LARGE MARGIN CLASSIFIERS

David Kauchak
CS 451 – Fall 2013

Admin

Assignment 5

Midterm

Download from course web page when you're ready to take it

2 hours to complete

Must hand-in (or e-mail in) by 11:59pm Friday Oct. 18

Can use: class notes, your notes, the book, your assignments and Wikipedia.

You may **not** use: your neighbor, anything else on the web, etc.

What can be covered

Anything we've talked about in class

Anything in the reading (not these are not necessarily the same things)

Anything we've covered in the assignments

Midterm topics

Machine learning basics
- different types of learning problems
- feature-based machine learning
- data assumptions/data generating distribution

Classification problem setup

Proper experimentation
- train/dev/test
- evaluation/accuracy/training error
- optimizing hyperparameters

Midterm topics

Learning algorithms
- Decision trees
- K-NN
- Perceptron
- Gradient descent

Algorithm properties
- training/learning
  - rational/why it works
- classifying
- hyperparameters
- avoiding overfitting
- algorithm variants/improvements

Midterm topics

Geometric view of data
- distances between examples
- decision boundaries

Features
- example features
- removing erroneous features/picking good features
- challenges with high-dimensional data
- feature normalization

Other pre-processing
- outlier detection

Midterm topics

Comparing algorithms
- n-fold cross validation
  - leave one out validation
- bootstrap resampling
- t-test

imbalanced data
- evaluation
  - precision/recall, F1, AUC
- subsampling
- oversampling
- weighted binary classifiers

Midterm topics

Multiclass classification
- Modifying existing approaches
- Using binary classifier
  - OVA
  - AVA
  - Tree-based
- micro- vs. macro-averaging

Ranking
- using binary classifier
- using weighted binary classifier
- evaluation

Midterm topics

Gradient descent
- 0/1 loss
- Surrogate loss functions
- Convexity
- minimization algorithm
- regularization
  - different regularizers
  - p-norms

Misc
- good coding habits
- JavaDoc

Midterm general advice

2 hours goes by fast!
- Don't plan on looking everything up
  - Lookup equations, algorithms, random details
- Make sure you understand the key concepts
- Don't spend too much time on any one question
  - Skip questions you're stuck on and come back to them
- Watch the time as you go

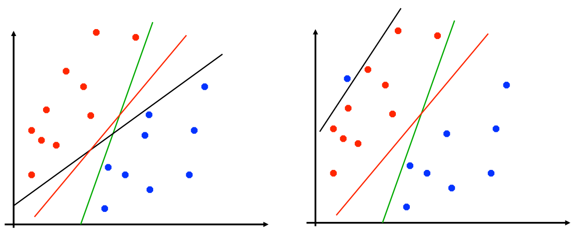
Be careful on the T/F questions

For written questions
- think before you write
- make your argument/analysis clear and concise

Midterm topics

General ML concepts
- avoiding overfitting
- algorithm comparison
- algorithm/model bias
- model-based machine learning
- online vs. offline learning

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:

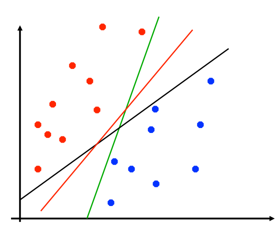
Linear approaches so far

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 is saw recently

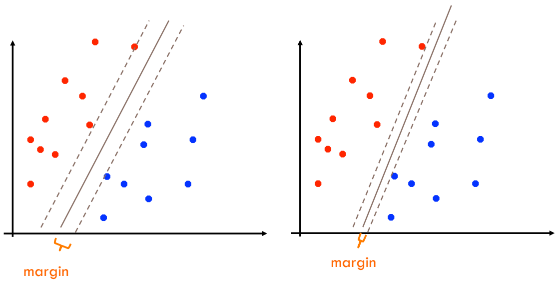
Gradient descent:
- separable and non-separable
  - finds the hyperplane that minimizes the objective function (loss + regularization)

Which hyperplane is this?

Which hyperplane would you choose?

