Linear Classification

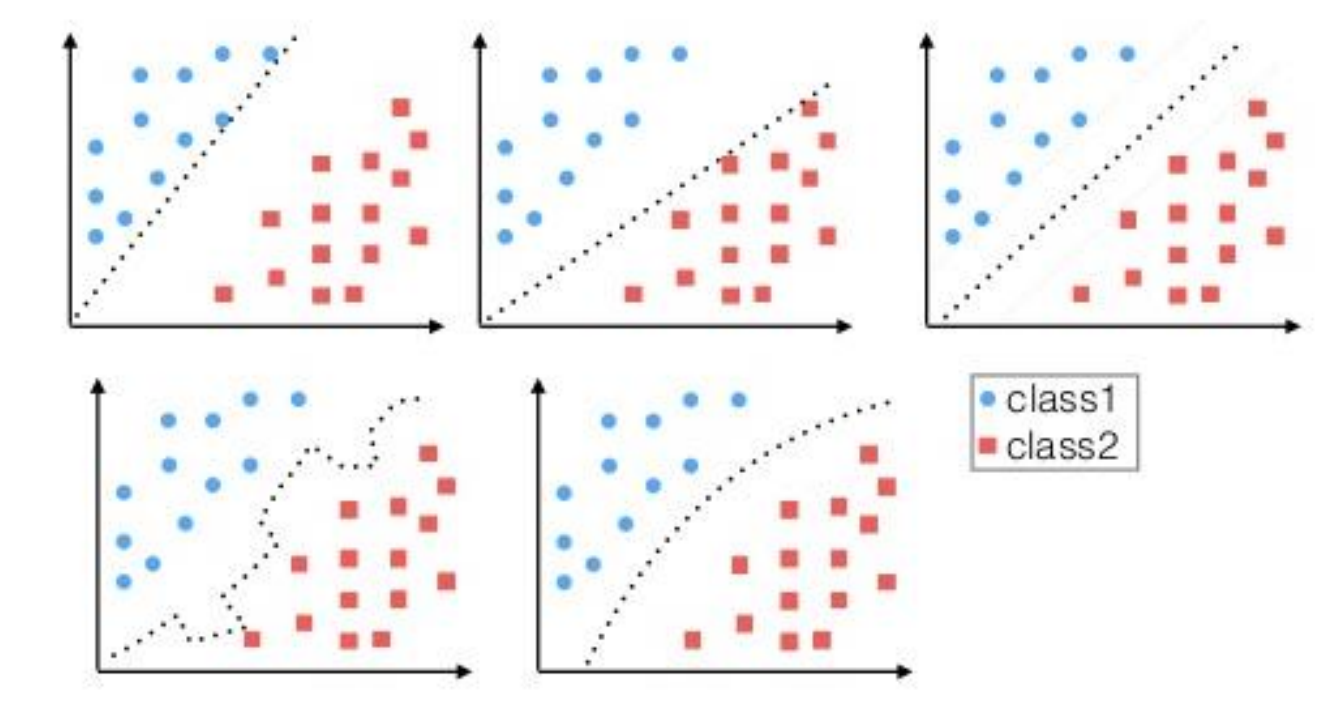
Machine Learning Sessions

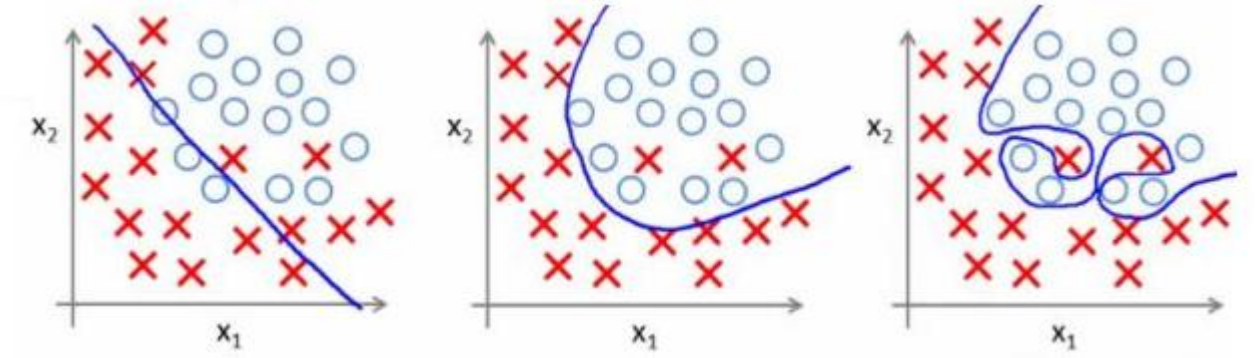
Parisa Abedi

Definition

- Classification:
  - Use an object characteristics to identify which class/category it belongs to
    - Example:
      - a new email is ‘spam’ or ‘non-spam’
      - A patient diagnosed with a disease or not
- Classification is an example of pattern recognition

