Machine Learning
ITCS 4156

Support Vector Machines

Razvan C. Bunescu

Department of Computer Science @ CCI

rbunescu@uncc.edu

Max-Margin Classifiers: Separable Case

- Linear model for binary classification:

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

- Training examples:

$$(\mathbf{x}_1, t_1),\ (\mathbf{x}_2, t_2),\ \ldots\ (\mathbf{x}_N, t_N), \text{ where } t_n \in \{+1, -1\}$$

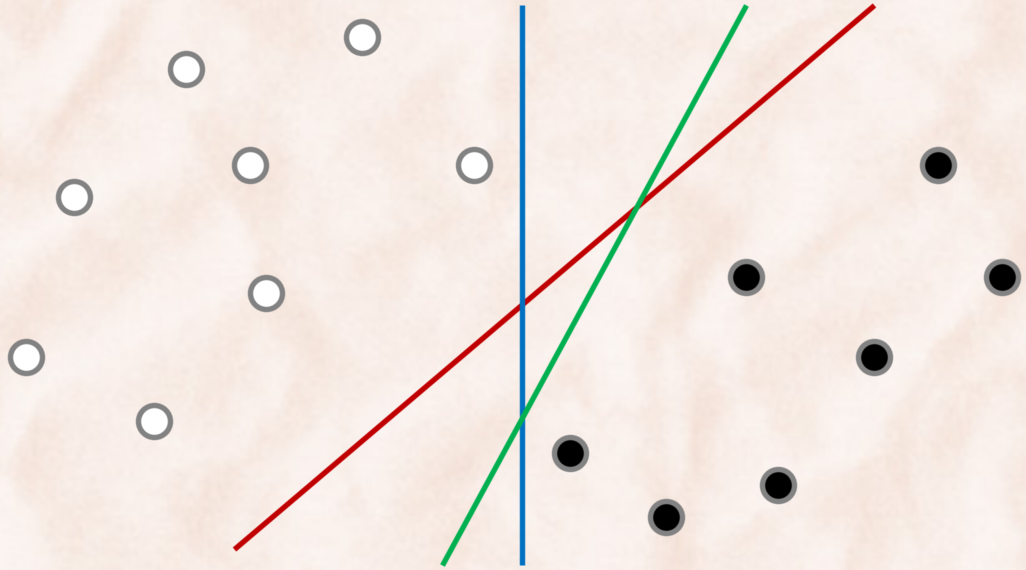
- Assume training data is linearly separable:

$$t_n y(x_n) > 0\text{, for all } 1 \leq n \leq N$$

$\Rightarrow$ perceptron solution depends on:

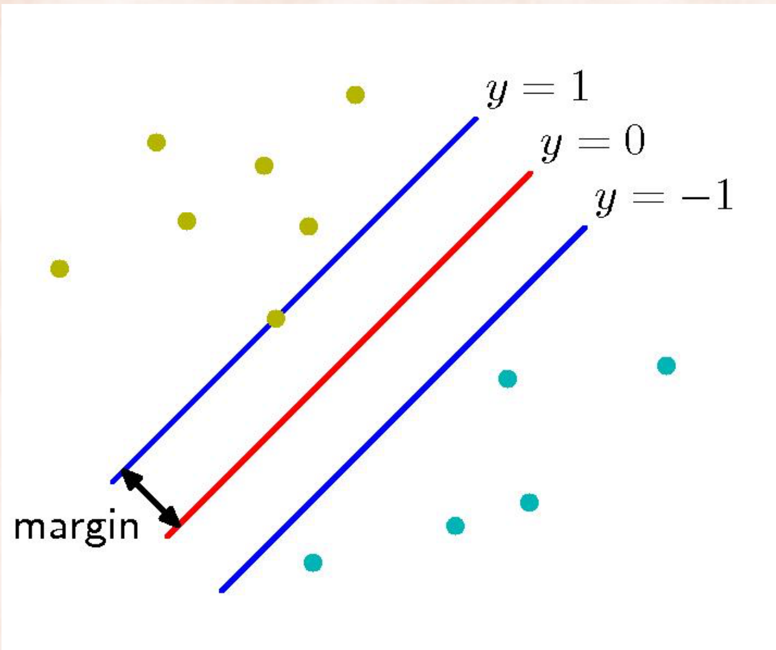
- initial values of **w** and b.
- order of processing of data points.

