≈ perceptron + regularization

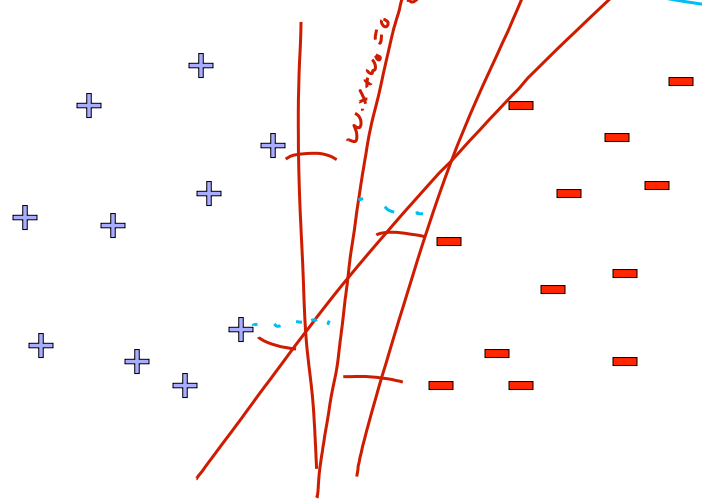
Support Vector Machines

Machine Learning – CSE446
Carlos Guestrin
University of Washington

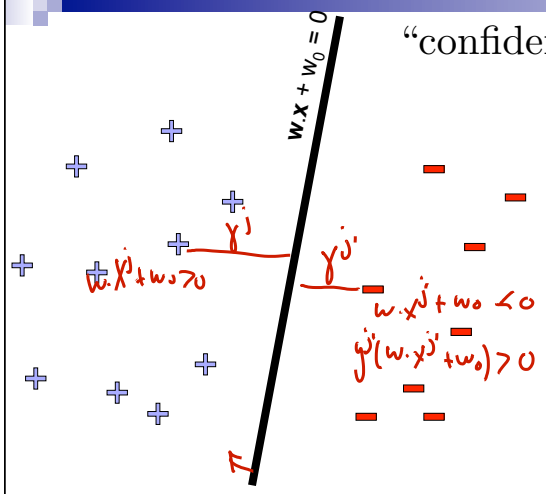
May 6, 2013

Linear classifiers – Which line is better?

farthest away from nearby points ≡ margin maximizing



Pick the one with the largest margin!



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

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

positive if no mistake

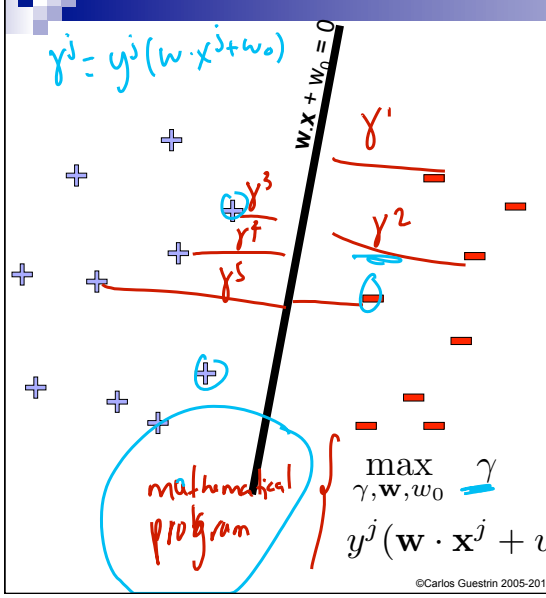
Maximize the margin

maximize worst case margin:

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

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

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



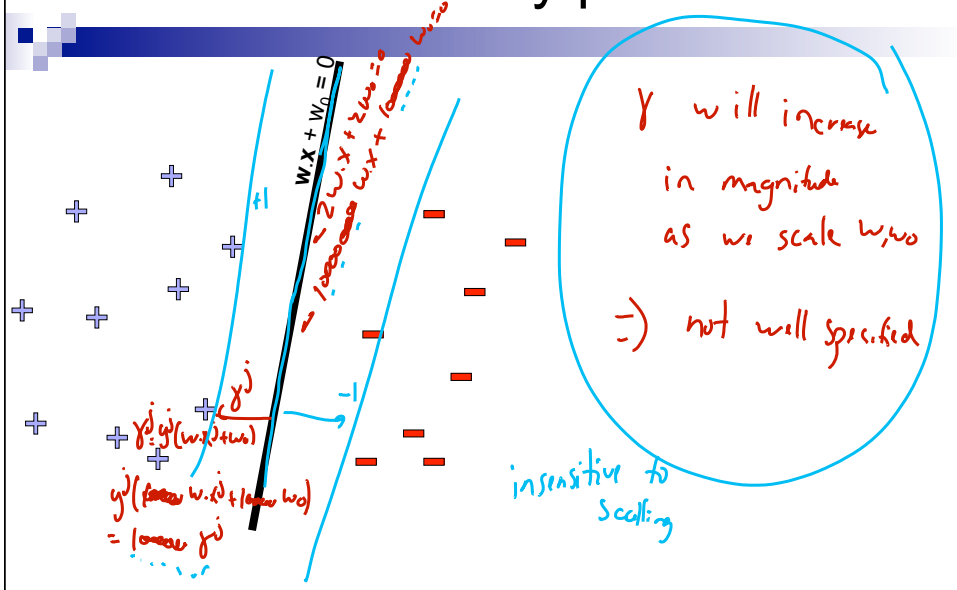
find me w, w_0 that gives max margin γ for all points j