Definition





Definition

- Linear classification
  - A classification algorithm (Classifier) that makes its classification based on a linear predictor function combining a set of weights with the feature vector

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right),$$

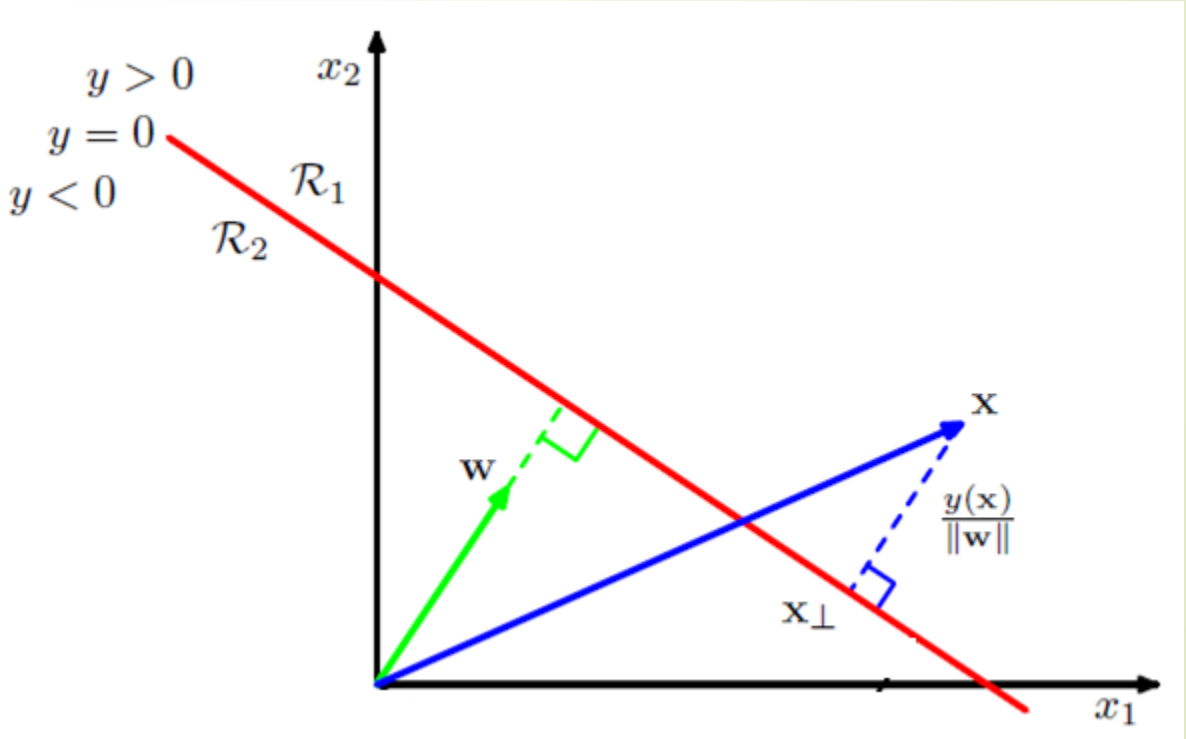
  - Decision boundaries is flat
    - Line, plane, ….
  - May involve non-linear operations

Different approaches

- Explicitly creating the discriminant function (Discriminant function)
  - Perceptron
  - Support vector machine
- Probabilistic approach
  - Model the posterior distribution
  - Algorithms
    - Logistic regression

Discriminant functions

- Two classes: $y(X) = w^T X + w_0$



- More than two classes (K classes):
  - There are K indicators
  - $Y_k = 1$ if G = K, else 0
  - $Y = (Y_1, \ldots, Y_K)$
  - Example: (0,0,1,0) for a feature in class 3 when there are 4 classes!