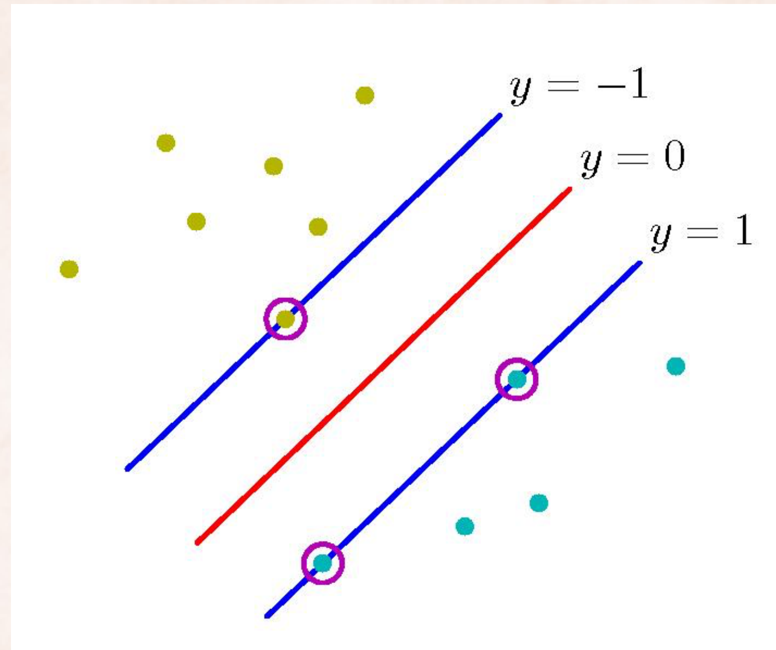
Maximum Margin Classifiers



- Which hyperplane has the smallest generalization error?
  - The one that maximizes the margin [Computational Learning Theory]
    - margin = the distance between the decision boundary and the closest sample.

Maximum Margin Classifiers





- The distance between a point $\mathbf{x}_n$ and a hyperplane $y(\mathbf{x})=0$ is:

$$\frac{|y(\mathbf{x}_n)|}{\|\mathbf{w}\|} = \frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T \varphi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}$$

Maximum Margin Classifiers

- Margin = the distance between hyperplane $y(\mathbf{x})$=0 and closest sample:

$$\min_n \left[\frac{t_n(\mathbf{w}^T\varphi(\mathbf{x}_n)+b)}{\|\mathbf{w}\|} \right]$$

- Find parameters **w** and b that maximize the margin:

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n(\mathbf{w}^T\varphi(\mathbf{x}_n)+b) \right] \right\}$$

- Rescaling **w** and b does not change distances to the hyperplane:

$\Rightarrow$ for the closest point(s), set $t_n(\mathbf{w}^T\varphi(\mathbf{x}_n)+b)=1$

$\Rightarrow t_n(\mathbf{w}^T\varphi(\mathbf{x}_n)+b) \geq 1, \quad \forall n \in \{1,\ldots,N\}$

Max-Margin: Quadratic Optimization

- Constrained optimization problem:

minimize:

$$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2$$

subject to:

$$t_n(\mathbf{w}^T \varphi(\mathbf{x}_n) + b) \geq 1, \quad \forall n \in \{1, \ldots, N\}$$

- Solved using the technique of Lagrange Multipliers.
  - [*derivation not shown in this class, but see end of slides if interested*].

Max-Margin: Quadratic Optimization

- Equivalent **dual representation**:

maximize:

$$L_D(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

subject to:

$$\alpha_n \geq 0, \quad n = 1, \ldots, N$$

$$\sum_{n=1}^{N} \alpha_n t_n = 0$$

  - $k(\mathbf{x}_n,\mathbf{x}_m) = \varphi(\mathbf{x}_n)^T \varphi(\mathbf{x}_n)$ is the *kernel* function.
  - where $\boxed{\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \varphi(x_n)}$ and $\sum_{n=1}^{N} \alpha_n t_n = 0$

Exactly like in the Kernel Perceptron!

