Linear Discriminant Functions

Discriminant functions

A discriminant function takes an input vector **x** and assigns it to one of the K classes (C_k)

Linear Discriminant Function for two classes

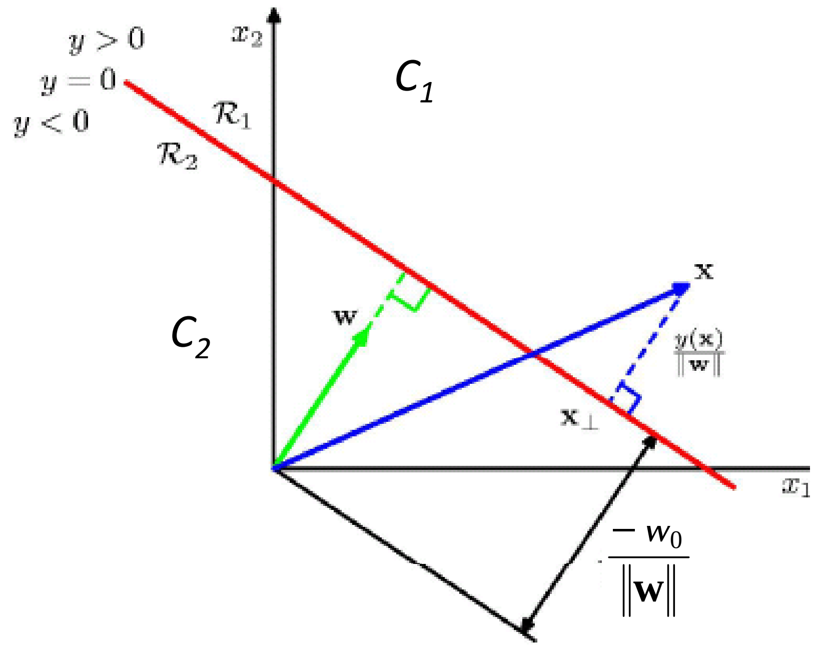
- Two classes

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

if $y(\boldsymbol{x}) \geq 0$, assign to C_1

otherwise, assign to C_2

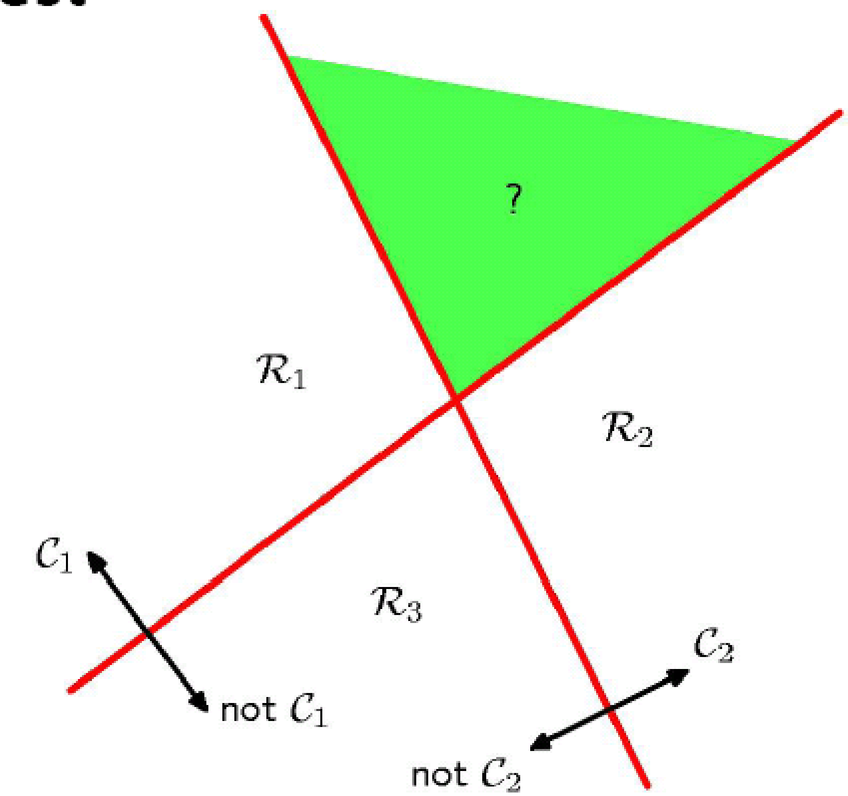
- Decision boundary: $y(\boldsymbol{x})=0$
- Decision boundary is perpendicular to $\boldsymbol{w}$



