LARGE MARGIN CLASSIFIERS

David Kauchak
CS 158 – Fall 2016

Which hyperplane?

Two main variations in linear classifiers:
- which hyperplane they choose when the data is linearly separable
- how they handle data that is not linearly separable

Linear approaches so far

Perceptron:
- separable:
- non-separable:

Gradient descent:
- separable:
- non-separable:

Admin

Assignment 5
- back soon
- write tests for your code!
- variance scaling uses standard deviation
- for this class

Assignment 6

Midterm

Course feedback
- Thank you!
- We’ll go over it at the end of class today or the beginning of next class

\[
\text{variance scaling uses standard deviation} = \sqrt{\sum \left( x - \text{mean}(\text{data}) \right)^2}
\]

\[
\text{for this class}
\]

\[
\text{variance scaling uses standard deviation} = \sqrt{\frac{1}{n} \sum \left( x - \text{mean}(\text{data}) \right)^2}
\]

\[
\text{for this class}
\]

Which hyperplane?

Two main variations in linear classifiers:
- which hyperplane they choose when the data is linearly separable
- how they handle data that is not linearly separable

Linear approaches so far

Perceptron:
- separable:
- non-separable:

Gradient descent:
- separable:
- non-separable:
Linear approaches so far

Perceptron:
- separable:
  - finds some hyperplane that separates the data
- non-separable:
  - will continue to adjust as it iterates through the examples
  - final hyperplane will depend on which examples it saw recently

Gradient descent:
- separable and non-separable
  - finds the hyperplane that minimizes the objective function (loss + regularization)

Which hyperplane is this?

Which hyperplane would you choose?

Large margin classifiers

Choose the line where the distance to the nearest point(s) is as large as possible

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 vectors

For any separating hyperplane, there exist some set of “closest points”
These are called the support vectors
For n dimensions, there will be at least n+1 support vectors

Measuring the margin

The margin is the distance to the support vectors, i.e. the “closest points”, on either side of the hyperplane

What are the equations for the margin lines?

- Negative examples: $w \cdot x_i + b < 0$
- Positive examples: $w \cdot x_i + b > 0$
What is $c$?

We know they’re the same distance apart (otherwise, they wouldn’t be support vectors!)

Depends! If we scale $w$, we vary the constant without changing the separating hyperplane.

Larger $w$ result in larger constants.

Smaller $w$ result in smaller constants.
For now, let’s assume $c = 1$.

Distance from the hyperplane

How far away is this point from the hyperplane?

Distance from the hyperplane

How far away is this point from the hyperplane?
Distance from the hyperplane

How far away is the point from the hyperplane?

\[ w = (1, 2) \]

\[ d(x) = w \cdot x + b \]

Does that seem right? What's the problem?

\[ w = (1, 2) \]

\[ d(x) = w \cdot x + b \]

\[ = w_1 x_1 + w_2 x_2 + b \]

\[ = 1 \cdot 1 + 1 \cdot 2 + 0 \]

\[ = 3? \]

Distance from the hyperplane

How far away is this point from the hyperplane?

\[ t_1 \]

\[ t_2 \]

\[ w = (2, 4) \]

\[ d(x) = w \cdot x + b \]

\[ = w_1 x_1 + w_2 x_2 + b \]

\[ = 2 \cdot 1 + 4 \cdot 2 + 0 \]

\[ = 10? \]
How far away is this point from the hyperplane?

Distance from the hyperplane

\[ w = (1, 2) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

The magnitude of the weight vector doesn’t matter

Distance from the hyperplane

\[ w = (2, 4) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

Distance from the hyperplane

\[ w = (0.5, 1) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

Distance from the hyperplane

\[ w = (1, 2) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

Distance from the hyperplane

\[ w = (1.5, 2) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

Distance from the hyperplane

\[ w = (0.5, 1) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]

Distance from the hyperplane

\[ w = (1.5, 2) \]

\[ d(x) = \frac{w \cdot x + b}{\|w\|} \]
Measuring the margin

For now, let’s just assume \( c = 1 \).

What is this distance?

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

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

Large margin classifier setup

Select the hyperplane with the largest margin where the points are classified correctly and outside the margin!

Setup as a constrained optimization problem:

\[
\max_{w,b} \text{margin}(w,b) \quad \text{subject to:} \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i
\]

what does this say?

\[
\frac{w \cdot x_i + b}{||w||} = \frac{1}{||w||}
\]
Maximizing the margin

\[
\begin{align*}
\min_{w,b} & \quad ||w|| \\
\text{subject to:} & \quad y_i(w \cdot x_i + b) \geq 1 \quad \forall i
\end{align*}
\]

Maximizing the margin is equivalent to minimizing \( ||w|| \) (subject to the separating constraints)

The constraints:
1. make sure the data is separable
2. encourages \( w \) to be larger (once the data is separable)

Measuring the margin

For now, let's just assume \( c = 1 \).

\[
\begin{align*}
w \cdot x_i + b = -1 \\
w \cdot x_i + b = 1
\end{align*}
\]

Claim: it does not matter what \( c \) we choose for the SVM problem. Why?