KKT conditions

1. primal constraints: $t_n y(x_n) - 1 \geq 0$

1. dual constraints: $\alpha_n \geq 0$

2. complementary slackness: $\alpha_n \{ t_n y(x_n) - 1 \} = 0$

$\Rightarrow$ for any data point, either $\alpha_n = 0$ or $t_n y(x_n) = 1$

S = $\{n \mid t_n y(x_n) = 1\}$ is the set of *support vectors*

Max-Margin Solution

- After solving the dual problem $\Rightarrow$ know α_n, for $n = 1\ldots N$

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \varphi(x_n) = \sum_{m \in S} \alpha_m t_m \varphi(x_m)$$

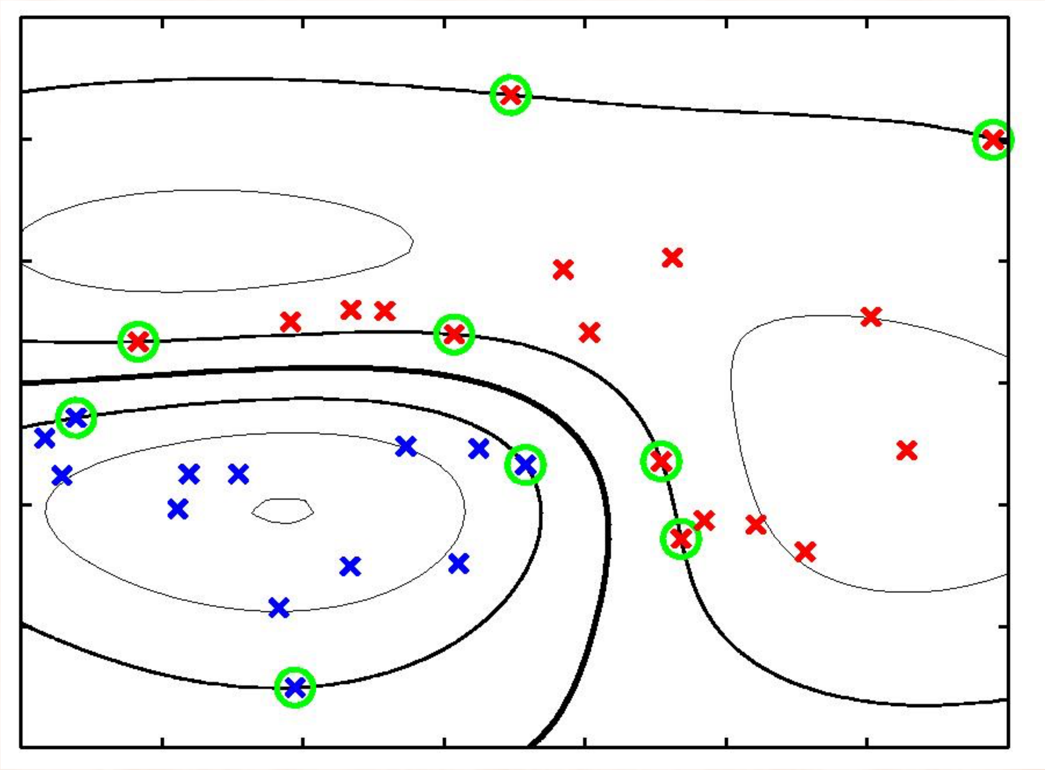
$$b = \frac{1}{|S|} \sum_{n \in S} \left(t_n - \sum_{m \in S} \alpha_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