- The Normal Distance from the origin to the decision boundary is $\dfrac{-w_0}{\|\mathbf{w}\|}$
- Signed distance from the decision boundary to any point **x** is $\dfrac{y(\mathbf{x})}{\|\mathbf{w}\|}$

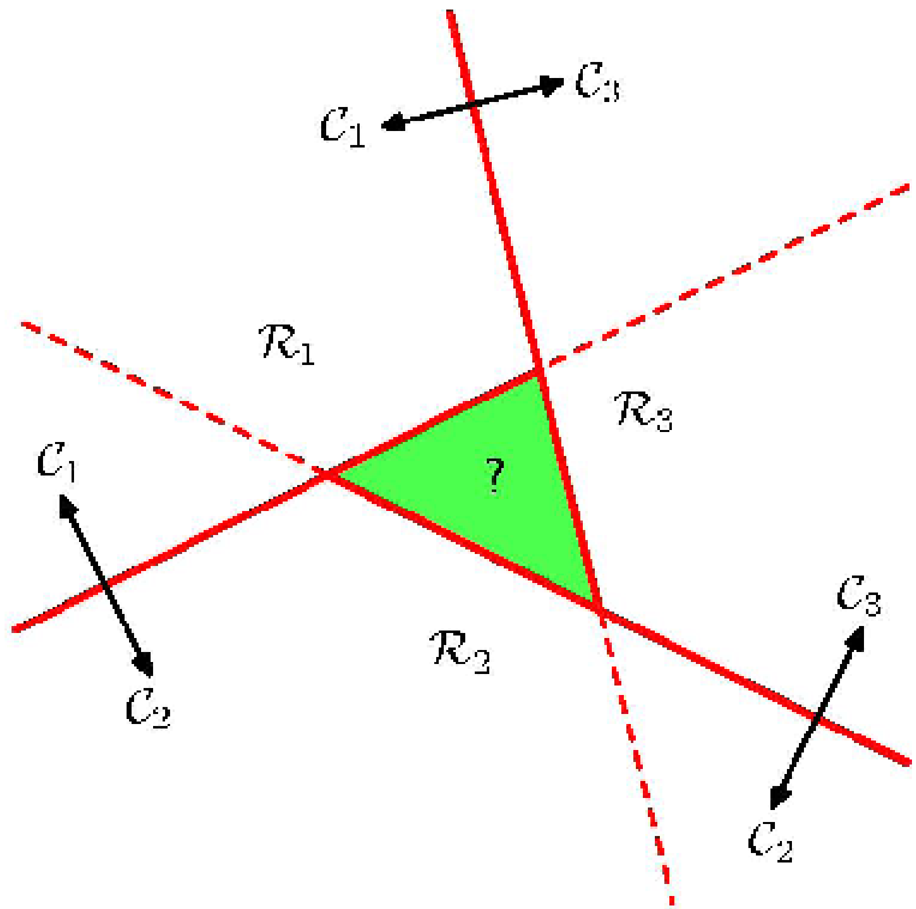
Multiple classes

- One-versus-the-rest
- K-1 discriminant functions



Multiple classes

- One-versus-one
- K(K-1)/ 2 binary discriminant functions.