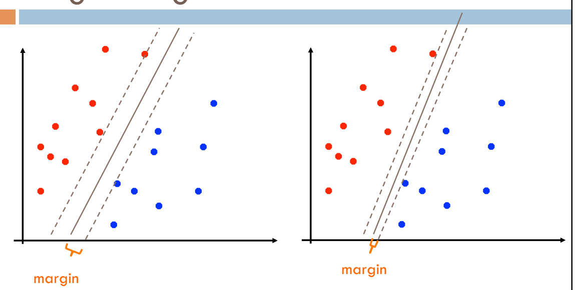


Large margin classifiers



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

Large margin classifier setup

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

Setup as a constrained optimization problem:

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

subject to:

$$y_i(w \cdot x_i + b) > 0 \quad \forall i$$

what does this say?

Large margin classifier setup

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

subject to:

$$y_i(w \cdot x_i + b) > 0 \quad \forall i$$

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

subject to:

$$y_i(w \cdot x_i + b) \geq c \quad \forall i$$
$$c > 0$$

Are these equivalent?

Large margin classifier setup

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

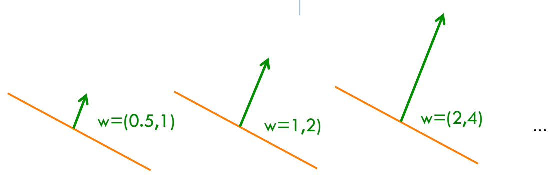
subject to:

$$y_i(w \cdot x_i + b) > 0 \quad \forall i$$

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

subject to:

$$y_i(w \cdot x_i + b) \geq c \quad \forall i$$
$$c > 0$$



Large margin classifier setup

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

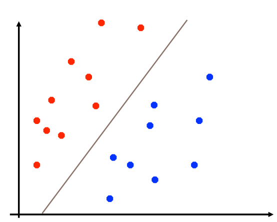
subject to:

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

We'll assume c =1, however, any c > 0 works

Measuring the margin

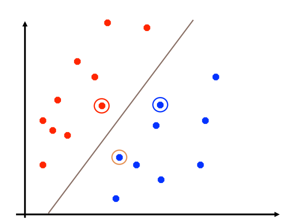
How do we calculate the margin?



Support vectors

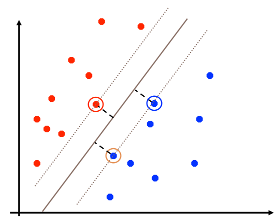
For any separating hyperplane, there exist some set of "closest points"

These are called the support vectors



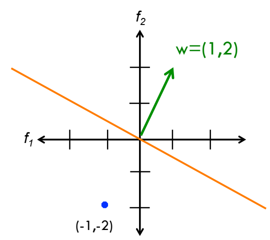
Measuring the margin

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



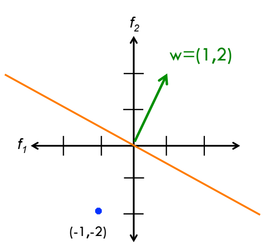
Distance from the hyperplane

How far away is this point from the hyperplane?



Distance from the hyperplane

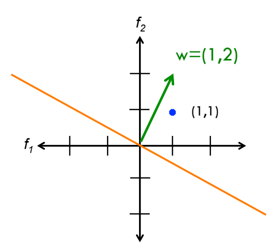
How far away is this point from the hyperplane?



$$d = \sqrt{1^2 + 2^2} = \sqrt{5}$$

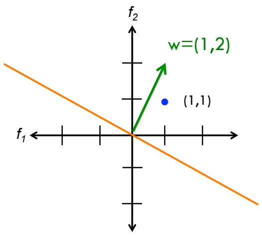
Distance from the hyperplane

How far away is this point from the hyperplane?



Distance from the hyperplane

How far away is this point from the hyperplane?

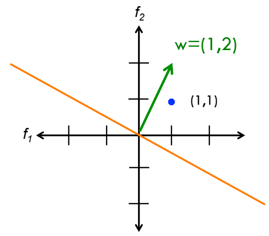


$$d(x) = \frac{w}{\|w\|} \cdot x + b$$

length normalized weight vectors

Distance from the hyperplane

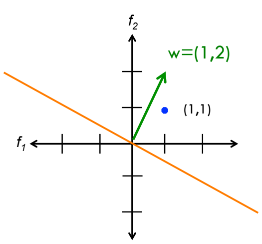
How far away is this point from the hyperplane?



$$\begin{aligned} d(x) &= \frac{w}{\|w\|} \cdot x + b \\ &= \frac{1}{\sqrt{5}}\left(w_1 x_1 + w_2 x_2\right) + b \\ &= \frac{1}{\sqrt{5}}\left(1*1 + 1*2\right) + 0 \\ &= 1.34 \end{aligned}$$

Distance from the hyperplane

Why length normalized?

