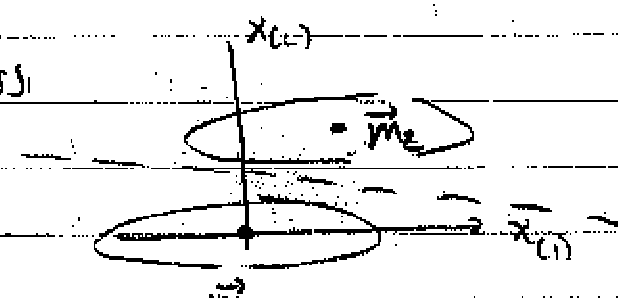
544NA
9/26/2013

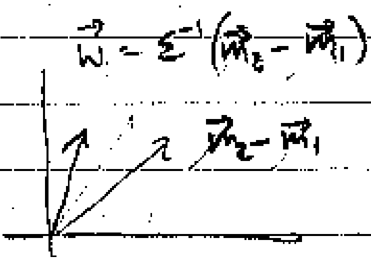
TODAY: FISHER'S LINEAR DISCRIMINANT

$$\boxed{\begin{aligned} W &= \text{eigenvectors}\left(S_W^{-1} S_B\right) \\ S_W &= \sum_{k=1}^{c} \sum_{n: t_k^n = 1} (x^n - \hat{\mu}_k)(x^n - \hat{\mu}_k)^T \\ S_B &= \sum_{k=1}^{c} \sum_{n: t_k^n = 1} (\hat{\mu}_k - \hat{\mu})(\hat{\mu}_k - \hat{\mu})^T \end{aligned}}$$

DERIVATION #1: LDA = BAYES OPTIMAL CLASSIFIER, GAUSSIANS w/ SAME COVARIANCE

TWO-CLASS





$$p(\vec{x} \mid C_1) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x} - \vec{m}_1)^T \Sigma^{-1} (\vec{x} - \vec{m}_1)}$$

$$p(\vec{x} \mid C_2) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{x} - \vec{m}_2)^T \Sigma^{-1} (\vec{x} - \vec{m}_2)}$$

$$p(C_1 \mid \vec{x}) = \frac{p(\vec{x} \mid C_1) \Pr(C_1)}{p(\vec{x} \mid C_1) \Pr(C_1) + p(\vec{x} \mid C_2) \Pr(C_2)}$$

$$= \frac{1}{1 + e^{\frac{1}{2}\left((\vec{x} - \vec{m}_1)^T \Sigma^{-1} (\vec{x} - \vec{m}_1) - (\vec{x} - \vec{m}_2)^T \Sigma^{-1} (\vec{x} - \vec{m}_2)\right) - \ln\left(\Pr(C_1)/\Pr(C_2)\right)}}$$

$$= \frac{1}{1 + e^{\vec{x}^T \vec{w} - w_0}}$$

$$\vec{w} = \Sigma^{-1}(\vec{m}_2 - \vec{m}_1), \quad w_0 = \vec{w}^T\left(\frac{\vec{m}_1 + \vec{m}_2}{2}\right) + \ln\left(\frac{Pr(C_2)}{Pr(C_1)}\right)$$

MULTI-CLASS

$$Pr(C_\ell|\vec{x}) = \frac{1}{1 + e^{\vec{x}^T w_{k\ell} - w_{0k\ell}} + e^{\vec{x}^T w_{\ell 3} - w_{0\ell 3}} + \cdots}$$

$$\vec{w}_{k\ell} = \Sigma^{-1}(\vec{m}_\ell - \vec{m}_k), \quad w_{0k\ell} = \vec{w}_{k\ell}^T\left(\frac{\vec{m}_k + \vec{m}_\ell}{2}\right) + \ln\left(\frac{Pr(C_\ell)}{Pr(C_k)}\right)$$

ALTERNATE DEFINITION

$$\vec{w}_k = \Sigma^{-1}\vec{m}_k$$

$$Pr(C_k|\vec{x}) = \frac{1}{1 + e^{\vec{x}^T(\vec{w}_2 - \vec{w}_k) - w_{0k2}} + e^{\vec{x}^T(\vec{w}_3 - \vec{w}_k) - w_{0k3}} + \cdots}$$

BAYES OPTIMAL CLASSIFIER

① COMPUTE $\vec{y} = W\vec{x}$, $W: [(C-1) \times M]$

$$= \left(\Sigma^{-1}[\vec{m}_2 - \vec{m}_1, \ldots, \vec{m}_C - \vec{m}_1]\right)^T$$

② LINEAR CLASSIFIER IN THE $(C-1)$-DIMENSIONAL $\vec{y}$ SPACE

WHAT IF YOU DON'T KNOW Σ, $\vec{\mu}_k$?