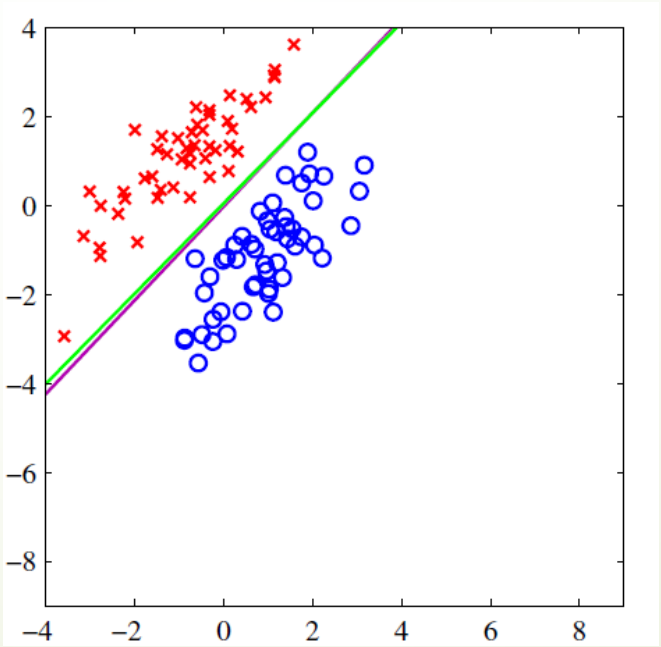
Discriminant functions
Least squares

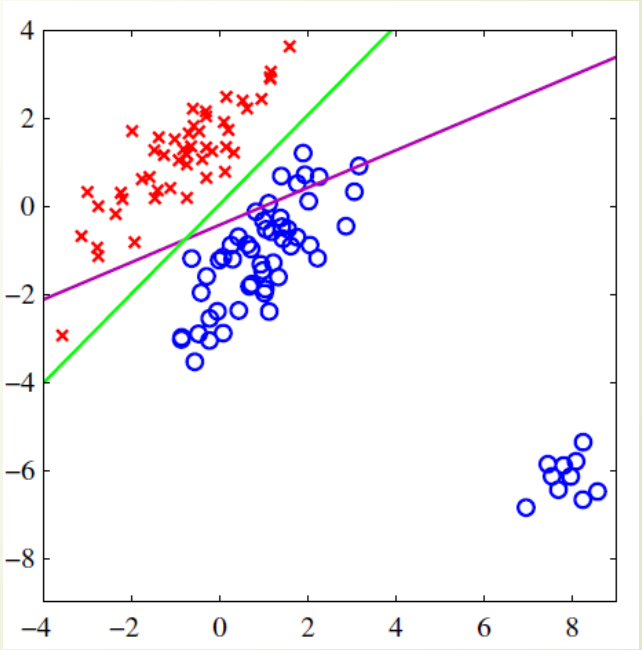
- Simultaneously fit a linear regression model to each of the columns of Y
  - Weights will have a close form $W = (X^T X)^{-1} X^T Y$

- Classify a new observation x:
  - For each class calculate the f(x) = W.X
  - Select the class with higher value for f(x)

