CS 2750 Machine Learning
Lecture 11

Support vector machines

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Outline

Outline:
- Fisher Linear Discriminant
- Algorithms for linear decision boundary
- **Support vector machines**
- Maximum margin hyperplane.
- Support vectors.
- Support vector machines.

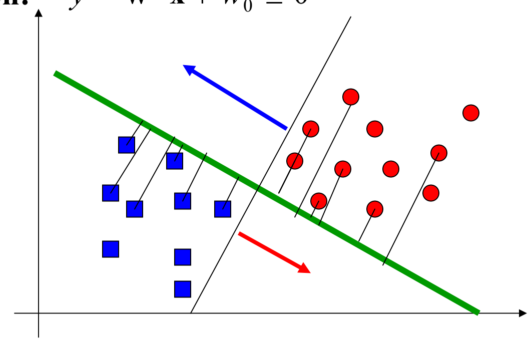
- Extensions to the non-separable case.
- Kernel functions.

Fisher linear discriminant

- Project data into one dimension

$$y = \mathbf{w}^T \mathbf{x}$$

Decision: $y = \mathbf{w}^T \mathbf{x} + w_0 \geq 0$

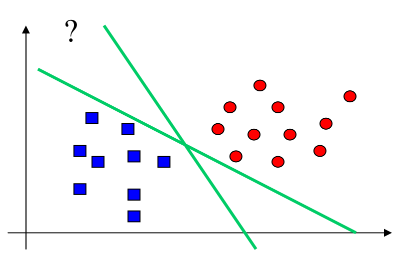


- How to find the projection line?

Fisher linear discriminant

How to find the projection line?

$$y = \mathbf{w}^T \mathbf{x}$$

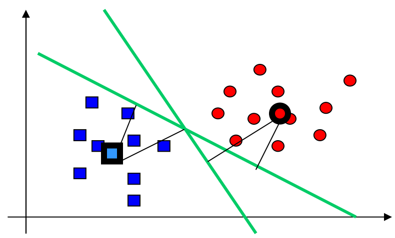


Fisher linear discriminant

Assume: $\mathbf{m}_1 = \frac{1}{N_1} \sum_{i \in C_1}^{N_1} \mathbf{x}_i \qquad \mathbf{m}_2 = \frac{1}{N_2} \sum_{i \in C_2}^{N_2} \mathbf{x}_i$

Maximize the difference in projected means:

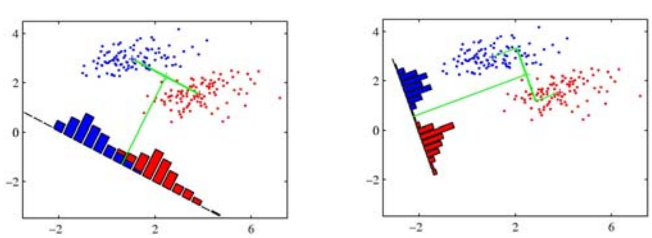
$$m_2 - m_1 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)$$



Fisher linear discriminant

Problem 1: $m_2 - m_1 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1)$ can be maximized by increasing $\mathbf{w}$

Problem 2: variance in class distributions after projection is changed



Fisher's solution: $$J(\mathbf{w}) = \frac{m_2 - m_1}{s_1^2 + s_2^2}$$

Within class variance $$s_k^2 = \sum_{i \in C_k} (y_i - m_k)^2$$

Fisher linear discriminant

Error:
$$J(\mathbf{w}) = \frac{m_2 - m_1}{s_1^2 + s_2^2}$$

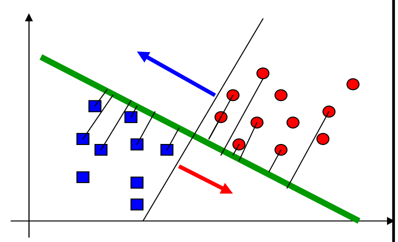
Within class variance after the projection

$$s_k^2 = \sum_{i \in C_k} (y_i - m_k)^2$$

Optimal solution:

$$\mathbf{w} \approx \mathbf{S}_{\mathbf{w}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

$$\mathbf{S}_{\mathbf{w}} = \sum_{i \in C_1} (\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{i \in C_2} (\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T$$



