# CS578 – Interactive and Transparent Machine Learning

## Topic: Support Vector Machines

**Mustafa Bilgic**

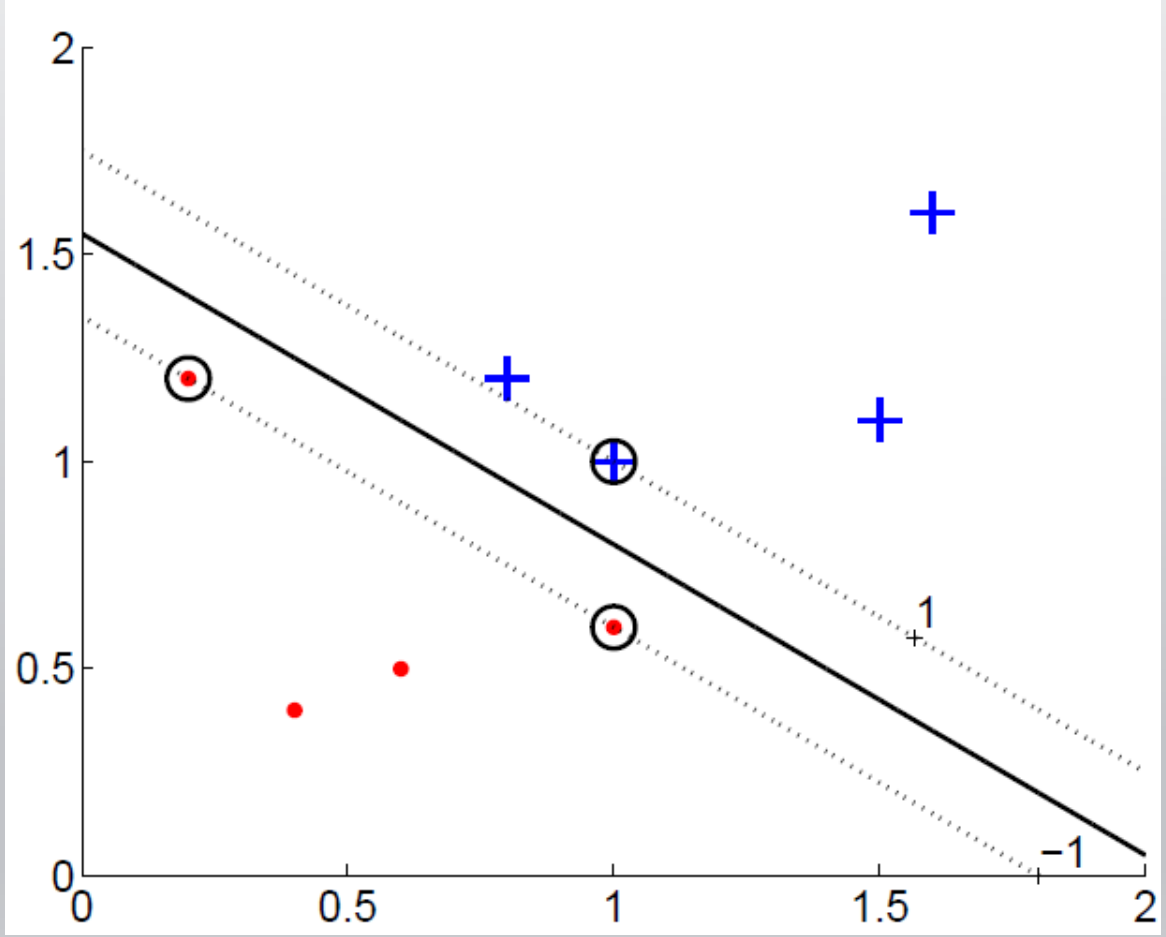
http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

# References

- Great notes, by Andrew Ng
  - http://cs229.stanford.edu/notes/cs229-notes3.pdf
- Great lecture, by Patrick Winston
  - https://www.youtube.com/watch?v=_PwhiWxHK8o

# Margin






Credit: Ethem Alpaydin. Introduction to Machine Learning. 3rd Edition. *http://www.cmpe.boun.edu.tr/~ethem/i2ml3e*

# Objective Function

- $D = \{\langle x^{(d)}, y^{(d)} \rangle\}$ where $y \in \{-1, +1\}$
- Find $w$ and $b$ such that
  - $w^T x^{(d)} + b \geq +1$ if $y^{(d)} = +1$ and
  - $w^T x^{(d)} + b \leq -1$ if $y^{(d)} = -1$
- Which can be written simply
  - $y^{(d)}\left(w^T x^{(d)} + b\right) \geq +1$

# Maximum Margin Classification

- Maximize (geometric margin)
  - $\min \frac{1}{2} w^T w$ subject to
  - $y^{(d)}\left(w^T x^{(d)} + b\right) \geq +1$ for $d \in D$
- A convex quadratic objective function with linear constraints
  - Quadratic programming solvers can easily solve it
  - For example: http://cvxopt.org/userguide/coneprog.html#quadratic-programming

