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

\[
\min_{w, b} \|w\|^2 \\
\text{subject to:} \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i
\]

This is 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

\[
\min_{w, b} \|w\|^2 \\
\text{subject to:} \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i
\]

What about this problem?

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

Soft Margin Classification

\[
\min_{w, b} \|w\|^2 \\
\text{subject to:} \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i
\]

Slack variables

\[
\min_{w, b} \|w\|^2 + C \sum \xi_i \\
\text{subject to:} \quad y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \\
\xi_i \geq 0
\]

slack variables (one for each example)

What effect does this have?
Slack variables

\[ \min_{w,b} \|w\|^2 + C \sum \xi_i \]
subject to:
\[ y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]
\[ \xi_i \geq 0 \]

allowed to make a mistake

trade-off between margin maximization and penalization

penalized by how far from “correct”

Soft margin SVM

\[ \min_{w,b} \|w\|^2 + C \sum \xi_i \]
subject to:
\[ y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]
\[ \xi_i \geq 0 \]

Still a quadratic optimization problem!

Demo

http://cs.stanford.edu/people/karpathy/svmjs/demo/
Solving the SVM problem

Understanding the Soft Margin SVM

\[ y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]

subject to:

\[ \sum \xi_i \leq C \]

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 \( w, b \)?

\[ w \cdot x_i + b = -1 \]

\[ w \cdot x_i + b = 1 \]

\[ y_i (w \cdot x_i + b) = 1 \]
What are the slack values for points outside (or on) the margin AND correctly classified?

Difference from point to the margin. Which is?

\( \xi_i = 1 - y_i (w \cdot x_i + b) \)
Understanding the Soft Margin SVM

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

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

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

Which is?

Why -?
Understanding the Soft Margin SVM

\[ y_i (w \cdot x_i + b) = 1 \]

subject to:
\[ y_i (w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]
\[ \xi_i \geq 0 \]

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

\[ -y_i (w \cdot x_i + b) \]

\[ \xi_i = 1 - y_i (w \cdot x_i + b) \]

\[ \min_{w,b} \|w\|^2 + C \sum_i \xi_i \]

10/16/13
Understanding the Soft Margin SVM

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

\[ \xi_i = \max(0, 1 - y_i(w \cdot x_i + b)) \]

\[ \xi_i = \max(0, 1 - yy') \]

Does this look familiar?

Hinge loss!

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

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

\[ \min_{w, b} \|w\|^2 + C \sum \xi_i \]

subject to:

\[ y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]

\[ \xi_i \geq 0 \]

Do we need the constraints still?

Understanding the Soft Margin SVM

\[ \min_{w, b} \|w\|^2 + C \sum \xi_i \]

subject to:

\[ y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \]

\[ \xi_i \geq 0 \]

\[ \min_{w, b} \|w\|^2 + C \sum \max(0, 1 - y_i(w \cdot x_i + b)) \]

Unconstrained problem!
Soft margin SVM as gradient descent

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

$$ \min_{w,b} \|w\|^2 + C \sum_i \text{loss}_{hinge}(y_i, y_i') $$

Does this look like something we’ve seen before?

$$ \arg\min_{w,b} \sum_i \text{loss}(y_i, y_i') + \lambda \text{regularizer}(w,b) $$

Gradient descent problem!

Multiply through by $1/C$ and rearrange:

$$ \min_{w,b} \|w\|^2 + C \sum_i \text{loss}_{hinge}(y_i, y_i') $$

Let $\lambda = 1/C$

What type of gradient descent problem?

$$ \arg\min_{w,b} \sum_i \text{loss}(y_i, y_i') + \lambda \text{regularizer}(w,b) $$

Soft margin SVM as gradient descent

Soft margin SVM as gradient descent

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)

$$ w_j = w_j - \eta \frac{d}{dw_j} (\text{loss}(w)+\text{regularizer}(w,b)) $$

$$ w_j = w_j + \eta \sum_i x_i [y_i (w \cdot x + b) < 1] - \eta \lambda w_j $$

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
- Support vector machine
- k nearest neighbor
- perceptron algorithm

Trends over time