≈ perceptron + regularization

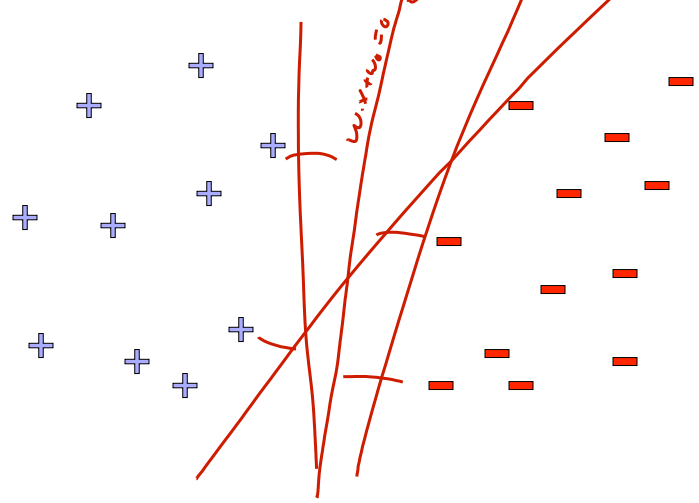
Support Vector Machines

Machine Learning – CSE446
Carlos Guestrin
University of Washington

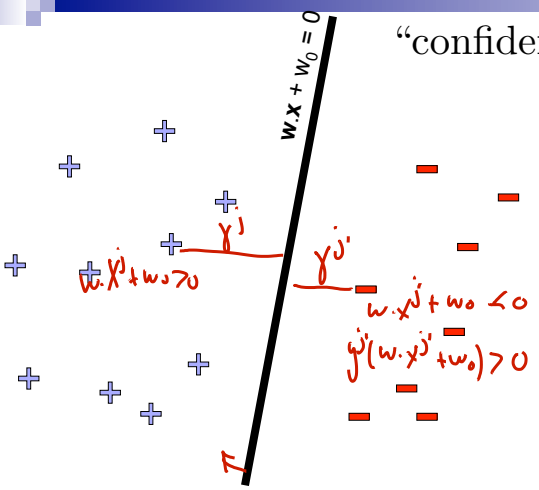
May 6, 2013

Linear classifiers – Which line is better?

furthest away from nearby points ≡ margin maximizing



Pick the one with the largest margin!

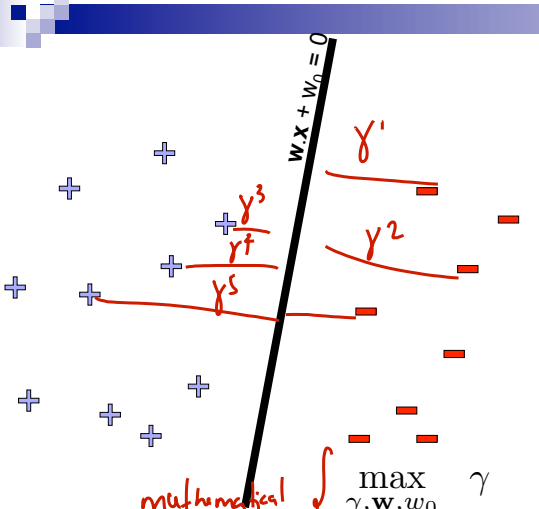


$$\text{“confidence”} = y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0)$$

$w\cdot x^j = \sum_{i=1}^{k} w_i x_i^j$

Maximize the margin

maximize worst case margin:



$$\max_{w,w_0} \min_j \gamma^j$$

$y^j(w\cdot x^j + w_0) = \gamma^j$

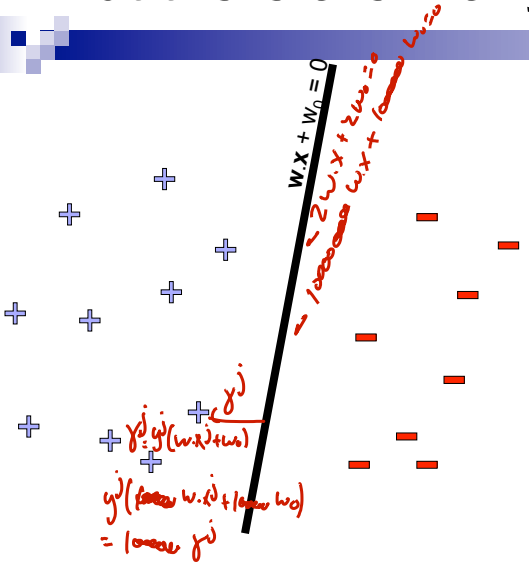
find me w, w_0 that gives max margin γ for all points j

