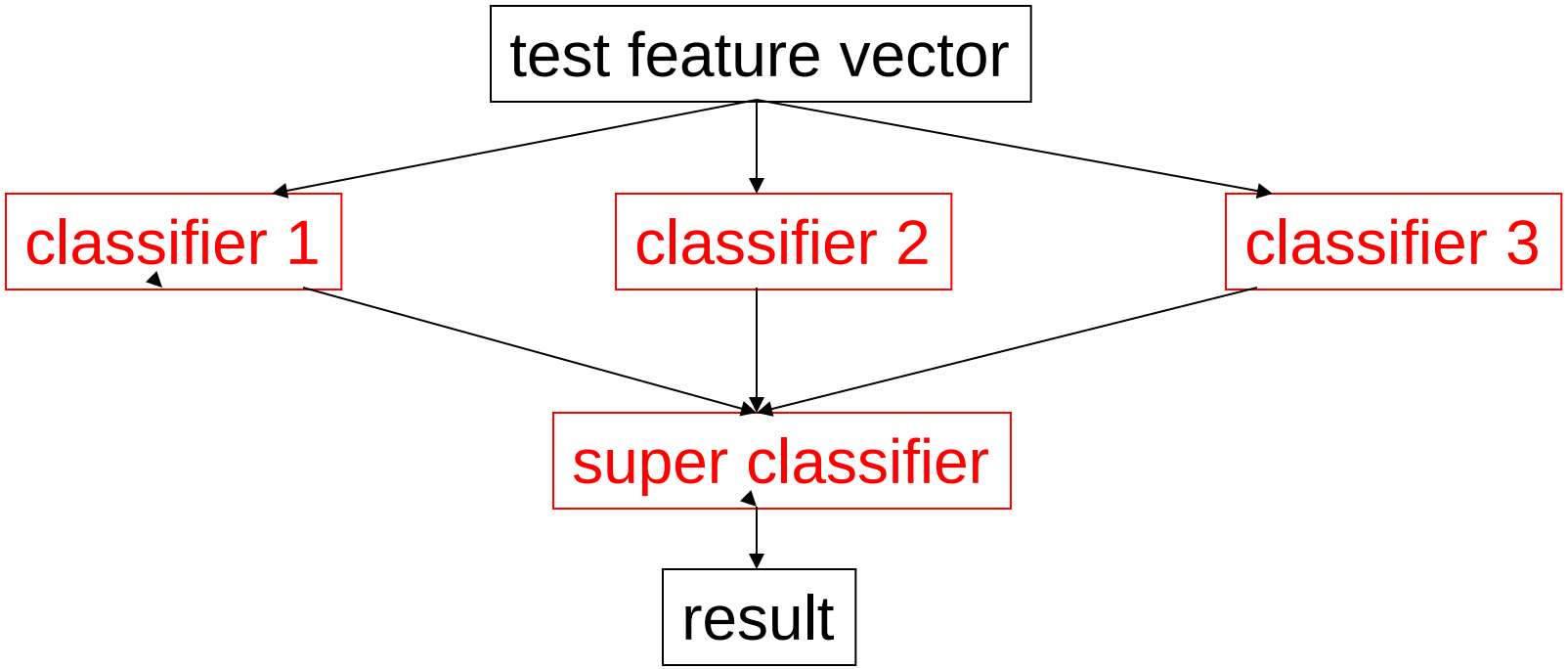
Ensembles

- An ensemble is a set of classifiers whose combined results give the final decision.