mathematical program

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

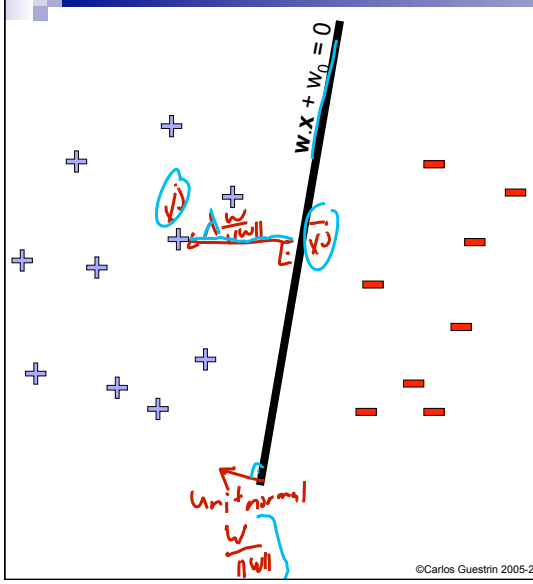
$$y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0) \geq \gamma, \forall j \in \{1,\ldots,N\}$$

But there are many planes…



Review: Normal to a plane

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



A Convention: Normalized margin – Canonical hyperplanes

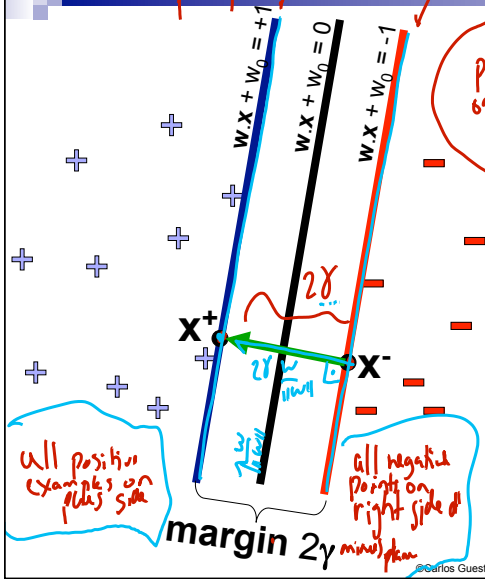
$$\mathbf{x}_j = \bar{\mathbf{x}}_j - \lambda \frac{\mathbf{w}}{||\mathbf{w}||}$$

$||w||^2 = w \cdot w$

$||w|| = \sqrt{\sum_i w_i^2}$

projection

margin



points on plane

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

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

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

$w \cdot \left(x^- + 2\gamma \frac{w}{||w||}\right) + w_0 = +1$

$\underbrace{w \cdot x^- + w_0}_{-1} + 2\gamma \frac{w \cdot w}{||w||} = +1$

$-1 + 2\gamma ||w|| = +1$

$\gamma = \frac{1}{||w||}$

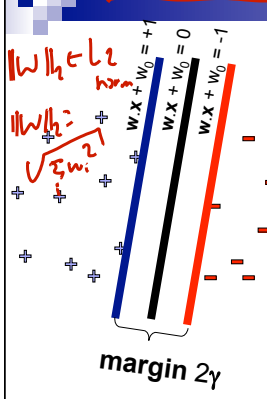
margin is inversely prop. to norm of w

©Carlos Guestrin 2005-2013 7

Margin maximization using canonical hyperplanes

$\max_x \frac{1}{f(x)} \equiv \min_x f(x)$

$\equiv \min_x f(x)^2$



$||w||_2 \leftarrow L_2$ norm

$||w||_2 = \sqrt{\sum_i w_i^2}$

Unnormalized problem:

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

bad, because unbounded

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

Normalized Problem:

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

canonical hyperplanes

$\max \gamma \equiv \max \frac{1}{||w||}$

$\equiv \min ||w|| \equiv \min ||w||_2^2$

just because derivatives will be simpler

standard SVM problem

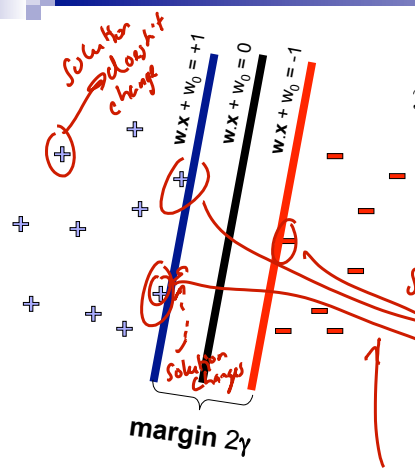
a.k.a. norm minimization or regularization!

