SOFT LARGE MARGIN CLASSIFIERS

David Kauchak
CS 451 – Fall 2013

Admin

Assignment 5

Midterm

Friday's class will be in MBH 632

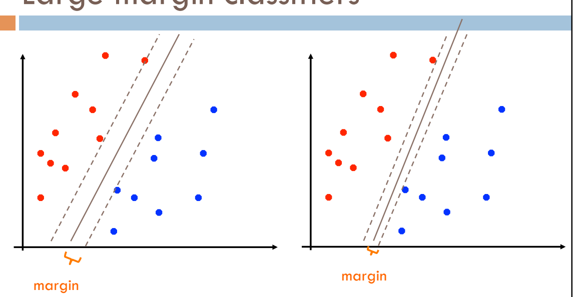
CS lunch talk Thursday

Java tips for the data

-Xmx
 -Xmx2g

http://www.youtube.com/watch?v=u0VoFU82GSw

Large margin classifiers



The margin of a classifier is the distance to the closest points of either class

Large margin classifiers attempt to maximize this

Support vector machine problem

$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

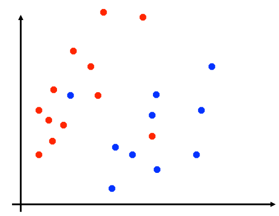
This is a a quadratic optimization problem

Maximize/minimize a quadratic function

Subject to a set of linear constraints

Many, many variants of solving this problem (we'll see one in a bit)

Soft Margin Classification



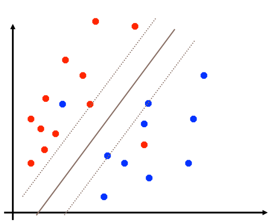
$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

What about this problem?

Soft Margin Classification



$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

We'd like to learn something like this, but our constraints won't allow it ☹

Slack variables

$$\min_{w,b} \quad \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

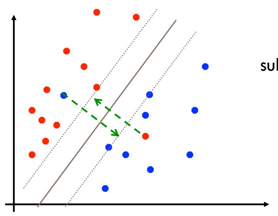
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \geq 0$$

slack variables
(one for each example)

What effect does this have?

Slack variables



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

slack penalties

Slack variables

margin

trade-off between margin maximization and penalization

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

penalized by how far from "correct"

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

allowed to make a mistake

$$\varsigma_i \geq 0$$

Soft margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

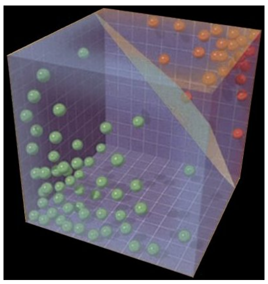
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

Still a quadratic optimization problem!

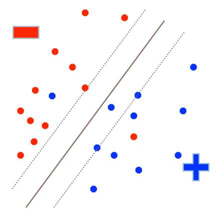
Demo

http://cs.stanford.edu/people/karpathy/svmjs/demo/

Solving the SVM problem



Understanding the Soft Margin SVM



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

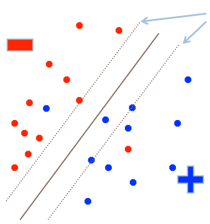
subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

Given the optimal solution, w, b:

Can we figure out what the slack penalties are for each point?

Understanding the Soft Margin SVM



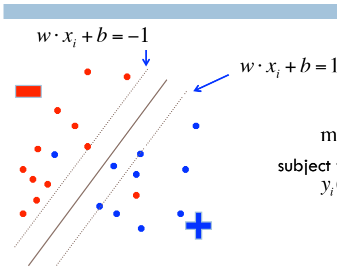
What do the margin lines represent wrt w,b?

