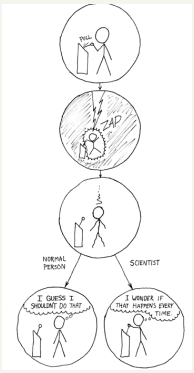


http://xkcd.com/242/

Text Classification 2

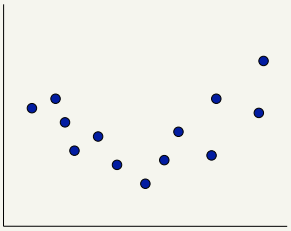
David Kauchak
cs459
Fall 2012

adapted from:
http://www.stanford.edu/class/cs276/handouts/lecture10-textcat-naivebayes.ppt
http://www.stanford.edu/class/cs276/handouts/lecture11-vector-classify.ppt
http://www.stanford.edu/class/cs276/handouts/lecture12-SVMs.ppt

Administrative

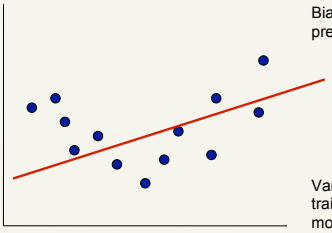
- Project status update
  - Due 11/27 (A week from today by midnight)
  - Take this seriously
  - I want to see some progress
- Quiz
  - Mean: 20.4
  - Median: 19.5
  - Will curve the scores up some (one example, add 10 divide by 35)
- Assignment 4 back soon…

Bias/variance trade-off



We want to fit a polynomial to this, which one should we use?

Bias/variance trade-off



Bias: How well does the model predict the training data?

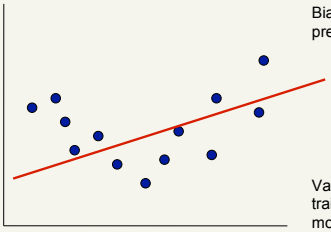
- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

High variance OR high bias?

Bias/variance trade-off



Bias: How well does the model predict the training data?

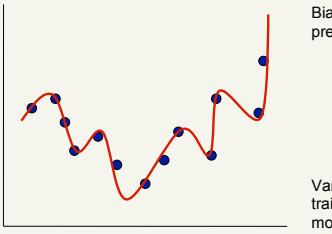
- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

High bias

Bias/variance trade-off



Bias: How well does the model predict the training data?

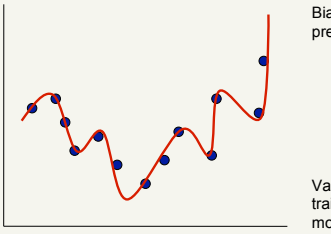
- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

High variance OR high bias?

Bias/variance trade-off



Bias: How well does the model predict the training data?

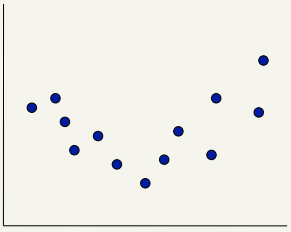
- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

High variance

Bias/variance trade-off



Bias: How well does the model predict the training data?

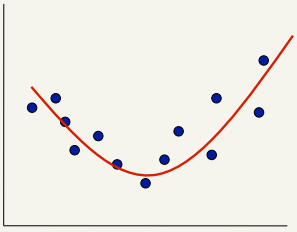
- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

What do we want?

Bias/variance trade-off



Bias: How well does the model predict the training data?

- high bias – the model doesn't do a good job of predicting the training data (high training set error)
- The model predictions are biased by the model

Variance: How sensitive to the training data is the learned model?

- high variance – changing the training data can drastically change the learned model

Compromise between bias and variance

k-NN vs. Naive Bayes

How do k-NN and NB sit on the variance/ bias spectrum?

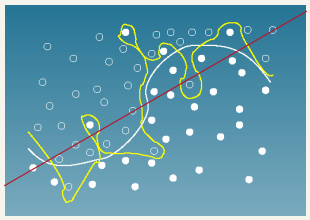
k-NN has high variance and low bias.

- more complicated model
- can model any boundary
- but very dependent on the training data

NB has low variance and high bias.

- Decision surface has to be linear
- Cannot model all data
- but, less variation based on the training data

Bias vs. variance: Choosing the correct model capacity

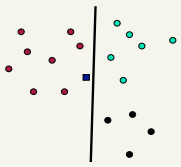


Which separating line should we use?