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

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

©Carlos Guestrin 2005-2013 8

Support vector machines (SVMs)



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

find hyperplane with maximum margin

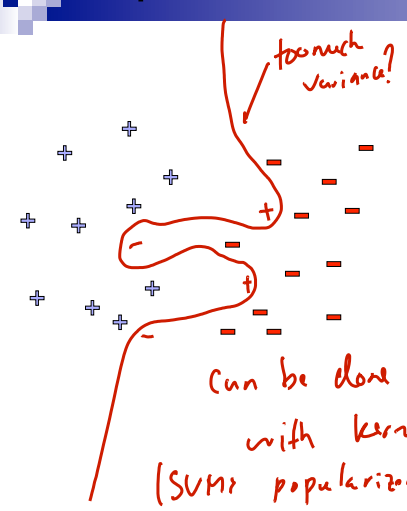
$$y^j(\mathbf{w}\cdot\mathbf{x}^j + w_0) \geq 1, \forall j \in \{1,\ldots,N\}$$

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

support vectors

only perturbations of these points will change solution

What if the data is not linearly separable?



Use features of features of features of features….

$$\phi(x) = \begin{pmatrix} x^2 \\ x \\ e^x \sin x \\ \vdots \end{pmatrix}$$

What if the data is still not linearly separable?

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

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

of this constraint violation

still make some mistakes

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

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

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

like the hinge loss

- How bad is the violation?

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

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

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

Now support vectors are
- on margin, or
- misclassified points

$C>0$

SVM objective

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

- Perceptron was minimizing the hinge loss:

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

both hinge loss

just convention

- SVMs minimizes the regularized hinge loss!!

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

extra regularization!!

$\nabla_w \|w\|_2^2 = \nabla_w w \cdot w = 2w \leftarrow$ direction of increase but we go towards $-\nabla_w \|w\|_2^2$

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:

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

update weights

step size = ?

- 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:

$$w^{(t+1)} \leftarrow w^{(t)} + \eta\left(C\,\mathbb{1}\left((1-y^{(t)}(w \cdot x^{(t)} + w_0)) > 0\right) y^{(t)} x^{(t)} - 2w^{(t)}\right)$$

step size needs to played with

pick C, η by cross validation

other way is the

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

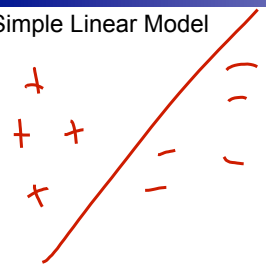
Attributes/ Observations	Features/ Basis Functions	Task	Hypothesis Class/ Model	Algorithm/ Optimization Method
X age gender wage ⋮	$h_i(x)$ $\phi(x)$ age^2 age ⋮	regression X or $\phi(x)$ $\to \mathbb{R}$ classification X or $\phi(x) \to \{1,\dots,r\}$ Density estimation X or $\phi(x)$ $\to$ probability Logistic regression	Linear models $w \cdot x$ NN, DTs Boosting	optimize a loss function - Gradient - set gradient $=0$ closed form - Stochastic Coordinate descent - SGD - search decision trees

Learning Task/Measuring Error

TASK	LOSS FUNCTIONS
Regression	Squared error
Classification	log loss for logistic regression hinge loss } approx of 0/1 loss ← intractable
Density Estimation	log loss for LR

Hypothesis Classes & Decision Boundaries

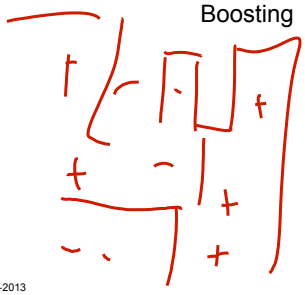
Simple Linear Model



Linear Model with Higher-Order Features or Kernels

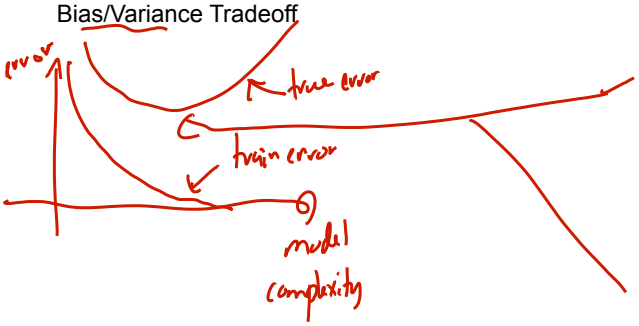
Nearest Neighbors

Boosting



The Power of Regularization

Overfitting
Bias/Variance Tradeoff



Regularization

$\|w\|_2^2 \leftarrow L_2$ regularization

$\|w\|_1$

find sparse solutions

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