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:
  - 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 would you choose?

Which hyperplane is this?

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

Measuring the margin

What are the equations for the margin lines?

Mathematically, the margin can be measured using the following equations:

Negative examples:

\[ w \cdot x_i + b < 0 \]

Positive examples:

\[ w \cdot x_i + b > 0 \]
Measuring the margin

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

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

What is \( c \)?

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

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

Larger \( w \) result in larger constants

Smaller \( w \) result in smaller constants
Measuring the margin

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 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 this point from the hyperplane?

\( d(x) = \langle w, x \rangle + b \)

Distance from the hyperplane

Does that seem right? What’s the problem?

\( d(x) = \langle w, x \rangle + b \)  
\( = w_1 x_1 + w_2 x_2 + b \)  
\( = 1 \cdot 1 + 1 \cdot 2 + 0 \)  
\( = 3 \) ?

Distance from the hyperplane

How far away is the point from the hyperplane?

\( d(x) = \langle w, x \rangle + b \)

Distance from the hyperplane

How far away is the point from the hyperplane?

\( d(x) = \langle w, x \rangle + b \)  
\( = w_1 x_1 + w_2 x_2 + b \)  
\( = 2 \cdot 1 + 4 \cdot 2 + 0 \)  
\( = 10 \) ?
Distance from the hyperplane

How far away is this point from the hyperplane?

How far away is this point from the hyperplane?

The magnitude of the weight vector doesn’t matter

The magnitude of the weight vector doesn’t matter

$\mathbf{w} = (1, 2)$

$d(x) = \frac{\mathbf{w} \cdot \mathbf{x} + b}{\|\mathbf{w}\|}$

$\mathbf{w} = (1, 2)$

$d(x) = \frac{\mathbf{w} \cdot \mathbf{x} + b}{\|\mathbf{w}\|}$

$= \frac{(1 \cdot 1 + 1 \cdot 2) + 0}{\sqrt{5}}$

$= \frac{3}{\sqrt{5}}$

$= 1.34$

$d(x) = \frac{\mathbf{w} \cdot \mathbf{x} + b}{\|\mathbf{w}\|}$

$= \frac{(1 \cdot 0.5 + 1 \cdot 1) + 0}{\sqrt{5}}$

$= \frac{0.5}{\sqrt{5}}$

$= 0.22$

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

For now, let's just assume $c = 1$.

$w \cdot x_i + b = -1$

What is this distance?

$w \cdot x_i + b = 1$

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:

$$\max_{w,b} \quad \text{margin}(w,b)$$

subject to:

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

what does this say?
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:

$$\max_{w,b} \frac{1}{\|w\|} \quad \text{subject to:} \quad y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

Maximizing the margin

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

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

Measuring the margin

The minimization criterion wants $w$ to be as small as possible

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

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

For now, let’s just assume $c = 1$.

Claim: it does not matter what $c$ we choose for the SVM problem. Why?
Measuring the margin

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

What is this distance?

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

Maximizing the margin

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

vs.

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

What's the difference?

Maximizing the margin

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

vs.

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

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...

Maximizing the margin: the real problem

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

Why the squared?

Maximizing the margin: the real problem

\[ \begin{align*}
\min_{w,b} & \quad \frac{1}{2} \sum_{i=1}^m w_i^2 \\
\text{subject to:} & \quad y_i (w \cdot x_i + b) \geq 1 \quad \forall i
\end{align*} \]

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

The sum of the squared weights is a convex function!

Support vector machine problem

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

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

What about this problem?

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

subject to:

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

\[ \xi_i \geq 0 \]

What effect does this have?
Slack variables

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

trade-off between margin maximization and penalization

margin

penalized by how far from “correct”

allowed to make a mistake

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

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

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

subject to:

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

\[ \xi_i \geq 0 \]

Given the optimal solution, \( w, b \):

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

What do the margin lines represent with respect to \( w, b \)?

Or:

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

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

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

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

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

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

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

If 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

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

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

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

Which is?

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

Why？

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

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

\[ 1 \]

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

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

Does this look familiar?
Hinge loss!

0/1 loss: \( l(y, y') = 1 \) if \( yy' \leq 0 \)

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

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

Do we need the constraints still?

Understanding the Soft Margin SVM

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

Understanding the Soft Margin SVM

\[
\begin{align*}
\min_{w,b} & \quad ||w||^2 + C \sum_i \text{loss}_{simp}(y_i, y'_i) \\
\text{subject to:} & \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i \\
& \quad \xi_i = \max(0, 1 - y_i(w \cdot x_i + b))
\end{align*}
\]

Does this look like something we’ve seen before?

\[
\arg\min_{w,b} \sum_i \text{loss}(yy') + \lambda \text{regularizer}(w, b)
\]

Gradient descent problem!
Soft margin SVM as gradient descent

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

multiply through by \(1/C\) and rearrange

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

let \(\lambda = 1/C\)

What type of gradient descent problem?

\[
\arg\min_{w,b} \sum_i \text{loss}(y y') + \lambda \text{regularizer}(w,b)
\]

Soft margin SVM as gradient descent

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

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

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)

\[
w_i = w_i - \eta \frac{d}{dw_i} (\text{loss}(w) + \text{regulizer}(w,b))
\]

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

hinge loss  
L2 regularization

Support vector machines: 2013

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

<table>
<thead>
<tr>
<th></th>
<th>2013</th>
<th>2016</th>
<th>2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>decision tree</td>
<td>About 2.180,000 results</td>
<td>About 2,480,000</td>
<td>About 3,000,000 r</td>
</tr>
<tr>
<td>Support vector machine</td>
<td>About 1,980,000 results</td>
<td>About 2,480,000</td>
<td>About 3,000,000</td>
</tr>
<tr>
<td>k nearest neighbor</td>
<td>About 1,740,000 results</td>
<td>About 979,000</td>
<td>About 1,380,000</td>
</tr>
<tr>
<td>perceptron algorithm</td>
<td>About 843,300 results</td>
<td>About 104,000</td>
<td>About 153,000 r</td>
</tr>
</tbody>
</table>

Finds the largest margin hyperplane while allowing for a soft margin