- Linear discriminant function becomes:

$$y(x) = \sum_{m \in S} \alpha_m t_m k(x, x_m) + b$$

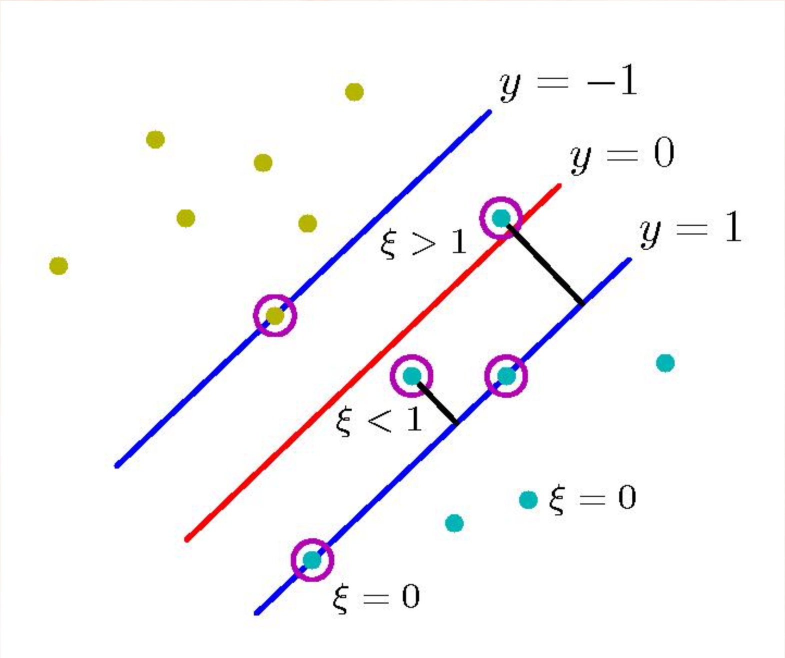
$\Rightarrow$ In both training and testing, examples are used only through the *kernel function*!

An SVM with Gaussian kernel



Max-Margin Classifiers: Non-Separable Case

- Allow data points to be on the wrong side of the margin boundary.
  - Penalty that increases with the distance from the boundary.



Max-Margin: Quadratic Optimization

- Optimization problem:

minimize:

$$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

subject to:

$$t_n(\mathbf{w}^T\varphi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \quad \forall n \in \{1, \ldots, N\}$$
$$\xi_n \geq 0$$

- Solve it using the technique of Lagrange Multipliers.

Max-Margin: Quadratic Optimization

- Dual representation:

maximize:

$$L_D(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

subject to:

$$0 \le \alpha_n \le C, \quad n = 1, \ldots, N$$

$$\sum_{n=1}^{N} \alpha_n t_n = 0$$

- $k(\mathbf{x}_n, \mathbf{x}_m) = \varphi(\mathbf{x}_n)^T \varphi(\mathbf{x}_n)$ is the *kernel* function.

(Some of the) KKT conditions

1. primal constraints: $t_n y(x_n) - 1 + \xi_n \geq 0$

1. dual constraints: $0 \leq \alpha_n \leq C$

2. complementary slackness: $\alpha_n \left\{ t_n y(x_n) - 1 + \xi_n \right\} = 0$

$\Rightarrow$ for any data point, either $\alpha_n = 0$ or $t_n y(x_n) = 1 - \xi_n$

$S = \{n \mid t_n y(x_n) = 1 - \xi_n\}$ is the set of *support vectors*

$M = \{n \mid 0 < \alpha_n < \mathrm{C}\}$ is the set of SVs that lie on the margin.

Max-Margin Solution

- After solving the dual problem $\Rightarrow$ know α_n, for $n = 1\ldots N$

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \varphi(x_n) = \sum_{m \in S} \alpha_m t_m \varphi(x_m)$$

$$b = \frac{1}{|M|} \sum_{n \in M} \left(t_n - \sum_{m \in S} \alpha_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

- Linear discriminant function becomes:

$$y(x) = \sum_{m \in S} \alpha_m t_m k(x, x_m) + b$$

$\Rightarrow$ In both training and testing, examples are used only through the *kernel function*!

Support Vector Machines

- Optimization problem:

upper bound on the **misclassification error** on the training data.

minimize:

$$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C \boxed{\sum_{n=1}^{N} \xi_n}$$

subject to:

$$t_n(\mathbf{w}^T \varphi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \quad \forall n \in \{1, \ldots, N\}$$
$$\xi_n \geq 0$$

  – Implemented in *sklearn*:
    - https://scikit-learn.org/stable/modules/svm.html
    - https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html