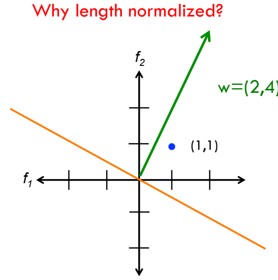


$$d(x) = \frac{w}{\|w\|} \cdot x + b$$

length normalized weight vectors

Distance from the hyperplane

Why length normalized?

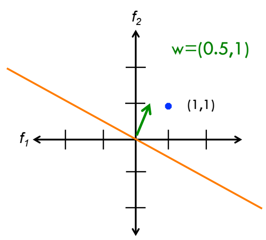


$$d(x) = \frac{w}{\|w\|} \cdot x + b$$

length normalized weight vectors

Distance from the hyperplane

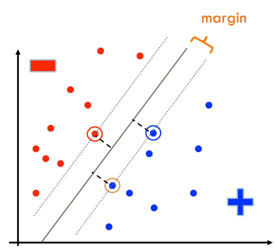
Why length normalized?



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

length normalized
weight vectors

Measuring the margin



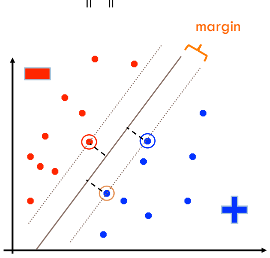
Thought experiment:

Someone gives you the optimal support vectors

Where is the max margin hyperplane?

Measuring the margin

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



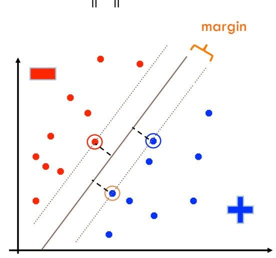
$\boxed{\text{Margin} = (d^+ - d^-)/2}$

Max margin hyperplane is halfway in between the positive support vectors and the negative support vectors

Why?

Measuring the margin

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



$\boxed{\text{Margin} = (d^+ - d^-)/2}$

Max margin hyperplane is halfway in between the positive support vectors and the negative support vectors

- All support vectors are the same distance
- To maximize, hyperplane should be directly in between

Measuring the margin

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

Margin = (d⁺-d⁻)/2

$$\text{margin } = \frac{1}{2}\left(\frac{w}{\|w\|} \cdot x^{+} + b - \left(\frac{w}{\|w\|} \cdot x^{-} + b\right)\right)$$

What is $wx+b$ for support vectors?

Hint:

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

subject to:

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

Measuring the margin

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

subject to:

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

The support vectors have $y_i(w \cdot x_i + b) = 1$

Otherwise, we could make the margin larger!

Measuring the margin

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

Margin = (d⁺-d⁻)/2

$$\text{margin } = \frac{1}{2}\left(\frac{w}{\|w\|} \cdot x^{+} + b - \left(\frac{w}{\|w\|} \cdot x^{-} + b\right)\right)$$

$$= \frac{1}{2}\left(\frac{1}{\|w\|} - \frac{-1}{\|w\|}\right)$$ negative example

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

Maximizing the margin

$$\max_{w,b} \quad \frac{1}{\|w\|}$$

subject to:

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

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

Maximizing the margin

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

subject to:

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

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

Maximizing the margin

The minimization criterion wants w to be as small as possible

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

subject to:

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

The constraints:

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

Maximizing the margin: the real problem

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

subject to:

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

Why the squared?

Maximizing the margin: the real problem

$$\min_{w,b} \quad \|w\| = \sqrt{\sum_i w_i^2}$$

subject to:

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

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

subject to:

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

Minimizing ||w|| is equivalent to minimizing $||w||^2$

The sum of the squared weights is a convex function!

Support vector machine problem

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

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)