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

Understanding the Soft Margin SVM



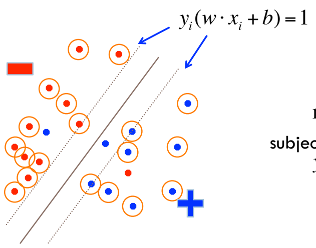
$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

Or: $\boxed{y_i(w \cdot x_i + b) = 1}$

Understanding the Soft Margin SVM



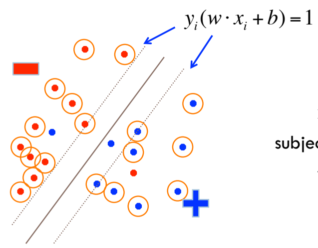
$$\min_{w,b} \; \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

What are the slack values for points outside (or on) the margin AND correctly classified?

Understanding the Soft Margin SVM



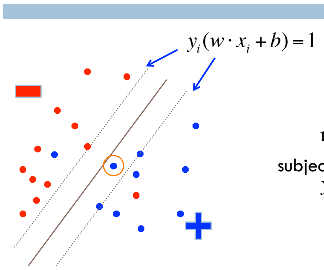
$$\min_{w,b} \; \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

0! The slack variables have to be greater than or equal to zero and if they're on or beyond the margin then $y_i(wx_i+b) \geq 1$ already

Understanding the Soft Margin SVM



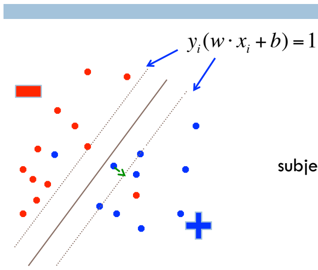
$$\min_{w,b} \; \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

What are the slack values for points inside the margin AND classified correctly?

Understanding the Soft Margin SVM



$$\min_{w,b} \; \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

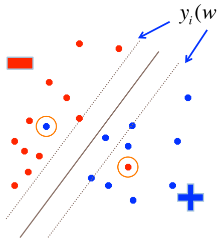
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

Difference from point to the margin. Which is?

$$\varsigma_i = 1 - y_i(w \cdot x_i + b)$$

Understanding the Soft Margin SVM

$y_i(w \cdot x_i + b) = 1$



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

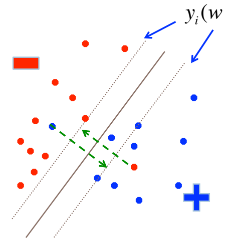
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \geq 0$$

What are the slack values for points that are incorrectly classified?

Understanding the Soft Margin SVM

$y_i(w \cdot x_i + b) = 1$



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

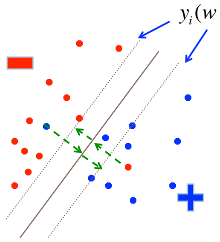
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \geq 0$$

Which is?

Understanding the Soft Margin SVM

$y_i(w \cdot x_i + b) = 1$



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

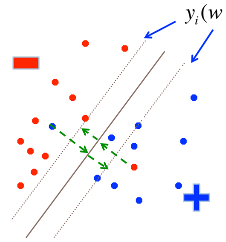
$$\varsigma_i \geq 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

?

Understanding the Soft Margin SVM

$y_i(w \cdot x_i + b) = 1$



$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

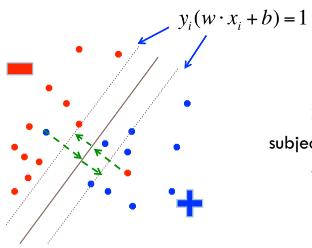
$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$

$$\varsigma_i \geq 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$ Why -?

Understanding the Soft Margin SVM



$$\min_{w,b} \ \|w\|^2 + C\sum_i \varsigma_i$$

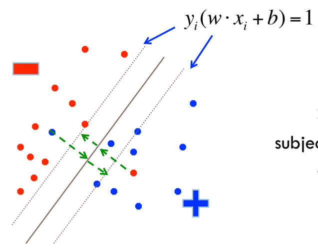
subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$?