Discriminant functions
Least squares

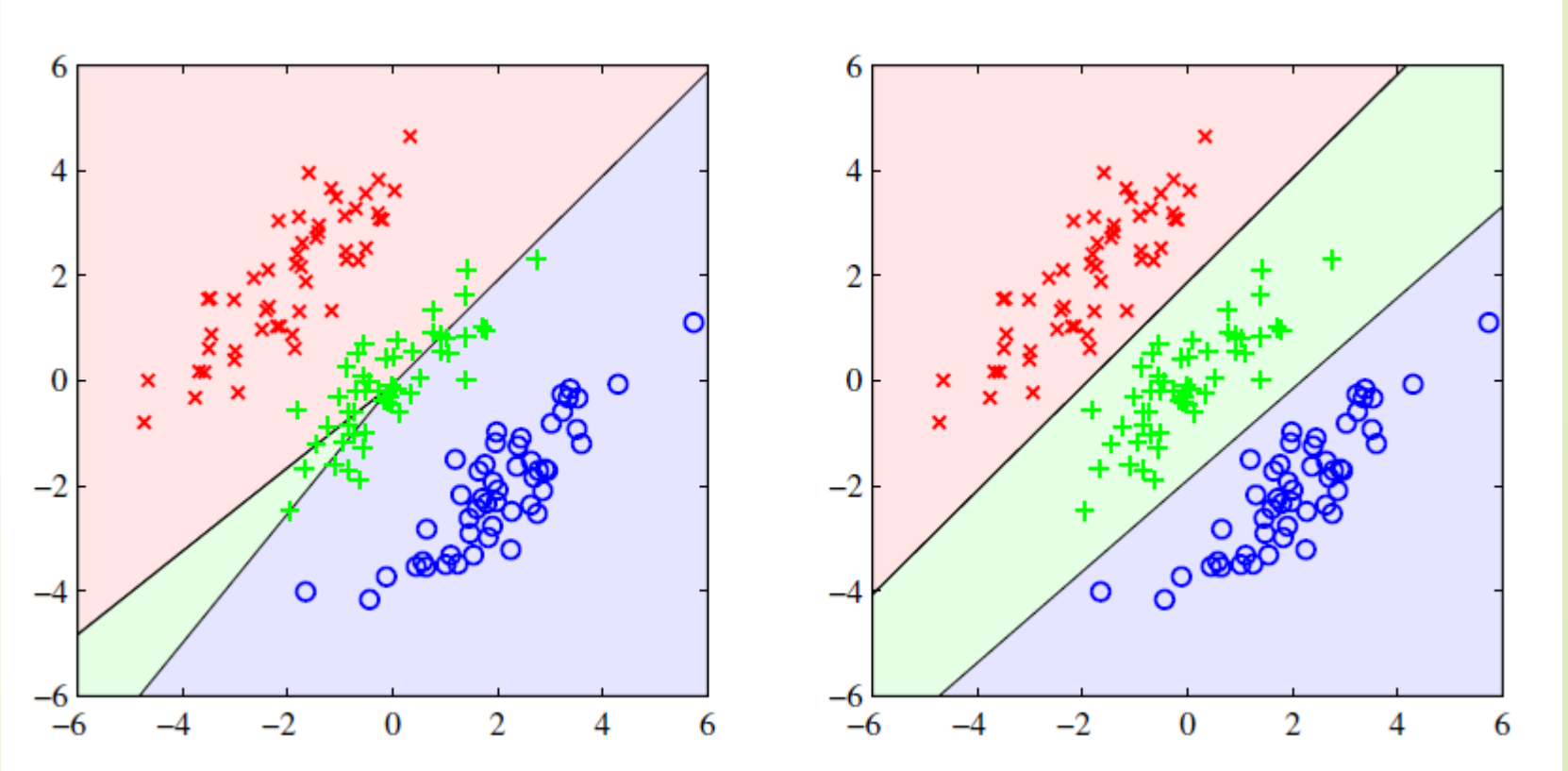
- Works well
  - Linearly separable
  - Few outliers
  - K = 2





