CSCI 5582
Artificial Intelligence

Lecture 20

Jim Martin

Today 11/9

- Review
- Boosting
- Support Vector Machines
- Break
- Neural Networks

Review

- Supervised machine learning
  - Naïve Bayes
  - Decision trees
  - Decision lists
  - Ensembles

Ensembles

- Ensemble classifiers are groups of classifiers that combine to vote on a classification.
  - Generally, the individual classifiers are weak classifiers. Classifiers that have a high error rate (barely better than chance).
- Boosting is a relatively simple method of creating effective ensembles.

Ensembles

- The key to ensemble learning is to induce different classifiers from the original training set
  - If the classifiers are all the same, there's no point to voting
- Boosting accomplishes this by altering the training set used to create each of the classifiers
- Bagging does it by re-sampling the training data.

Why?

- Why the heck should voting work? Here's a rough intuition.
- Let's assume you have three classifiers each with an error rate of .2 (Correct 80% of the time).
- Let's assume you do a majority vote of them to get your answer.
- What should your error rate be on the combined classifier?

Why?

- Majority vote yields 4 situations (out of 8 total) where the wrong answer is produced.
  - All three guess wrong
    - .2*.2*.2 = .008
  - 2 of the three guess wrong
    - .2*.2*.8 = .032
    - But there are three ways for one to be right and 2 wrong. 3* .032 = .096.
  - So… .008+.096 = .104
  - You cut your error rate in half.

Boosting

1. Start with a training set where all instances have an equal weight
2. Train a weak classifier with that set
3. Generate an error rate for that classifier; and a goodness metric for that classifier.
4. Add this classifier to the ensemble

Boosting

5. Use the goodness measure to
   1. Reduce the importance of the instances that the classifier got right.
   2. Increase the importance of the instances that the classifier got wrong.
6. Go to 2 (Generate another classifier using the new distribution of instances)

Boosting: Major Points

- What's a weak learner?
- What's the distribution mean?
- Error on the training set
- Altering the distribution
- The number of trials
- Final voting scheme

The Weighted Distribution

- Basic idea:
  - If you start with m instances, give each a weight of 1/m.
  - With each re-weighting, reduce the weight of the examples gotten right and increase the weight of the ones gotten wrong. (Normalizing so it still sums to 1).

Using the Weights

- The learning scheme must be sensitive to the weights on the instances.
- Some schemes can handle it directly, others like DT and DL learning have to get at it indirectly.
  - Any ideas?

Weighted Voting

- Each classifier in the ensemble gets a voted proportional to its error rate
- Majority rules

Major Theoretical Results

- The training error can be reduced to any desired amount given enough rounds of boosting
- The testing error can be reduced to any desired amount given enough rounds and enough data.

Major Practical Results

- Extremely dumb learning algorithms can be boosted above the performance of smarter algorithms.
- Even smart algorithms can be boosted to higher levels even though they weren't weak to begin with
- Boosting is trivial to implement

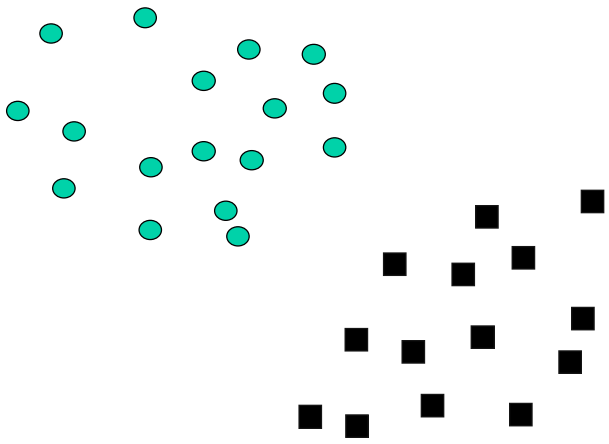
Practicality

- You don't need access to the insides of a learning algorithm/system to improve it via boosting.

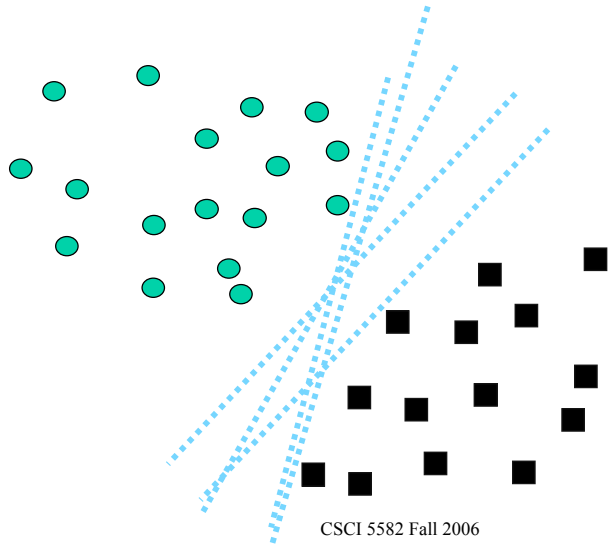
Support Vector Machines

- Two key ideas
  - The notion of the margin
    - Support vectors
  - Mapping to higher dimensional spaces
    - Kernel functions
- Don't sweat the details of the math