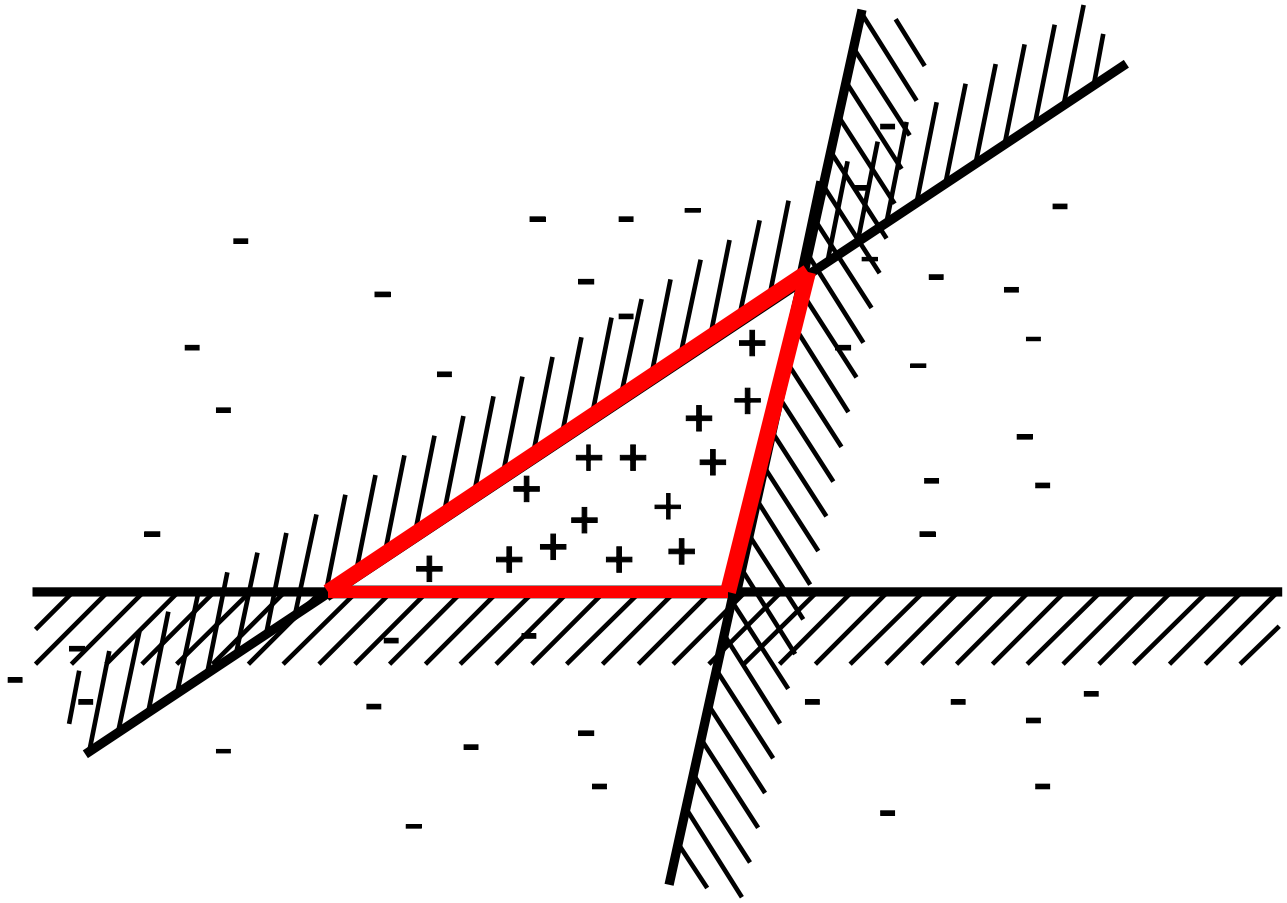


Model* Ensembles

- **Basic idea:**
  Instead of learning one model,
  Learn several and combine them
- Typically improves accuracy, often by a lot
- **Many methods:**
  - Bagging
  - Boosting
  - ECOC (error-correcting output coding)
  - Stacking
  - Etc.