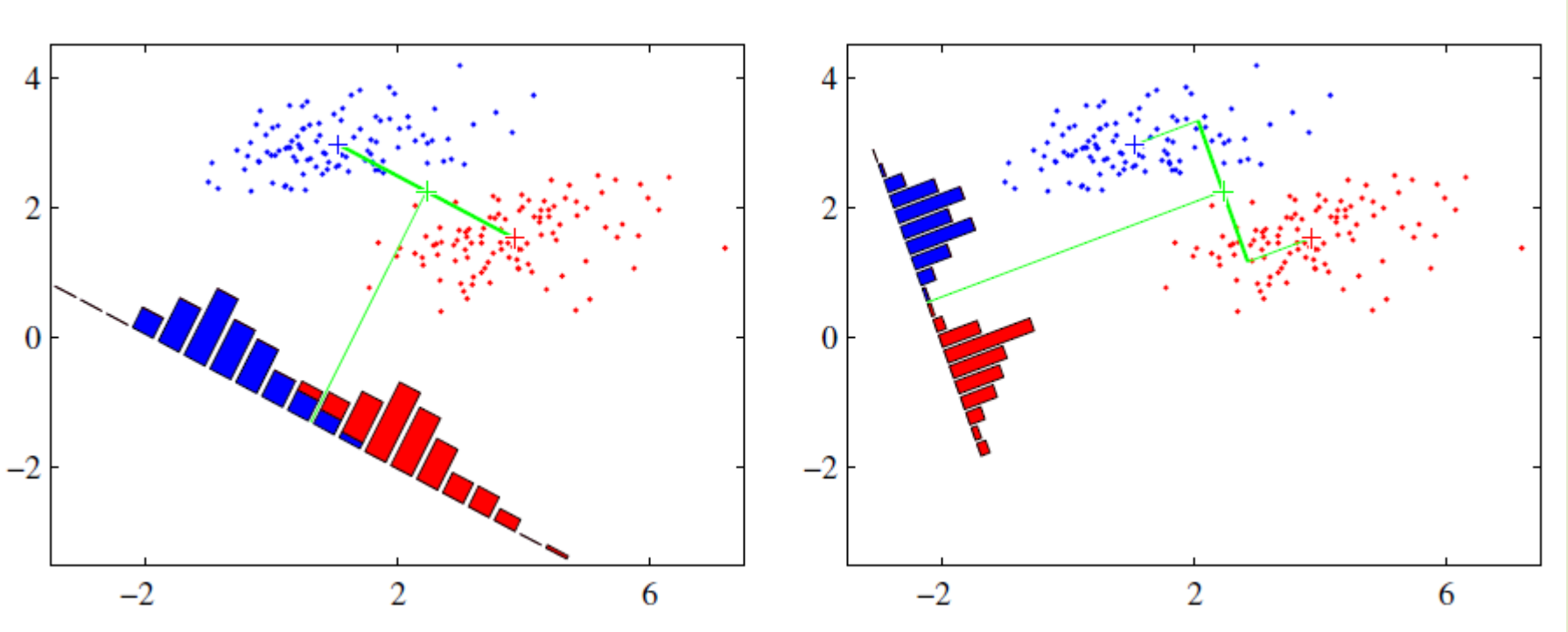
Discriminant functions
Least squares

- Not good for
  - $K \geq 3$



Discriminant functions
Fischer's method

- Dimensionality reduction



- Still need a decision rule

Discriminant functions Perceptron

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right), f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0. \end{cases}$$

How to drive w?

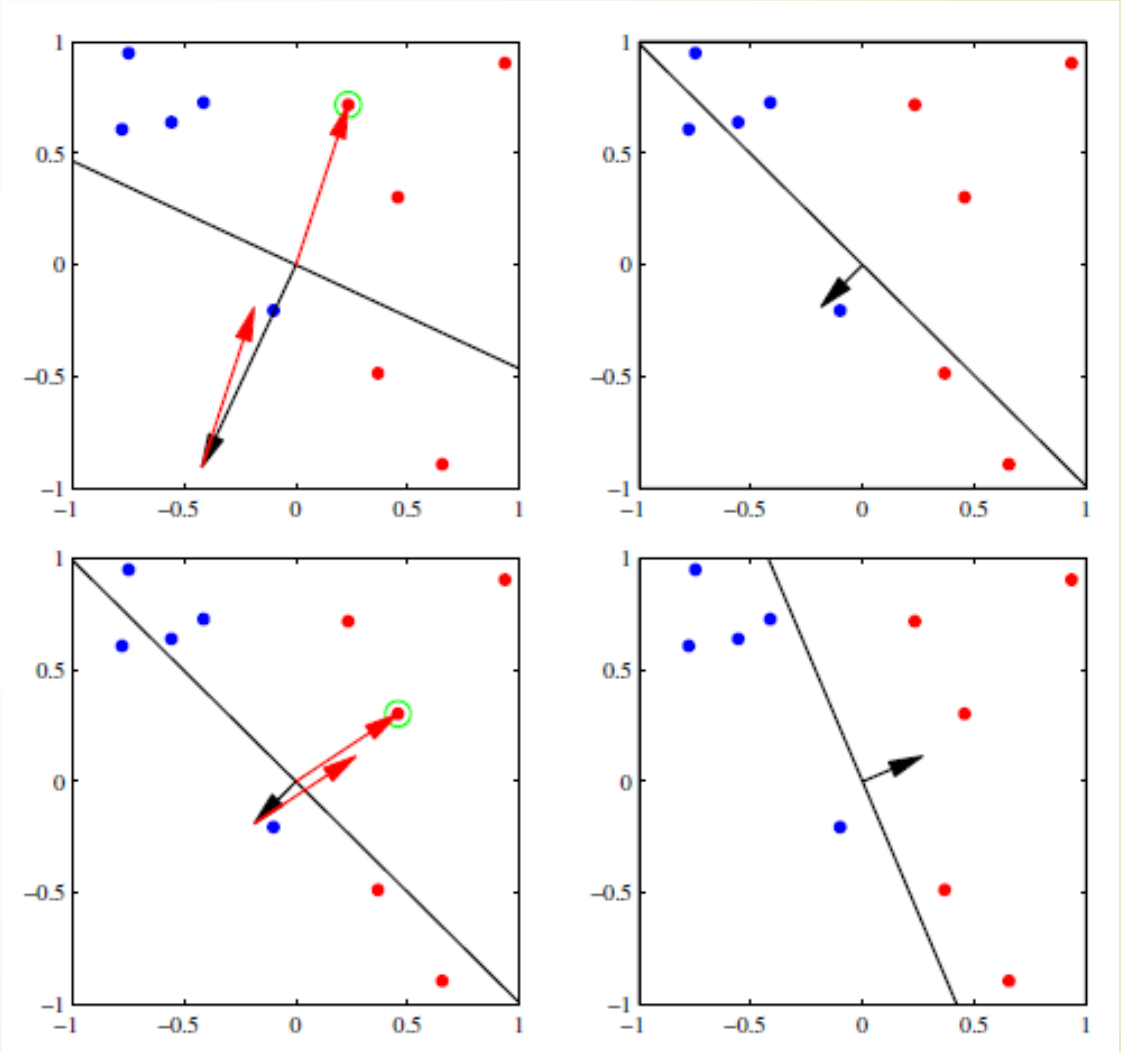
- Choose some initial weights
- For each example of x_j in training set calculate the y