ML ESTIMATES

$$\hat{\Sigma}_k = \frac{1}{N_k - 1} \sum_{n: t_k^n = 1} (\vec{x}^n - \vec{m}_k)(\vec{x}^n - \vec{m}_k)^T \qquad N_k = \text{\# TOKENS IN CLASS } k$$

$$\hat{\Sigma} = \frac{1}{N - C} \sum_{k=1}^{C} (N_k - 1) \hat{\Sigma}_k = \frac{1}{N-C} S_W$$

$$S_W = \sum_{k=1}^{C} \sum_{n: t_k^n = 1} (\vec{x}^n - \vec{m}_k)(\vec{x}_n - \vec{m}_k)^T$$

$$\vec{m}_k = \frac{1}{N_k} \sum_{n: t_k^n = 1} \vec{x}^n$$

C−1 DIMENSIONS: $W = \left(\Sigma^{-1} \left[(\vec{m}_2 - \vec{m}_1), \dots (\vec{m}_C - \vec{m}_1) \right] \right)^T$

IF $\vec{m}_1$ IS BAD, W SUFFERS.

SOLUTION: "BETWEEN-CLASS COVARIANCE"

BALANCES ALL ESTIMATION ERRORS

HOW: TREAT $\vec{m}_k$ AS AN RV:

$$p(\mu_k) = \frac{1}{|2\pi \Sigma_B|^{\frac{1}{2}}} e^{-\frac{1}{2} (\vec{m}_k - \vec{m})^T \Sigma_B^{-1} (\vec{m}_k - \vec{m})}$$

$$\vec{m} = \frac{1}{N} \sum_{k=1}^{C} N_k \vec{m}_k \qquad \text{= GLOBAL MEAN ESTIMATE}$$

ML ESTIMATE OF $\hat{\Sigma}_B = \frac{1}{N} \sum_{k=1}^{C} N_k (\vec{m}_k - \vec{m})(\vec{m}_k - \vec{m})^T$

$$= \frac{1}{N} S_B$$

$$S_B = \sum_{k=1}^{C} N_k (\vec{m}_k - \vec{m})(\vec{m}_k - \vec{m})^T = \underbrace{\sum_{k=1}^{C} \sum_{n: t_k^n = 1} \underbrace{(\vec{m}_k - \vec{m})(\vec{m}_k - \vec{m})^T}_{\text{RANK ONE MATRIX}}}_{\text{RANK} \le C-1}$$

- S_B HAS AT MOST $C-1$ EIGENVECTORS W/ NONZERO EIGENVALUES
- THOSE EIGENVECTORS ARE ORTHONORMAL BASIS ($V^T V = I$) FOR THE SPACE SPANNED BY $[\vec{m}_2 - \vec{m}_1, \ldots, \vec{m}_c - \vec{m}_1]$
- EIGENVECTORS OF $\boxed{\hat{\Sigma}^{-1} \hat{\Sigma}_B = S_W^{-1} S_B}$ ARE ORTHONORMAL BASIS OF $\hat{\Sigma}^{-1} [\vec{m}_2 - \vec{m}_1, \ldots, \vec{m}_c - \vec{m}_1]$