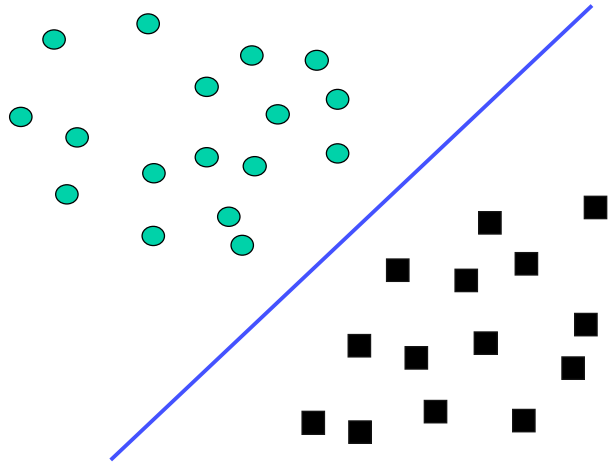
Best Linear Separator?



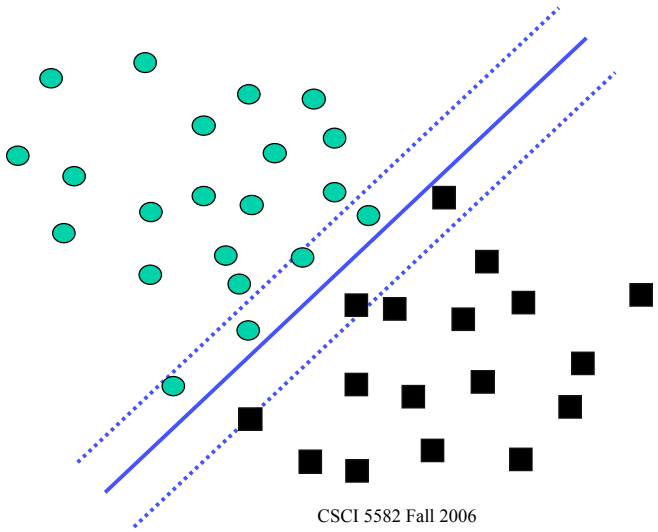
Best Linear Separator?



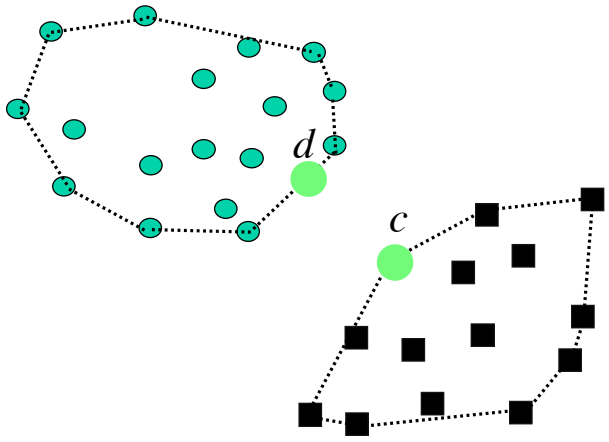
Best Linear Separator?



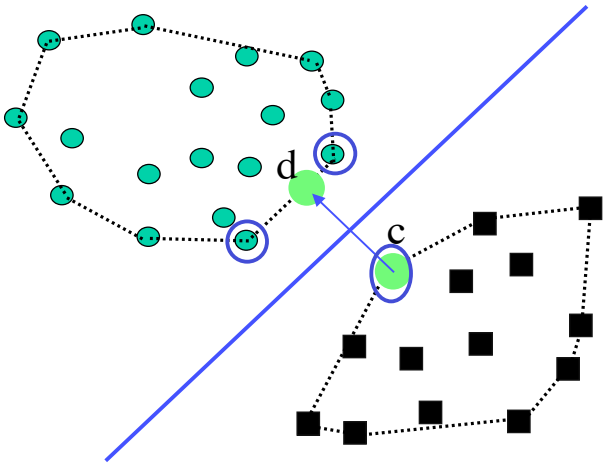
Why is this good?



