Support Vector Machines

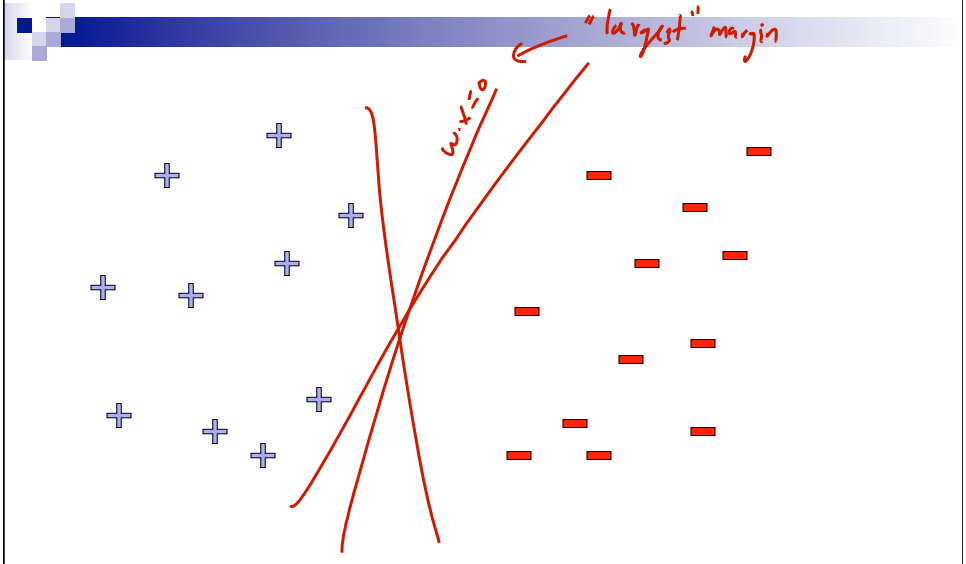
Machine Learning – CSE546

Carlos Guestrin

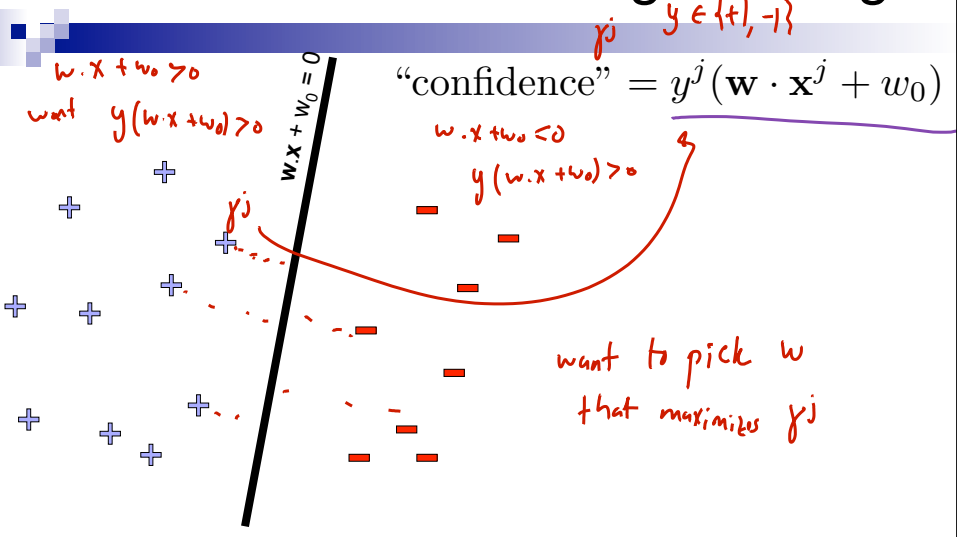
University of Washington

October 28, 2014

Linear classifiers – Which line is better?

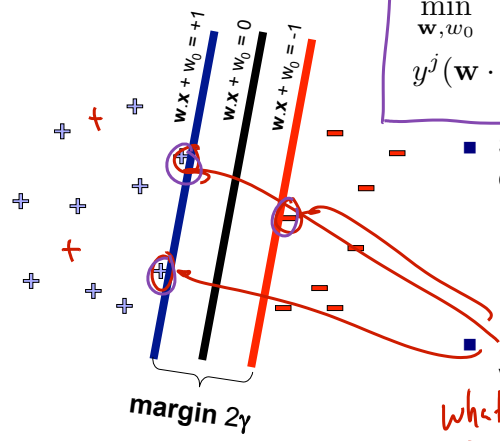


Pick the one with the largest margin!



$$\text{“confidence”} = y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0)$$