mathematical program

$$\max_{\gamma,\mathbf{w},w_0} \gamma$$

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

But there are many planes…



γ will increase in magnitude as we scale w, w_0

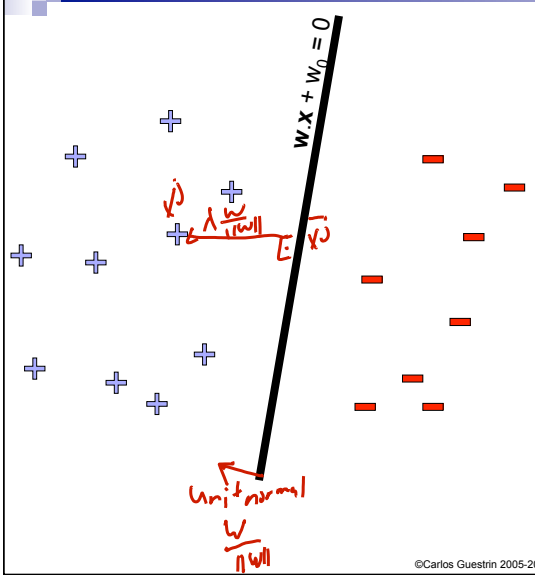
$\Rightarrow$ not well specified

Review: Normal to a plane

$$\mathbf{x}^j = \bar{\mathbf{x}}^j + \lambda \frac{\mathbf{w}}{||\mathbf{w}||}$$

projection onto plane

margin

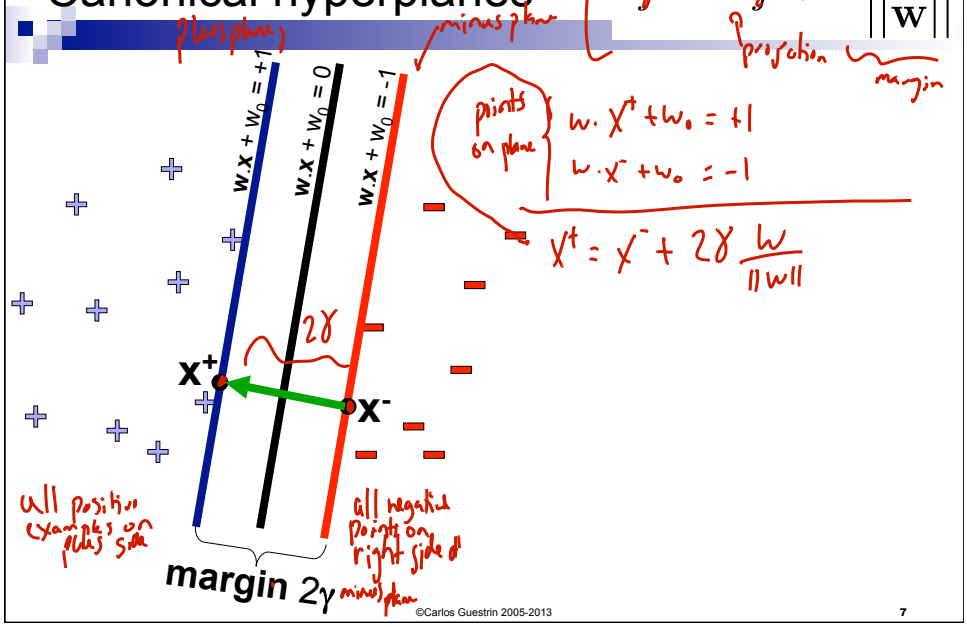


A Convention: Normalized margin – Canonical hyperplanes

$$\mathbf{x}_j = \bar{\mathbf{x}}_j + \lambda \frac{\mathbf{w}}{||\mathbf{w}||}$$