Find Closest Points in Convex Hulls

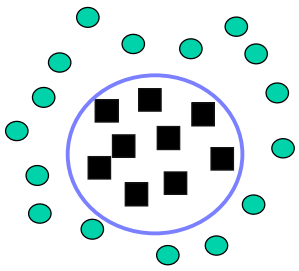


Plane Bisect Support Vectors



Higher Dimensions

- That assumes that there is a linear classifier that can separate the data.



One Solution

- Well, we could just search in the space of non-linear functions that will separate the data
- Two problems
  - Likely to overfit the data
  - The space is too large

Kernel Trick

- Map the objects to a higher dimensional space.
- Book example
  - Map an object in two dimensions (x1 and x2) into a three dimensional space
    - F1 = x1^2, F2 = x2^2, and F3 = Sqrt(2*x1*x2)
- Points not linearly separable in the original space will be separable in the new space.

But

- In the higher dimensional space, there are gazillion hyperplanes that will separate the data cleanly.
  - How to choose among them?
    - Use the support vector idea

Break

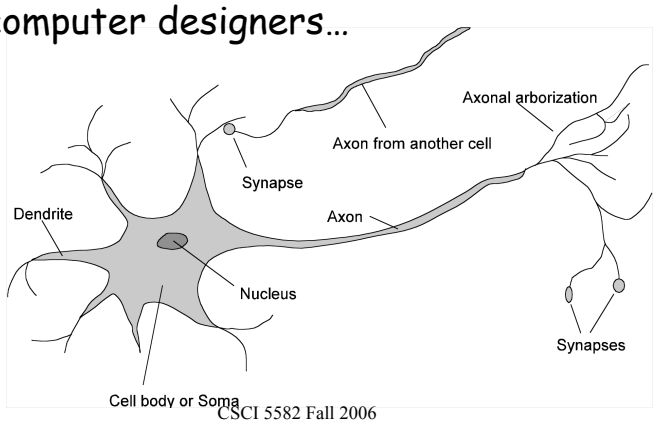
- The next quiz will be on 11/28.
- It will cover the ML material and the probabilistic sequence material.
- The readings for this quiz are:
  - Chapter 18
  - Chapter 19
  - Chapter 20: 712-718
  - HMM chapter posted on the web

Break

- The ML code I mentioned last time is called Weka. It's all java.
- There's also a package called PyML. It's a python package for using SVMs.

Neural Networks

- The brain has always been an inspiration to computer designers…



Some Relevant Factoids

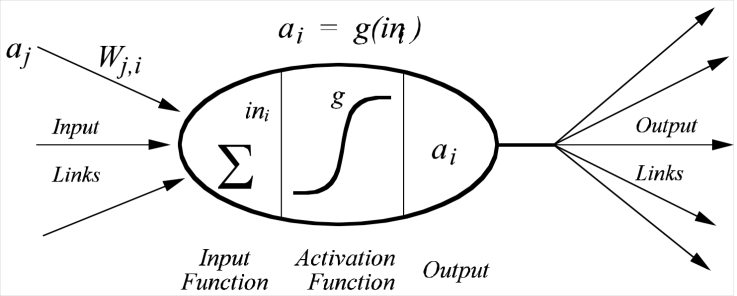
- Basic units are slow
- Basic units are highly interconnected
- Processing (largely) occurs in parallel
- There is little central control
- Complex tasks (face recognition, language processing) can occur in under a second

Motivation

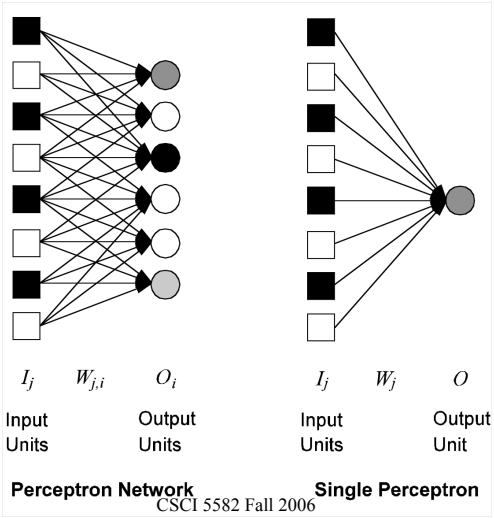
- Roughly neurons can fire once every 5 milliseconds (200 times a second)
- You can recognize someone's face in about 1/2 second.
- How many steps does your face recognition algorithm have?