What is this distance?
Measuring the margin

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

Maximizing the margin

\[ \min_{w, b} \frac{\|w\|}{c} \]
subject to:
\[ y_i (w \cdot x_i + b) \geq c \forall i \]

vs.

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

What's the difference?

Learn the exact same hyperplane just scaled by a constant amount.

Because of this, often see it with \( c = 1 \)

For those that are curious...

\[ \frac{1}{c} \sqrt{w_1^2 + w_2^2 + \ldots + w_m^2 + b^2} \]
\[ = \sqrt{\frac{w_1^2 + w_2^2 + \ldots + w_m^2}{c^2}} \]
\[ = \sqrt{\frac{w_1^2 + w_2^2 + \ldots + w_m^2}{c^2}} \]
\[ = \sqrt{\frac{w_1^2}{c^2} + \frac{w_2^2}{c^2} + \ldots + \frac{w_m^2}{c^2}} \]

scaled version of \( w \)
Maximizing the margin: the real problem

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

Why the squared?

Minimizing \( \|w\| \) is equivalent to minimizing \( \|w\|^2 \)

The sum of the squared weights is a convex function!

Support vector machine problem

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

This is a version of 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

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

What about this problem?
Soft Margin Classification

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

subject to:
\[ \min_{w,b} \|w\|^2 \]

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

Slack variables

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

subject to:
\[ \min_{w,b} \|w\|^2 + C \sum \varsigma_i \]

What effect does this have?

Slack variables

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

subject to:
\[ \min_{w,b} \|w\|^2 + C \sum \varsigma_i \]

slack penalties

\[ \varsigma_i \geq 0 \]

Slack variables

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

subject to:
\[ \min_{w,b} \|w\|^2 + C \sum \varsigma_i \]

slack variables

\[ \varsigma_i \geq 0 \]

slack penalties

\[ \varsigma_i \geq 0 \]

What effect does this have?

Slack variables

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

subject to:
\[ \min_{w,b} \|w\|^2 + C \sum \varsigma_i \]

penalized by how far from “correct”

allowed to make a mistake
Soft margin SVM

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

Still a quadratic optimization problem!

Demo

http://cs.stanford.edu/people/karpathy/wnmp/demos/

Solving the SVM problem

Understanding the Soft Margin SVM

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?**

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

On: \(y_i(w \cdot x_i + b) = 1\)

Understanding the Soft Margin SVM

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

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

On: \(y_i(w \cdot x_i + b) = 1\)

01. The slack variables have to be greater than or equal to zero and if they're on or beyond the margin then \(y_i(w \cdot x_i + b) \geq 1\) already.
Understanding the Soft Margin SVM

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

\[ \min_{\omega, b} \|\omega\| + C \sum \xi_i \]

subject to:

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

\[ \xi_i \geq 0 \]

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

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

Difference from the point to the margin. Which is?

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

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

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

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

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

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

Why -?

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

1
Understanding the Soft Margin SVM

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

min \[ \|w\| + C \sum \varsigma_i \]
subject to:
\[ y_i(w \cdot x_i + b) \geq 1 - \varsigma_i \forall i \]
\[ \varsigma_i \geq 0 \]

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

"distance" to the hyperplane plus the "distance" to the margin
\[ \varsigma_i = 1 - y_i(w \cdot x_i + b) \]

Does this look familiar?

Hinge loss!

0/1 loss: \[ l(y, y') = \begin{cases} 1 & \text{if } yy' < 0 \\ 0 & \text{otherwise} \end{cases} \]

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

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

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

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

Unconstrained problem!

Understanding the Soft Margin SVM

\[ \arg \min_{w, b} \sum_{i=1} \text{loss}(y_i, y_i') + \lambda \text{regularizer}(w, b) \]

Does this look like something we’ve seen before?

Soft margin SVM as gradient descent

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

multiply through by 1/C and rearrange

\[ \min_{w, b} \sum_i \text{loss}_{\text{margin}}(y_i, y_i') + \frac{1}{C} \|w\|^2 \]

let \( \lambda = 1/C \)

\[ \min_{w, b} \sum_i \text{loss}_{\text{margin}}(y_i, y_i') + \lambda \|w\|^2 \]

What type of gradient descent problem?

\[ \arg \min_{w, b} \sum_{i=1} \text{loss}(y_i, y_i') + \lambda \text{regularizer}(w, b) \]
One way to solve the soft margin SVM problem is using gradient descent:

\[ \min_{w,b} \sum_i \text{loss}_i(y_i, y_i') + \lambda \|w\|^2 \]

- hinge loss
- L2 regularization

**Gradient descent SVM solver**

1. **pick a starting point** \(w\)
2. repeat until loss doesn’t decrease in all dimensions:
   1. **pick a dimension**
   2. move a small amount in that dimension towards decreasing loss (using the derivative)

\[ w_j = w_j + \eta \left[ y_i (w \cdot x + b) < 1 \right] - \eta \lambda w_j \]

Finds the largest margin hyperplane while allowing for a soft margin.
Trends over time