Multiple classes: solution

- Consider a k-class single discriminant consisiting of k linear functions of the form

$$y_j(\mathbf{x}) = \mathbf{w}_j^T \mathbf{x} + w_{j0}$$

- Assign a point x to class C_k if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$
  - The decision boundary between class i and j is given by

$$y_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} = y_j(\mathbf{x}) = \mathbf{w}_j^T \mathbf{x} + w_{j0}$$

$$\Rightarrow (\mathbf{w}_i - \mathbf{w}_j)^T \mathbf{x} + (w_{i0} - w_{j0}) = 0$$

Learning Linear Discriminants

- We will see two approaches
  - LDA- linear discriminant analysis (fisher's linear discriminant)
  - Perceptron

Fishers discriminant analysis

- One way to view a linear classification model is to see it as dimension reduction
- For two classes case:
  - We project **x** onto a single dimension using projection vector **w**: $y=\mathbf{w}^T\mathbf{x}$
  - Set a threshold t:
    - If y> t, predict class 1
    - Other wise predict class 2

Intuition

- Find a project direction so that the separation between classes is maximized
- In other words, we are looking for a projection that best discriminates different classes

- How to find such project directions?

LDA

- Data: $\{(x_i, c_i) : i = 1,\ldots, N\} \quad x_i \in R^q \quad c_i \in \{c_1, c_2\}$
- N1 samples of c_1
- N2 samples of c_2
- Consider $\mathbf{w} \in R^q$, with the constrain that $\|\mathbf{w}\| = 1$
- Then $\mathbf{w}^T\mathbf{x}$ is the projection of **x** onto the direction of **w**
- we want the projected points of x from c_1 to be well separated from those of x from c_2

LDA

- One way to measure separation is to look at the projected class means

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{\mathbf{x}\in c_1}\mathbf{x} \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{\mathbf{x}\in c_2}\mathbf{x}$$

$$m'_1 = \frac{1}{N_1}\sum_{\mathbf{x}\in c_1}\mathbf{w}^T\mathbf{x} \qquad m'_2 = \frac{1}{N_2}\sum_{\mathbf{x}\in c_2}\mathbf{w}^T\mathbf{x}$$

$$\left|m'_1 - m'_2\right|^2 = \left|\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right|^2$$

  - This quantity is the distance between projected centers, the larger the better

LDA

- We would further like to make the projected points within a class to be as compact as possible
- This is typically measured by **variance**

$$s_i^2 = \sum_{x \in c_i} (\mathbf{w}^T \mathbf{x} - m'_i)^2$$

variance r for the projected class c_i

$$s_1^2 + s_2^2$$

Total within class **variance of the** projected data

$$\boxed{\arg\max_{\mathbf{w}} \frac{|m'_1 - m'_2|^2}{s_1^2 + s_1^2}}$$

Final objective function

LDA

- To optimize this objective function $\arg\max_{\mathbf{w}} \frac{|m'_1 - m'_2|^2}{s_1^2 + s_1^2}$
- We rewrite it by noticing that:

$$|m'_1 - m'_1|^2 = \left(\mathbf{w}^{\mathbf{T}}\mathbf{m}_1 - \mathbf{w}^{\mathbf{T}}\mathbf{m}_2\right)^2 = \mathbf{w}^{\mathbf{T}}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{w}$$

$$= \mathbf{w}^{\mathbf{T}}\mathbf{S}_B\mathbf{w}$$

S_B: ***Between-class*** **covariance (scatter) matrix**

$$s_i^2 = \sum_{x \in c_i} (\mathbf{w}^T\mathbf{x} - m'_i)^2 = \sum_{x \in c_i} (\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{m}_i)^2$$

$$= \sum_{x \in c_i} \mathbf{w}^T(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T\mathbf{w} = \mathbf{w}^T S_i \mathbf{w}$$

$$s_1^2 + s_2^2 = \mathbf{w}^T(S_1 + S_2)\mathbf{w} = \mathbf{w}^T S_w \mathbf{w}$$

$$\boxed{J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}}$$

S_w: **total *within-class* covariance (scatter) matrix**

$$\boxed{J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}}$$

- J(w) is maximized when

$$(\mathbf{w}^T S_w \mathbf{w}) S_B \mathbf{w} = (\mathbf{w}^T S_B \mathbf{w}) S_w \mathbf{w}$$