Separation by Hyperplanes

A strong high-bias assumption is *linear separability*:

- in 2 dimensions, can separate classes by a line
- in higher dimensions, need hyperplanes



Lots of linear classifiers

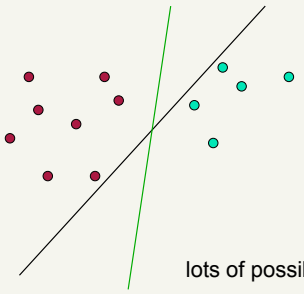
Many common text classifiers are linear classifiers

- Naïve Bayes
- Perceptron
- Rocchio
- Logistic regression
- Support vector machines (with linear kernel)
- Linear regression

Despite this similarity, noticeable performance difference

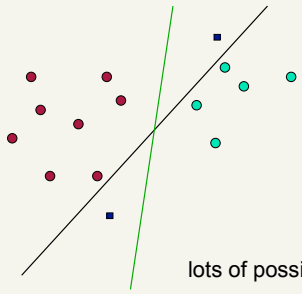
How might algorithms differ?

Which Hyperplane?



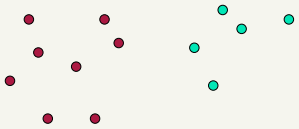
lots of possible solutions

Which Hyperplane?

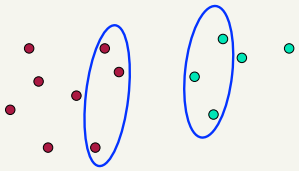


lots of possible solutions

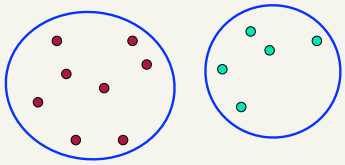
Which examples are important?



Which examples are important?

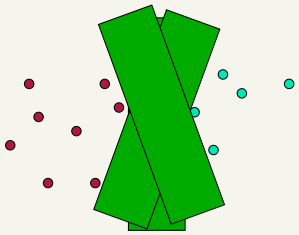


Which examples are important?



Another intuition

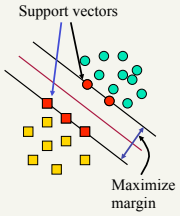
- If you have to place a fat separator between classes, you have less choices, and so the capacity of the model has been decreased



20

Support Vector Machine (SVM)

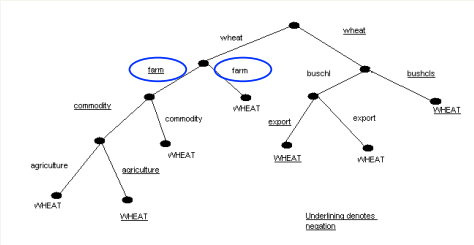
- SVMs maximize the *margin* around the separating hyperplane.
  - A.k.a. large margin classifiers
- The decision function is fully specified by a subset of training samples, *the support vectors*.
- Solving SVMs is a *quadratic programming* problem
- Seen by many as the most successful current text classification method*



*but other discriminative methods often perform very similarly

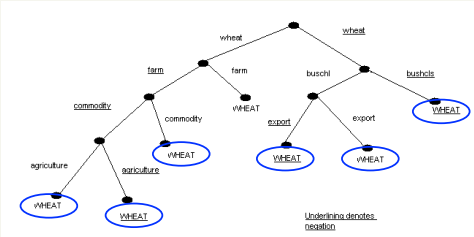
Decision trees

- Tree with internal nodes labeled by terms/features
- Branches are labeled by tests on the weight that the term has
  - farm vs. not farm
  - x > 100



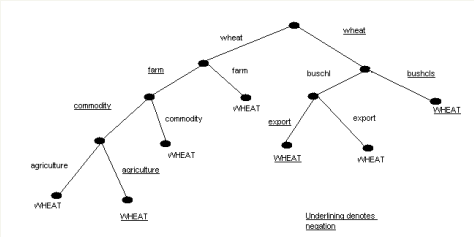
Decision trees

- Roots are labeled with the class



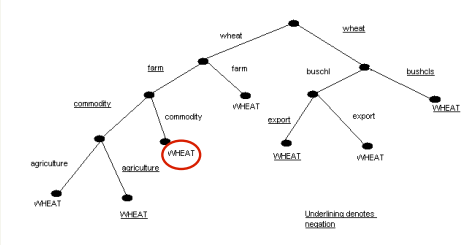
Decision trees

- Classifier categorizes a document by descending tree following tests to leaf
- The label of the leaf node is then assigned to the document



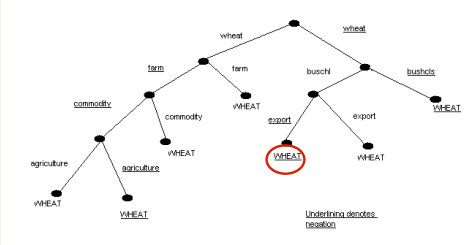
Decision trees

wheat, not(farm), commodity, not(agriculture)?



Decision trees

not(wheat), not(farm), commodity, export, buschi?



Decision trees

- Most decision trees are binary trees
- DT make good use of a few high-leverage features
- Linear or non-linear classifier?