# Lagrangian - Primal

- The objective function
  - $\min \frac{1}{2} w^T w$ subject to $y^{(d)}\left(w^T x^{(d)} + b\right) \geq +1$
- Form its Lagrangian
  - $L_p = \frac{1}{2} w^2 - \sum_{d \in D} \alpha^{(d)} \left(y^{(d)}\left(w^T x^{(d)} + b\right) - 1\right)$
  - $\alpha^{(d)} \geq 0 \; \forall d \in D$
- Take its derivative w.r.t $w$ and $b$
  - $\frac{\partial L_p}{\partial w} = w - \sum_{d \in D} \alpha^{(d)} y^{(d)} x^{(d)} = 0 \Rightarrow w = \sum_{d \in D} \alpha^{(d)} y^{(d)} x^{(d)}$
  - $\frac{\partial L_p}{\partial b} = \sum_{d \in D} \alpha^{(d)} y^{(d)} = 0$

# Lagrangian – Primal – Solution

- Solution of the primal: find $\alpha^{(d)}$. Then
  - $w = \sum_{d \in D} \alpha^{(d)} y^{(d)} x^{(d)}$
- $\alpha^{(d)}$ is non-zero for instances that have a functional margin of $+1$ or $-1$ and zero for others
- The ones that have a functional margin of $+1$ or $-1$ are called the support vectors
- $b$ can be calculated using support vectors
- Classify new object $x$ using
  - $\text{sign}(w^T x + b)$

# Lagrangian - Dual

- Enforce the derivatives in the primal itself
- $L_{dual} =$

  $-\frac{1}{2}\sum_{i\in D}\sum_{j\in D}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\,x^{(i)}x^{(j)} + \sum_{d\in D}\alpha^{(d)}$
- Subject to
  - $\sum_{d\in D}\alpha^{(d)}\,y^{(d)} = 0$
  - $\alpha^{(d)} \geq 0\ \forall d \in D$
- Maximize $L_{dual}$ with respect to α

# Primal VS Dual

- Primal and dual formulation lead to the same solution under certain conditions, called
  - Karush–Kuhn–Tucker conditions
  - Often abbreviated as KKT conditions
  - We will not go into details of KKT in this class
- Dual has the nice formalism that it enables the "kernel" trick

# Kernel Trick

- $L_{dual} =$

  $-\frac{1}{2}\sum_{i\in D}\sum_{j\in D}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\,x^{(i)}x^{(j)} + \sum_{d\in D}\alpha^{(d)}$

- $L_{dual} =$

  $-\frac{1}{2}\sum_{i\in D}\sum_{j\in D}\alpha^{(i)}\alpha^{(j)}y^{(i)}y^{(j)}\,K(x^{(i)},x^{(j)}) + \sum_{d\in D}\alpha^{(d)}$

  - where $K(x^{(i)},x^{(j)})$ is a kernel

- Remember $w = \sum_{d\in D}\alpha^{(d)}y^{(d)}x^{(d)}$

- So $\text{sign}\left(w^T x + b\right) =$

  - $\text{sign}\left(\left(\sum_{d\in D}\alpha^{(d)}y^{(d)}x^{(d)}\right)x + b\right) =$
  - $\text{sign}\left(\sum_{d\in D}\alpha^{(d)}y^{(d)}K(x^{(d)},x) + b\right)$

# Some Kernels

- Linear kernel
  - $K\left(x^{(i)}, x^{(j)}\right) = x^{(i)^T} x^{(j)}$
- Polynomial kernel degree d
  - $K\left(x^{(i)}, x^{(j)}\right) = \left(x^{(i)^T} x^{(j)} + 1\right)^d$
- RBF kernel
  - $K\left(x^{(i)}, x^{(j)}\right) = e^{-\gamma \left\| x^{(i)} - x^{(j)} \right\|^2}$

https://scikit-learn.org/stable/modules/metrics.html

# Soft Margin

- What if the data is not linearly separable?
  - One solution is obviously to use non-linear kernels
- What if the data is not separable even with a kernel?
- Or, what if, we do not want to overfit?
- A solution is to relax the hard constraints a bit
- New objective function
  - $\min \frac{1}{2} w^T w + C \sum_{d \in D} \xi^{(d)}$
  - subject to
  - $y^{(d)}\left(w^T x^{(d)} + b\right) \geq +1 - \xi^{(d)}$
- Formulate Lagrangian Primal and Dual and solve

# Transparency – Linear Kernel

- $w$ can be calculated
  - $w = \sum_{d \in D} \alpha^{(d)} y^{(d)} x^{(d)}$
- A new object $x$ is classified as follows:
  - $\text{sign}(w^T x + b)$
- Transparency is similar to other linear models, e.g., logistic regression

# Transparency – Non-Linear Kernel

- $w$ cannot be calculated (except under some specific conditions)
- A new object $x$ is classified as follows:
  - $\text{sign}\left(\left(\sum_{d \in D} \alpha^{(d)} y^{(d)} K(x^{(d)}, x)\right) + b\right)$
- Model prediction can be explained in terms of
  - Similarities to support vectors: i.e, $K(x^{(d)}, x)$, and
  - Their weights $\alpha^{(d)}$
- For a "short" explanation, we need a sparse solution
  - i.e., $\alpha^{(d)}$ should be non-zero for only a handful of objects

# Scikit-Learn

- http://scikit-learn.org/stable/modules/svm.html