Scalar

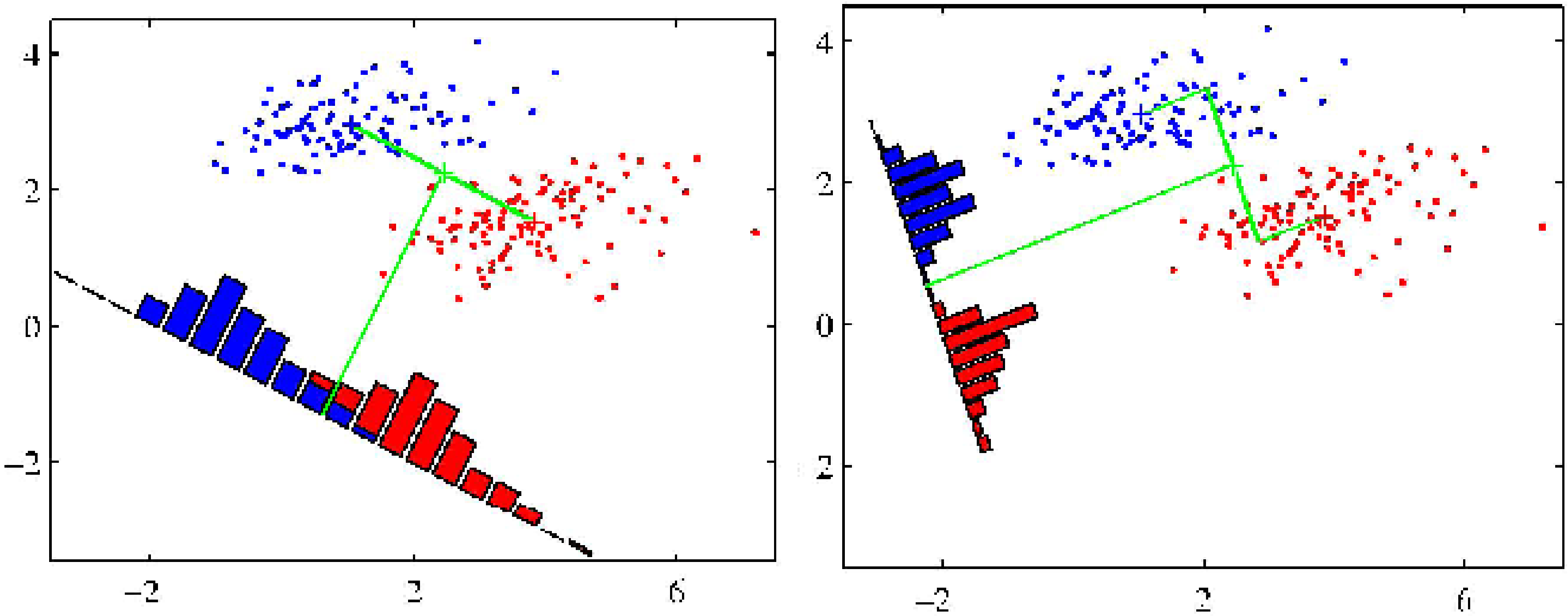
- Noticing that $S_B \mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}$

  always take direction $\mathbf{m}_1 - \mathbf{m}_2$

And we don't care about the magnitude of **w**, we can drop off the scalar's in the equation

$$(\mathbf{m}_1 - \mathbf{m}_2) = S_w \mathbf{w} \quad \Rightarrow \quad \mathbf{w} = S_w^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

Visual depiction of LDA



LDA

$$\mathbf{w} = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Gives the linear function with maximum ratio of between class-scatter and within-class scatter
- The classification problem is reduced from a q-dimensional problem to a 1-dimensional problem
- Can then learn a threshold for the final determinant function

Generalizing to Multi-Class

- For K >2

$$\mathbf{S}_w = \sum_{k=1}^{K} \mathbf{S}_k \quad \mathbf{S}_k = \sum_{n \in c_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$$

$$\mathbf{S}_B = \sum_{k=1}^{K} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T$$

$$J(\mathbf{w}) = Tr\{(\mathbf{w}\mathbf{S}_W \mathbf{w}^T)^{-1}(\mathbf{w}\mathbf{S}_B \mathbf{w}^T)\}$$

$\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_{k-1}$: the k-1 largest eigen-vectors of $\mathbf{S}_W{}^{-1}\mathbf{S}_B$