$$\begin{aligned} y_j(t) &= f[\mathbf{w}(t) \cdot \mathbf{x}_j] \\ &= f[w_0(t)x_{j,0} + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + w_n(t)x_{j,n}] \end{aligned}$$

- Update the weights

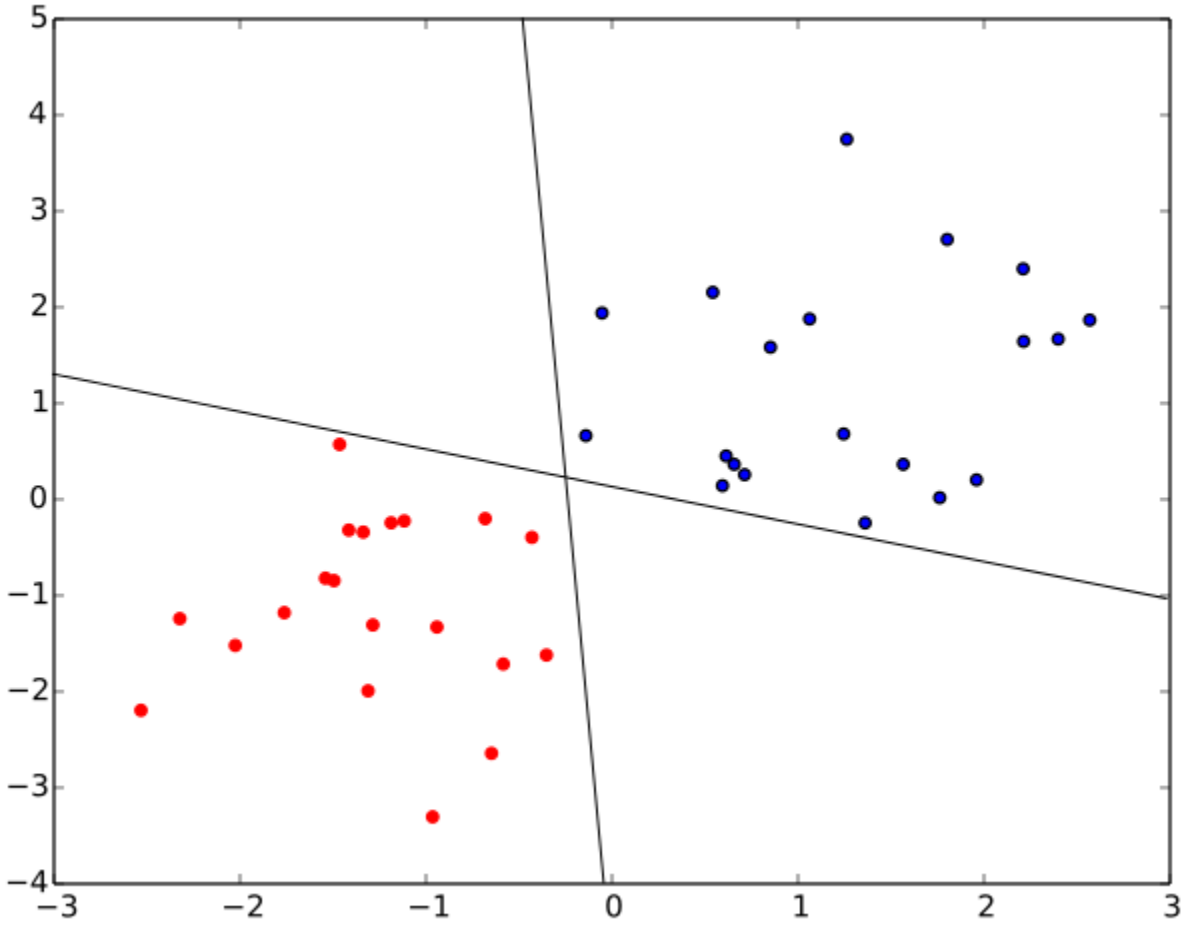
$$w_i(t+1) = w_i(t) + \alpha(d_j - y_j(t))x_{j,i} \quad 0 \leq i \leq n$$