Support vector machines (SVMs)



$$\min_{\mathbf{w},w_0} \; ||w||_2^2$$

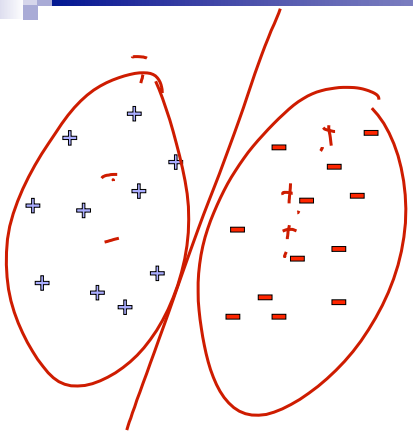
$$y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0) \geq 1, \forall j \in \{1,\dots,N\}$$

Primal SVM formulation

- Solve efficiently by many methods, e.g.,
  - quadratic programming (QP)
    - Well-studied solution algorithms
  - Stochastic gradient descent
- Hyperplane defined by support vectors

what if I forget all data but the support vectors ⇒ same solution!

What if the data is not linearly separable?



Use features of features of features….

→ Kernels

What if the data is still not linearly separable?

Constraint violation

or outliers

$$\min_{\mathbf{w},w_0} \; ||w||_2^2$$

$$(z)_+ = \begin{cases} 0 & \text{if } z \le 0 \\ z & \text{otherwise} \end{cases}$$

$$\rightarrow \; y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0) \ge 1 \qquad ,\forall j \leftarrow$$

$y^j(w\cdot x^j + w_0) < 1$

not correctly classified

- If data is not linearly separable, some points don't satisfy margin constraint:

$\exists j \;\; y^j(w\cdot x^j + w_0) < 1 \Rightarrow 1 - y^j(w\cdot x^j + w_0) > 0$

- How bad is the violation?

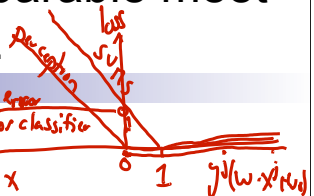
$$\text{constraint violation} = \begin{cases} 0 & \text{if } 1 - y^j(w\cdot x^j + w_0) \le 0 \\ 1 - y^j(w\cdot x^j + w_0) & \text{otherwise} \end{cases} = \left(1 - y^j(w\cdot x^j + w_0)\right)_+$$

- Tradeoff margin violation with ||**w**||:

$$\min_{w,w_0} \underbrace{||w||_2^2}_{\text{margin}} + C \underbrace{\sum_{j=1}^{N} \left(1 - y^j(w\cdot x^j + w_0)\right)_+}_{\text{slacks}}$$

penalty $C > 0$

SVMs for Non-Linearly Separable meet my friend the Perceptron…



- Perceptron was minimizing the hinge loss:

$$\min_{w, w_0} \sum_{j=1}^{N} \left(-y^j(\mathbf{w} \cdot \mathbf{x}^j + w_0)\right)_+$$

hinge loss

- SVMs minimizes the regularized hinge loss!!

$$||\mathbf{w}||_2^2 + C \sum_{j=1}^{N} \left(1 - y^j(\mathbf{w} \cdot \mathbf{x}^j + w_0)\right)_+$$

SVMs convention

regularization term

like $1/\lambda$

convex upper bound on train error ⇒ "easy" to optimize

Stochastic Gradient Descent for SVMs

- Perceptron minimization:

$$\sum_{j=1}^{N} \left(-y^j(\mathbf{w} \cdot \mathbf{x}^j + w_0)\right)_+$$

- SGD for Perceptron:

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \mathbb{1}\left[y^{(t)}(\mathbf{w}^{(t)} \cdot \mathbf{x}^{(t)}) \le 0\right] y^{(t)}\mathbf{x}^{(t)}$$

mistake

$\eta = 1$

$\nabla\left(-y^{(t)}(w^{(t)} x^{(t)} + w_0)\right)$

update weight

- SVMs minimization:

$$||\mathbf{w}||_2^2 + C \sum_{j=1}^{N} \left(1 - y^j(\mathbf{w} \cdot \mathbf{x}^j + w_0)\right)_+$$

sum over "mistakes" $x^{(j)}$

- SGD for SVMs:

$w^{(t+1)} \leftarrow w^{(t)}$

$+ \eta \left[- 2 w^{(t)} + C \mathbb{1}\left[y^{(t)}(w^{(t)} \cdot x^{(t)} + w_0^{(t)}) \le 1\right] y^{(t)} x^{(t)} \right]$

step size is important

margin violation

can use kernels because $w(t)$ can be expressed as a weighted

What you need to know

- Maximizing margin
- Derivation of SVM formulation
- Non-linearly separable case
  - Hinge loss
  - A.K.A. adding slack variables
- SVMs = Perceptron + L2 regularization
- Can also use kernels with SVMs
- Can optimize SVMs with SGD
  - Many other approaches possible