Artificial NNs

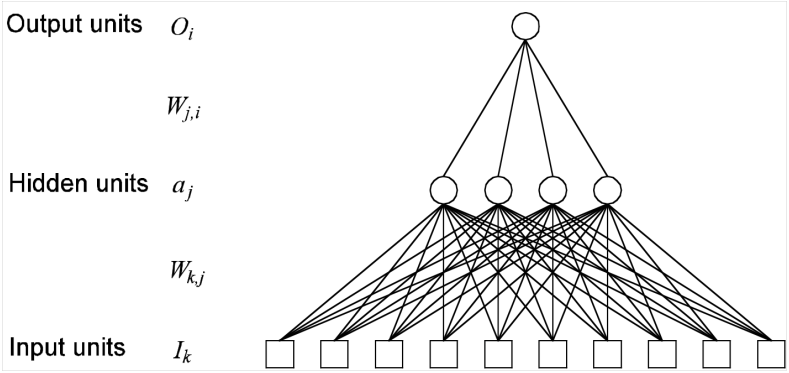
- Abstract away from almost everything except connectivity…



Single Layer Networks



Multi-Layer Networks

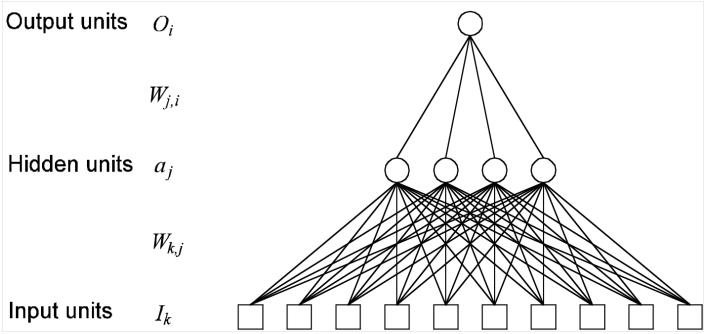


Feed Forward Processing

- Input is presented to the input layer
- Activations are computed for the hidden layers
- Activations are then computed for the output layer

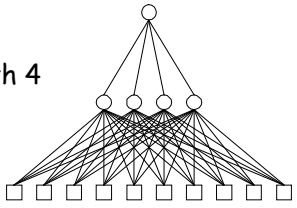
Reality Check

- What is this network really?



Reality Check

- What is this network really?
  - The output unit is just a floating point number
  - The input units are just a vector of floating point numbers of length 10.
  - The hidden units are just a vector of length 4
  - The weights
    - The first layer is 4x10 matrix
    - The second layer is a vector of length 4



Knowledge Representation

- Typical representations are essentially feature-vectors of the type we've been discussing.
  - Food: Thai, French, Italian, etc
  - Patrons: Some, None, Full
  - Hungry: Yes/No
  - Wait estimate: ?

Knowledge Representation

- A unit for each feature/value combination
- An encoding of that based on the number of values per feature
- A micro-encoding that captures the similarity structure
  - I.e. use sub-features to represent features.

Knowledge Representation

- But using clever encodings you can pretty much encode anything you want
  - Recursive tree structures
  - Logical statements
  - Role-filler representations

Knowledge Representation

- But just saying you can encode something is not the same thing as saying that you can learn it.

Learning in Neural Networks

- Learning in NNs involves induction (supervised training) using labeled examples.
- This typically involves setting the weights so that the network does the right thing.
- Doing the right thing? Reflecting the training set.

NN Learning

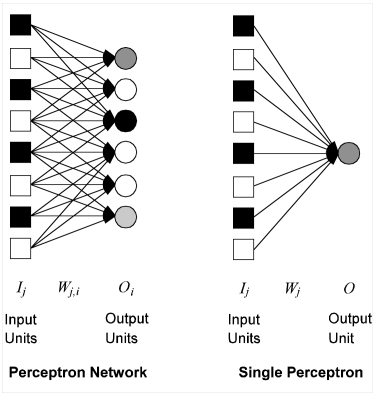
- Initialize the weights (randomly).
- Examine a single example
  - Compute the difference between the right answer and the answer given
- Use the difference to adjust the weights in the right direction
- Get another training instance

Generic Learning

```
function NEURAL-NETWORK-LEARNING(examples) returns network

  network ⇐ a network with randomly assigned weights
  repeat
     for each e in examples do
        O ⇐ NEURAL-NETWORK-OUTPUT(network, e)
        T ⇐ the observed output values from e
        update the weights in network based on e, O, and T
     end
  until all examples correctly predicted or stopping criterion is reached
  return network
```


Single-Layer Learning



$$W_j = W_j + \alpha * I_j * Err$$

Next Time

- Learning with knowledge
  - You go to Brazil. You hear someone speaking Portuguese.
    - What do the rest of the people in Brazil speak?
  - You go to New York. You hear someone speaking Polish.
    - What do the rest of the people in New York speak?

Next Time

- Learning with Knowledge
  - Chapter 19