Discriminant functions
Perceptron



Discriminant functions
Perceptron

- Works well for linearly separable
- It's fast
- Can work online
- Not able to choose the best boundaries



Probabilistic approaches

- Determine the class-conditional densities for each class

$$p(\mathbf{x}|\mathcal{C}_k)$$

- Bayes' theorem

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

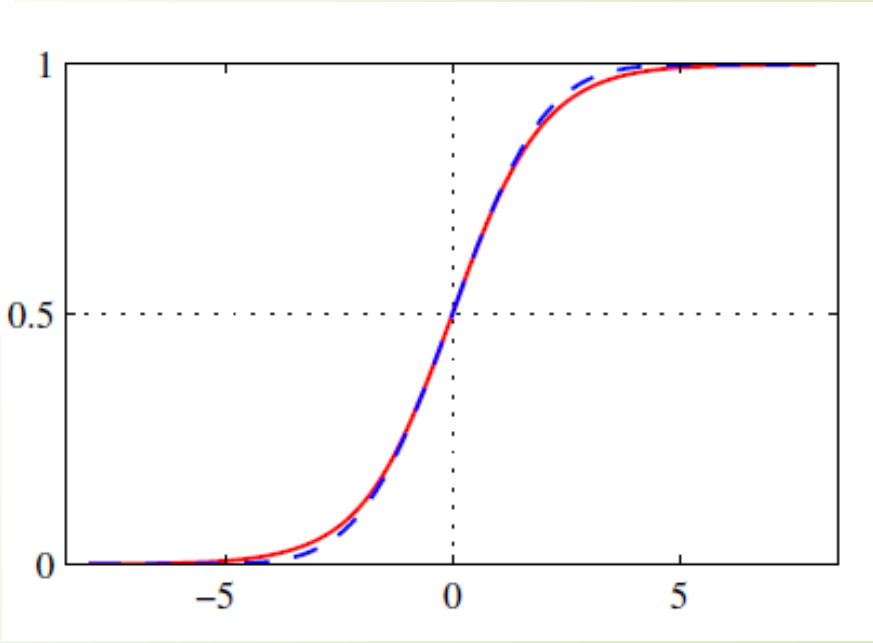
Probabilistic approaches

- Use logistic sigmoid function

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a)$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$



Probabilistic approaches

- Gaussian input with common covariance matrix

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

- In the case of two classes

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

$$\begin{aligned} \mathbf{w} &= \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. \end{aligned}$$

Probabilistic approaches

- Estimate the parameters
  - Maximum likelihood estimation
    - Data set $\{\mathbf{x}_n, t_n\}$
    - $t_n = 1$ denotes class $\mathcal{C}_1$

$$p(\mathbf{x}_n, \mathcal{C}_1) = p(\mathcal{C}_1)p(\mathbf{x}_n|\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$p(\mathbf{x}_n, \mathcal{C}_2) = p(\mathcal{C}_2)p(\mathbf{x}_n|\mathcal{C}_2) = (1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$$