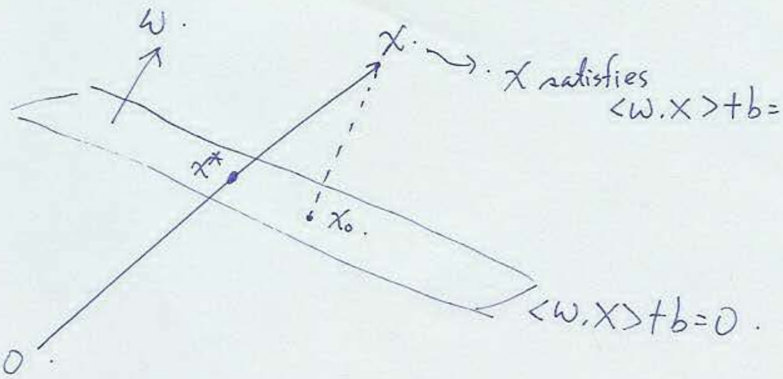
— $a_1 = \widehat{W}^{-\frac{1}{2}} v_1^*$

$a_2 = \widehat{W}^{-\frac{1}{2}} v_2^*$

$\vdots$

Support Vector Machines.

$$x \in \mathbb{R}^p, \quad f(x) = \langle \overset{w\cdot x}{\overset{\shortparallel}{w}} . x \rangle + b$$



$$x^* = cx \qquad w^T x + b = 1. \qquad w^T x^* + b = 0. \qquad c\,w^T x + b = 0.$$

$$\Rightarrow w^T x = \frac{-b}{c} = 1-b.$$

$$\therefore x^* = \frac{-b}{1-b} \cdot x$$

$$\left|\overrightarrow{x x_0}\right| = \langle \overrightarrow{x x^*}, w \rangle \frac{1}{|w|}$$

$$= \langle x - \frac{-b}{1-b} x, w \rangle \frac{1}{|w|}$$

$$= \langle \frac{1}{1-b} x, w \rangle \frac{1}{|w|}$$

$$= \frac{1}{1-b} \cdot 1-b \cdot \frac{1}{|w|} = \frac{1}{|w|}.$$

---

maximizing margin:

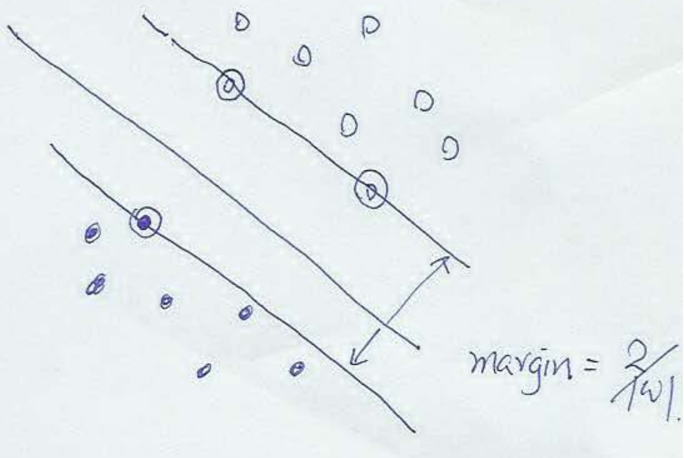
$$\min \frac{1}{2} \|w\|^2.$$

under the condition that.

$$\boxed{\begin{cases} w^T x_i + b \geq 1 & \text{for } y_i = +1 \\ w^T x_i + b \leq -1 & \text{for } y_i = -1 \end{cases}}$$

$$\Downarrow$$

equivalent to. $\quad y_i \left[(w^T x_i) + b \right] - 1 \geq 0. \quad i = 1, 2, \cdots n.$



Kuhn-Tucker Theorem.

$$L(\omega, \alpha) = f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega)$$

Given an optimization problem:

$$\min f(\omega)$$

subject to $g_i(\omega) \le 0$. $i = 1, 2, \cdots k$.

Then

$$\frac{\partial L(\omega^*, \alpha^*)}{\partial \omega} = 0.$$

$$\frac{\partial L(\omega^*, \alpha^*)}{\partial \alpha} = 0.$$

$$\alpha_i^* g_i(\omega^*) = 0, \quad i = 1, \cdots k$$

$$g_i(\omega^*) \le 0, \quad i = 1, \cdots k.$$

$$\alpha_i^* \ge 0, \quad i = 1, \cdots k.$$

$$\min \frac{1}{2} \langle \omega, \omega \rangle.$$

subject to $y_i (\langle \omega, x_i \rangle + b) \ge 1$.

$$L(\omega, b, \alpha) = \frac{1}{2} \langle \omega, \omega \rangle - \sum_{i=1}^{n} \alpha_i \left[y_i (\langle \omega \cdot x_i \rangle + b) - 1 \right].$$

where $\alpha_i \ge 0$ are the Lagrange multipliers.

$$\frac{\partial L(\omega^*, b^*, \alpha^*)}{\partial \omega} = \omega^* - \sum_{i=1}^{n} y_i \alpha_i^* x_i = 0. \quad \Rightarrow \quad \omega^* = \sum_{i=1}^{n} y_i \alpha_i^* x_i$$

$$\frac{\partial L(\omega^*, b^*, \alpha^*)}{\partial b} = \sum_{i=1}^{n} y_i \alpha_i^* = 0.$$

$$\alpha_i^* \left[y_i (\langle \omega^* \cdot x_i \rangle + b^*) - 1 \right] = 0 \quad \text{———} \quad (1). \quad \left(\begin{array}{l} y_i(\langle \omega^*, x_i \rangle + b^*) > 1 \\ \Rightarrow \alpha_i^* = 0 \end{array} \right)$$

classifier

$$f(x) = \langle \omega^* \cdot x \rangle + b^*$$

all the n data points

$$= \sum_{i=1}^{n} y_i \alpha_i^* \langle x_i \cdot x \rangle + b^*$$

$$= \sum_{i \in SV} y_i \alpha_i^* \langle x_i \cdot x \rangle + b^*$$

(from (1), only the support vectors have non-zero α_i^*)