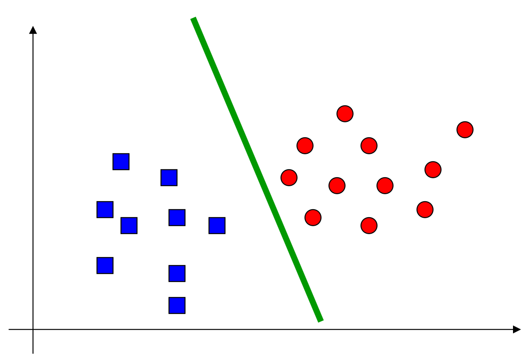
Linearly separable classes

There is a **hyperplane** that separates training instances with no error

Hyperplane:

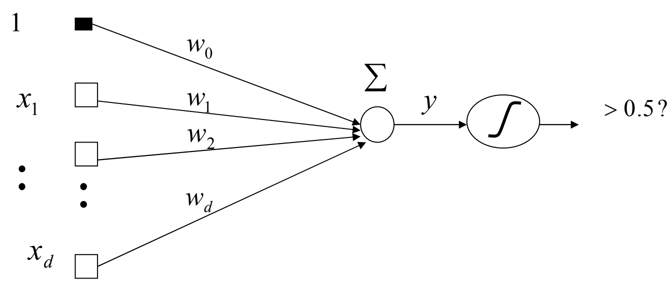
$$\mathbf{w}^T\mathbf{x} + w_0 = 0$$

Class (+1) $\mathbf{w}^T\mathbf{x} + w_0 > 0$
Class (-1) $\mathbf{w}^T\mathbf{x} + w_0 < 0$



Algorithms for linearly separable set

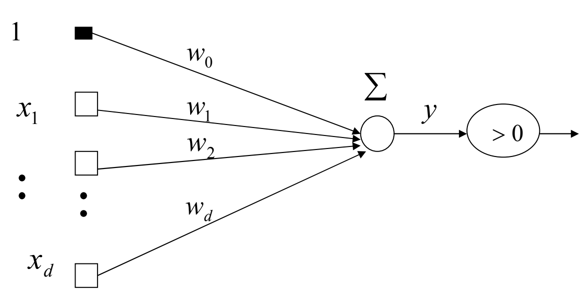
- **Separating hyperplane** $\mathbf{w}^T\mathbf{x} + w_0 = 0$



- We can use **gradient methods** or Newton Rhapson for sigmoidal switching functions and learn the weights
- Recall that we learn the linear decision boundary

Algorithms for linearly separable set

- **Separating hyperplane** $\mathbf{w}^T\mathbf{x} + w_0 = 0$



Algorithms for linearly separable sets

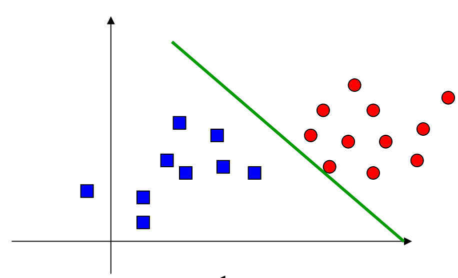
- **Perceptron algorithm:**
  Simple iterative procedure for modifying the weights of the linear model
  Initialize weights $\mathbf{w}$
  Loop through examples $(\mathbf{x}, y)$ in the dataset D
    1. Compute $\hat{y} = \mathbf{w}^T \mathbf{x}$
    2. If $y \neq \hat{y} = -1$ then $\mathbf{w}^T \leftarrow \mathbf{w}^T + \mathbf{x}$
    3. If $y \neq \hat{y} = +1$ then $\mathbf{w}^T \leftarrow \mathbf{w}^T - \mathbf{x}$

  Until all examples are classified correctly

Properties:

guaranteed convergence

Algorithms for linearly separable sets



Linear program solution:

- Finds weights that satisfy the following constraints:

$\mathbf{w}^T \mathbf{x}_i + w_0 \geq 0$ For all i, such that $y_i = +1$

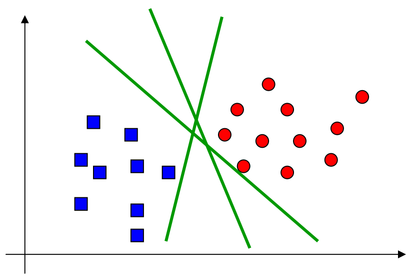
$\mathbf{w}^T \mathbf{x}_i + w_0 \leq 0$ For all i, such that $y_i = -1$

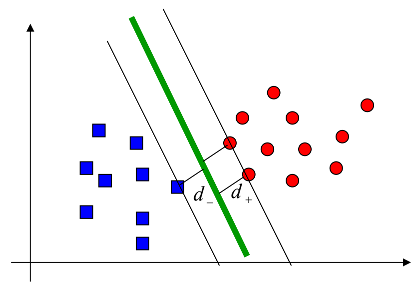
Together: $y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 0$