projection

margin



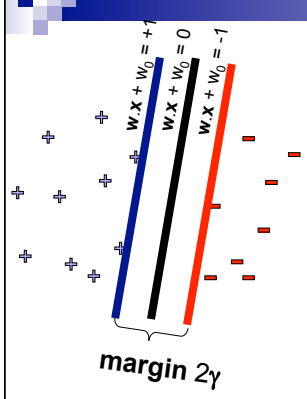
points on plane

$w \cdot x^+ + w_0 = +1$

$w \cdot x^- + w_0 = -1$

$x^+ = x^- + 2\gamma \frac{w}{||w||}$

Margin maximization using canonical hyperplanes



Unnormalized problem:

$$\max_{\gamma, \mathbf{w}, w_0} \quad \gamma$$

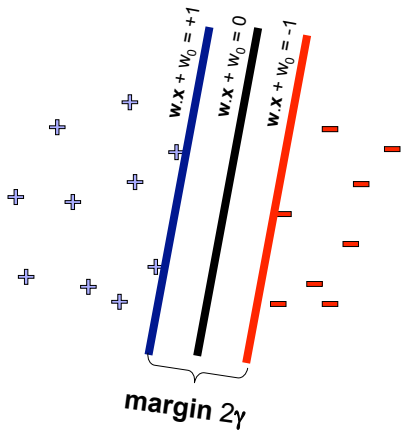
$$y^j(\mathbf{w} \cdot \mathbf{x}^j + w_0) \geq \gamma, \forall j \in \{1, \ldots, N\}$$

Normalized Problem:

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

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

Support vector machines (SVMs)

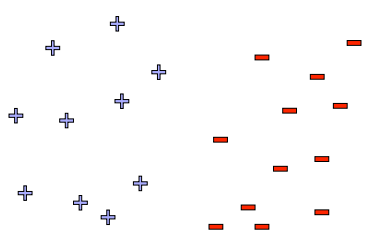


$$\min_{\mathbf{w},w_0} \quad ||w||_2^2$$

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

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

What if the data is not linearly separable?

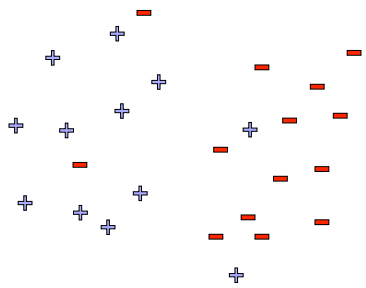


Use features of features of features of features….

What if the data is still not linearly separable?

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

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



- If data is not linearly separable, some points don't satisfy margin constraint:
- How bad is the violation?
- Tradeoff margin violation with ||**w**||:

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

- Perceptron was minimizing the hinge loss:

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

- SVMs minimizes the regularized hinge loss!!

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

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:

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

- SVMs minimization:

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

- SGD for SVMs:

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 optimize SVMs with SGD
  - Many other approaches possible

Big Picture

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 6, 2013

What you have learned thus far

- Learning is function approximation
- Point estimation
- Regression
- LASSO
- Logistic regression
- Bias-Variance tradeoff
- Regularization
- Decision trees
- Cross validation
- Boosting
- Instance-based learning
- Online learning
- Perceptron
- SVMs
- Kernel trick

Review material in terms of…

- Types of learning problems
- Hypothesis spaces
- Loss functions
- Optimization algorithms

ML Pipeline

Attributes/ Observations	Features/ Basis Functions	Task	Hypothesis Class/ Model	Algorithm/ Optimization Method

Learning Task/Measuring Error

TASK	LOSS FUNCTIONS
Regression	
Classification	
Density Estimation	

Hypothesis Classes & Decision Boundaries

Simple Linear Model

Linear Model with
Higher-Order Features or Kernels

Nearest Neighbors

Boosting

The Power of Regularization

Overfitting
Bias/Variance Tradeoff

Regularization

Your Midterm…

- Content: Everything up to today…
- Only 50mins, so arrive early and settle down quickly
- "Open book"
  - Textbook, Course notes, Personal notes
- No:
  - Computer, phone, other materials,…
- The exam:
  - Covers key concepts and ideas, work on understanding the big picture, and differences between methods