Understanding the Soft Margin SVM



$$\min_{w,b} \ \|w\|^2 + C\sum_i \varsigma_i$$

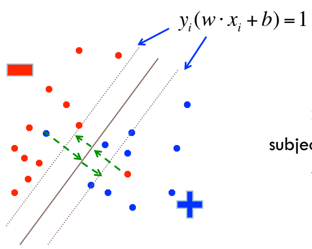
subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$-y_i(w \cdot x_i + b)$ 1

Understanding the Soft Margin SVM



$$\min_{w,b} \ \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

"distance" to the hyperplane *plus* the "distance" to the margin

$$\varsigma_i = 1 - y_i(w \cdot x_i + b)$$

Understanding the Soft Margin SVM

$$\min_{w,b} \ \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \geq 0$$

$$\varsigma_i = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \geq 1 \\ 1 - y_i(w \cdot x_i + b) & \text{otherwise} \end{cases}$$

Understanding the Soft Margin SVM

$$\varsigma_i = \begin{cases} 0 & \text{if } y_i(w \cdot x_i + b) \ge 1 \\ 1 - y_i(w \cdot x_i + b) & \text{otherwise} \end{cases}$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$
$$= \max(0, 1 - yy')$$

Does this look familiar?

Hinge loss!

0/1 loss: $l(y, y') = 1[yy' \le 0]$

Hinge: $l(y, y') = \max(0, 1 - yy')$

Exponential: $l(y, y') = \exp(-yy')$

Squared loss: $l(y, y') = (y - y')^2$

Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$

Do we need the constraints still?

Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \varsigma_i$$

subject to:

$$y_i(w \cdot x_i + b) \ge 1 - \varsigma_i \quad \forall i$$
$$\varsigma_i \ge 0$$

$$\varsigma_i = \max(0, 1 - y_i(w \cdot x_i + b))$$

$$\min_{w,b} \quad \|w\|^2 + C\sum_i \max(0, 1 - y_i(w \cdot x_i + b))$$

Unconstrained problem!

Understanding the Soft Margin SVM

$$\min_{w,b} \quad \|w\|^2 + C\sum_i loss_{hinge}(y_i, y_i')$$

Does this look like something we've seen before?

$$\mathrm{argmin}_{w,b} \sum_{i=1}^{n} loss(yy') + \lambda \; regularizer(w,b)$$

Gradient descent problem!

Soft margin SVM as gradient descent

$$\min_{w,b} \quad \|w\|^2 + C\sum_i loss_{hinge}(y_i, y_i')$$

multiply through by 1/C and rearrange

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \frac{1}{C}\|w\|^2$$

let $\lambda = 1/C$

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \lambda\|w\|^2$$

What type of gradient descent problem?

$$\mathrm{argmin}_{w,b} \sum_{i=1}^{n} loss(yy') + \lambda \; regularizer(w,b)$$

Soft margin SVM as gradient descent

One way to solve the soft margin SVM problem is using gradient descent

$$\min_{w,b} \quad \sum_i loss_{hinge}(y_i, y_i') + \lambda\|w\|^2$$

hinge loss

L2 regularization

Gradient descent SVM solver

- pick a starting point (w)
- repeat until loss doesn't decrease in all dimensions:
  - pick a dimension
  - move a small amount in that dimension towards decreasing loss (using the derivative)

$$w_i = w_i - \eta \frac{d}{dw_i}(loss(w) + regularizer(w,b))$$

$$w_j = w_j + \eta \sum_{i=1}^{n} y_i x_i 1[y_i(w \cdot x + b) < 1] - \eta\lambda w_j$$

hinge loss L2 regularization

Finds the largest margin hyperplane while allowing for a soft margin

Support vector machines

One of the most successful (if not the most successful) classification approach:

decision tree	About 2,160,000 results (**0.05** sec)
Support vector machine	About 1,960,000 results (**0.04** sec)
k nearest neighbor	About 746,000 results (**0.04** sec)
perceptron algorithm	About 84,300 results (**0.04** sec)

Google scholar

Trends over time