Property: if there is a hyperplane separating the examples, the linear program finds the solution

Optimal separating hyperplane

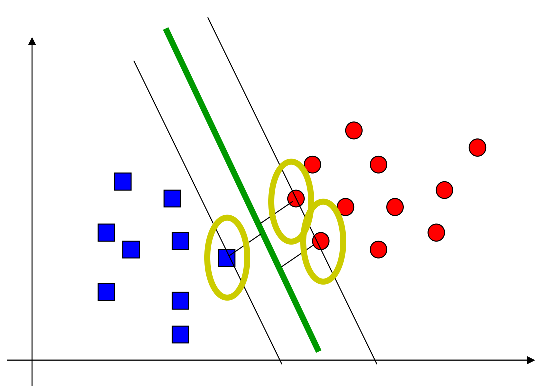
- There are multiple hyperplanes that separate the data points
  - Which one to choose?
- **Maximum margin** choice: the maximum distance of $d_+ + d_-$
  - where d_+ is the shortest distance of a positive example from the hyperplane (similarly d_- for negative examples)





Maximum margin hyperplane

- For the maximum margin hyperplane only examples on the margin matter (only these affect the distances)
- These are called **support vectors**



Finding maximum margin hyperplanes

- **Assume** that examples in the training set are $(\mathbf{x}_i, y_i)$ such that $y_i \in \{+1,-1\}$
- **Assume** that all data satisfy:

$$\mathbf{w}^T\mathbf{x}_i + w_0 \geq 1 \quad \text{for} \quad y_i = +1$$

$$\mathbf{w}^T\mathbf{x}_i + w_0 \leq -1 \quad \text{for} \quad y_i = -1$$

- The inequalities can be combined as:

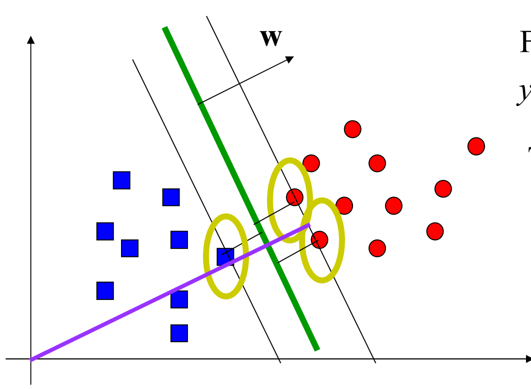
$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 \geq 0 \quad \text{for all} \quad i$$

- Equalities define two hyperplanes:

$$\mathbf{w}^T\mathbf{x}_i + w_0 = 1 \qquad \mathbf{w}^T\mathbf{x}_i + w_0 = -1$$

Finding the maximum margin hyperplane

- **Geometrical margin:** $\rho_{\mathbf{w},w_0}(\mathbf{x},y) = y(\mathbf{w}^T\mathbf{x} + w_0)/\|\mathbf{w}\|_{L2}$
  - measures the distance of a point $\mathbf{x}$ from the hyperplane
    $\mathbf{w}$ - normal to the hyperplane $\|..\|_{L2}$ - Euclidean norm



For points satisfying:

$$y_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 = 0$$

The distance is $\dfrac{1}{\|\mathbf{w}\|_{L2}}$

Width of the margin:

$$d_+ + d_- = \frac{2}{\|\mathbf{w}\|_{L2}}$$

Maximum margin hyperplane

- **We want to maximize** $d_+ + d_- = \dfrac{2}{\|\mathbf{w}\|_{L2}}$
- We do it by **minimizing**

$$\|\mathbf{w}\|_{L2}^{2} / 2 = \mathbf{w}^T \mathbf{w} / 2$$

$\mathbf{w}, w_0$ - variables

  – But we also need to enforce the constraints on points:

$$\left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 \right] \geq 0$$

Maximum margin hyperplane

- **Solution: Incorporate constraints into the optimization**
- **Optimization problem** (Lagrangian)

$$J(\mathbf{w}, w_0, \alpha) = \|\mathbf{w}\|^2 / 2 - \sum_{i=1}^{n} \alpha_i \left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 \right]$$

$\alpha_i \geq 0$ - **Lagrange multipliers**

- **Minimize** with respect to $\mathbf{w}, w_0$ (primal variables)
- **Maximize** with respect to $\boldsymbol{\alpha}$ (dual variables)

Lagrange multipliers enforce the satisfaction of constraints

If $\left[y_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 \right] > 0 \implies \alpha_i \to 0$

Else $\implies \alpha_i > 0$ Active constraint