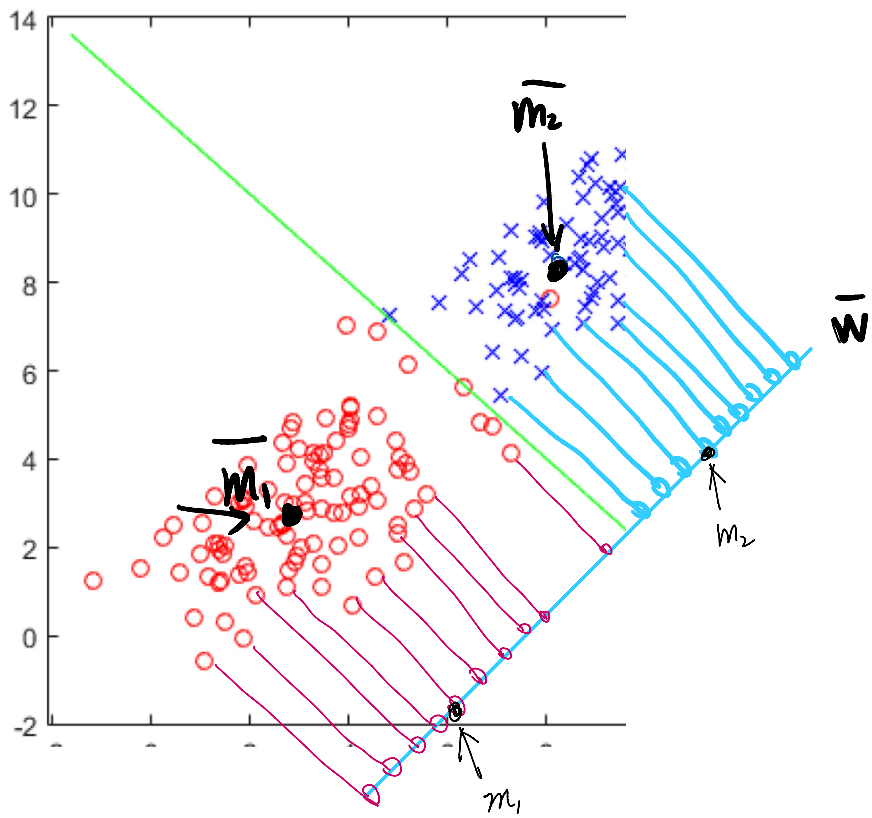
Lecture 17

Fisher's discriminant (cont'd)



$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$$

Since $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})$, $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})$ and $(m_2 - m_1)$ are all scalar factors, we can drop them if we care only about the direction of the weight vector $\mathbf{w}$, instead of its magnitude. Thus we can obtain

$$\underset{n\times 1}{\mathbf{w}} \propto \underset{n\times n}{\mathbf{S}_{\mathrm{W}}^{-1}} \underbrace{(\overline{\mathbf{m}}_2 - \overline{\mathbf{m}}_1)}_{n\times 1}$$

$$\mathbf{S}_{\mathrm{W}} = \sum_{n\in\mathcal{C}_1} (\mathbf{x}_n - \overline{\mathbf{m}}_1)(\mathbf{x}_n - \overline{\mathbf{m}}_1)^{\mathrm{T}} + \sum_{n\in\mathcal{C}_2} (\mathbf{x}_n - \overline{\mathbf{m}}_2)(\mathbf{x}_n - \overline{\mathbf{m}}_2)^{\mathrm{T}}$$

Extension of Fisher's discriminant to multiple classes:

- This criterion can then be rewritten as an explicit function of the projection matrix $\mathbf{W}$ in the form:

$$J(\mathbf{w}) = \mathrm{Tr}\left\{(\mathbf{W}\mathbf{S}_{\mathrm{W}}\mathbf{W}^{\mathrm{T}})^{-1}(\mathbf{W}\mathbf{S}_{\mathrm{B}}\mathbf{W}^{\mathrm{T}})\right\}$$

- It can be shown that the weight values are determined by those eigenvectors of $\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$, which correspond to the D' largest eigenvalues.
- It can be shown $\mathbf{S}_{\mathrm{B}}$ has rank at most equal to $(K-1)$ and so there are at most $(K-1)$ nonzero eigenvalues. So we are therefore unable to find more than $(K-1)$ linear "features".

- LDA
  In general, discriminant analysis assumes that the class
  conditional densities to have multivariate Gaussian distributions.

  the model assumes **the same covariance matrix** for each class only the means vary

  Shrinkage: covariance matrix estimation

  http://www.ledoit.net/honey.pdf

Two Classes As an Example

Decision functions (with a common covariance matrix $\boldsymbol{C}$, where $\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{C}$):

$$d_1(\mathbf{x}) = \ln P(\omega_1) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}[(\mathbf{x}-\mathbf{m_1})^{\mathrm{T}}\mathrm{C}^{-1}(\mathbf{x}-\mathbf{m_1})]$$

$$d_2(\mathbf{x}) = \ln P(\omega_2) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{1}{2}[(\mathbf{x}-\mathbf{m_2})^{\mathrm{T}}\mathrm{C}^{-1}(\mathbf{x}-\mathbf{m_2})]$$

Decision Boundary (assuming equal class probabilities): $d_1(\mathbf{x}) = d_2(\mathbf{x})$

$$(\mathbf{x}-\mathbf{m_1})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m_1}) = (\mathbf{x}-\mathbf{m_2})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m_2})$$

$$\mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{m_1} - \mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{x} + \mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1}$$
$$= \mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{m_2} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{x} + \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2}$$

$$x^T C^{-1} m_1 = (x^T C^{-1} m_1)^T$$
$$\underbrace{1\times n \quad n\times n \quad n\times 1}_{1\times 1}$$
$$= m_1^T (C^{-1})^T (x^T)^T$$
$$= m_1^T C^{-1} x$$

Cancellation due to the assumption of same covariance (LDA); otherwise quadratic function of $\mathbf{x}$, thus QDA results.

$$(\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x} = \frac{1}{2}(\mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2})$$

Thus

$$-x^T C^{-1} m_1 - m_1^T C^{-1} x$$
$$= -2 m_1^T C^{-1} x$$

$$\overbrace{(\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}\mathbf{C}^{-1}}^{w^T}\mathbf{x} = \frac{1}{2}(\mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2})$$

Let weight vector $\mathbf{w} = \mathbf{C}^{-1}(\mathbf{m_1}-\mathbf{m_2})$, then

$$(\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{x} = \mathbf{w}^{\mathrm{T}}\mathbf{x} \quad \text{and}$$

$$\mathbf{w}^{\mathrm{T}}(\mathbf{m_1}+\mathbf{m_2}) = (\mathbf{m_1}-\mathbf{m_2})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{m_1}+\mathbf{m_2})$$

$$= \mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1} + \underbrace{\mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1}}_{=0} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2}$$

$$= \mathbf{m_1^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_1} - \mathbf{m_2^{\mathrm{T}}}\mathbf{C}^{-1}\mathbf{m_2}$$

Thus the decision function is a function of a linear combination of the observations:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} = \frac{1}{2}\mathbf{w}^{\mathrm{T}}(\mathbf{m_1}+\mathbf{m_2}), \text{ where } \mathbf{w} = \mathbf{C}^{-1}(\mathbf{m_1}-\mathbf{m_2})$$

Consistent with Fisher's linear discriminant with projection:
$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2-\mathbf{m}_1)$, where $\mathbf{S}_{\mathrm{W}} = 2\boldsymbol{C}$