$\Rightarrow$ CHOOSE W = EIGENVECTORS $(S_W^{-1} S_B)$ W/ NONZERO EIGENVALUES

DERIVATION #2: CLASSES NOT GAUSSIAN W/SAME COVARIANCE. MAYBE WE KNOW NOTHING ABOUT PDF

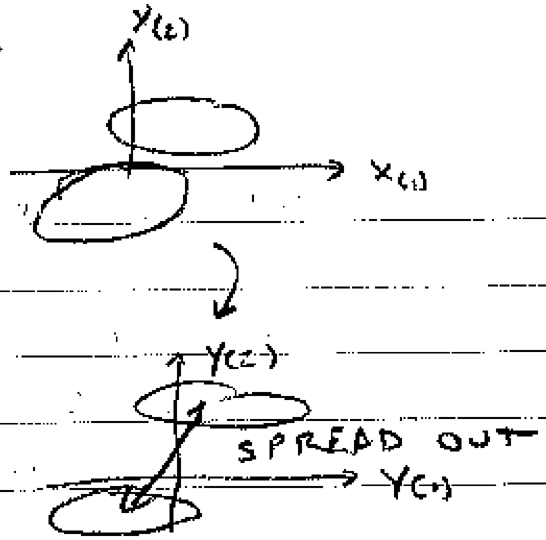
NON-BAYESIAN GOAL: "SPREAD OUT" THE CLASSES AS MUCH AS POSSIBLE

CHOOSE $\boxed{\vec{y} = W\vec{x}}$ SO THAT FOR

$$\vec{\mu}_k = E[\vec{y}^n \mid C_k], \quad \vec{\mu} = E[\vec{y}^n]$$

$$S_B = \sum_{k=1}^{c} N_k (\vec{\mu}_k - \vec{\mu})(\vec{\mu}_k - \vec{\mu})^T \quad \text{LARGE}$$

$$S_W = \sum_{k} \sum_{n: t_k^n = 1} (\vec{y}^n - \vec{\mu}_k)(\vec{y}^n - \vec{\mu}_k)^T \quad \text{SMALL}$$



NOTICE $\vec{y} = W\vec{x}$

$\vec{\mu}_k = W\vec{m}_k$

$\vec{\mu} = W\vec{m}$

$S_B = W S_B W^T$

$S_W = W S_W W^T$

SUPPOSE WE MAXIMIZE

$$J(W) = \frac{\text{trace}(W S_B W^T)}{\text{trace}(W S_W W^T)} = \frac{\sum_{k=1}^{c-1} \vec{w}_k^T S_B \vec{w}_k}{\sum_{k=1}^{c-1} \vec{w}_k^T S_W \vec{w}_k}$$

SUBJECT TO $W = \begin{bmatrix} \vec{w}_1^T \\ \vdots \\ \vec{w}_{c-1}^T \end{bmatrix}$ ORTHONORMAL VECTORS

$(WW^T = I)$

METHOD

① FIND $\vec{w}_1$ TO MAXIMIZE $\dfrac{\vec{w}_1^T S_B \vec{w}_1}{\vec{w}_1^T S_W \vec{w}_1} = J(\vec{w}_1)$

② FIND $\vec{w}_2 \perp \vec{w}_1$ TO MAXIMIZE $\dfrac{\vec{w}_2^T S_B \vec{w}_2}{\vec{w}_2^T S_W \vec{w}_2}$

③ AND SO ON

$$\nabla_{\vec{w}_1} J(\vec{w}) = \frac{\nabla_{\vec{w}}(\vec{w}^T S_B \vec{w})}{\vec{w}^T S_W \vec{w}} - \frac{\vec{w}^T S_B \vec{w}}{(\vec{w}^T S_W \vec{w})^2} \nabla_{\vec{w}}(\vec{w}^T S_W \vec{w})$$

$$= \frac{\cancel{2} S_B \vec{w}}{\cancel{\vec{w}^T S_W \vec{w}}} - \cancel{2} \underbrace{\frac{\vec{w}^T S_B \vec{w}}{(\vec{w}^T S_W \vec{w})^{\cancel{2}}}}_{\text{SCALAR}} \underbrace{S_W}_{\text{ASSUME THIS IS NONSINGULAR}} \vec{w} = 0$$

$$S_W^{-1} S_B \vec{w} - \underbrace{\left(\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}\right)}_{\text{SCALAR}} \vec{w} = 0$$

THIS IS JUST $S_W^{-1} S_B \vec{w} = \lambda \vec{w}$

FOR $\lambda = J(\vec{w})$! TO MAXIMIZE λ,

CHOOSE $\vec{w}_1 = \text{EIGENVECTOR}(S_W^{-1} S_B)$

WITH LARGEST EIGENVALUE

CHOOSE $\vec{W}_2 \perp \vec{W}_1$ AS

$\vec{w}_2 = \text{EIGENVECTOR}(S_W^{-1} S_B)$

W/SECOND LARGEST λ

AND SO ON