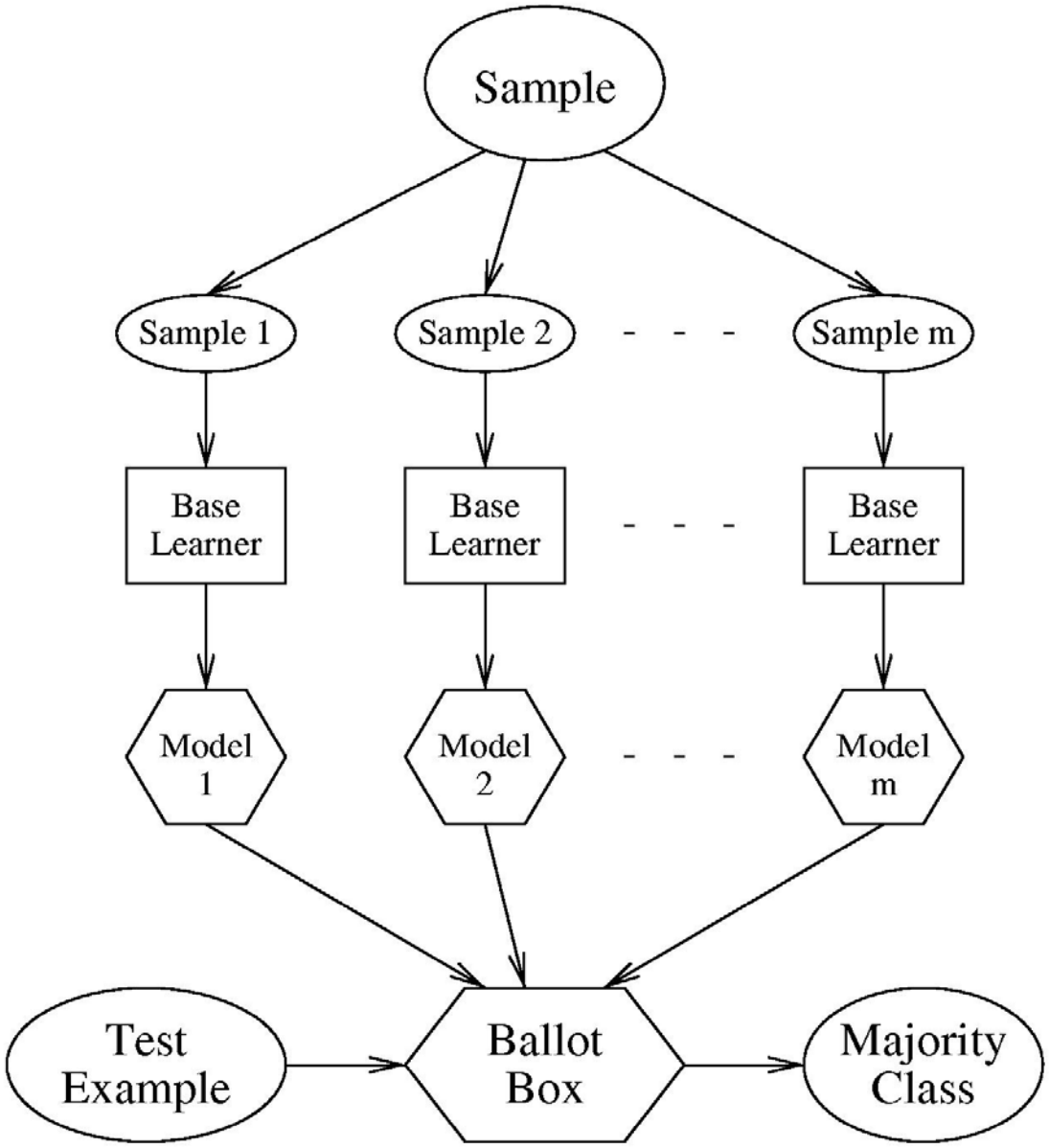
*A model is the learned decision rule. It can be as simple as a hyperplane in n-space (ie. a line in 2D or plane in 3D) or in the form of a decision tree or other modern classifier.

Majority Vote for Several Linear Models



Bagging

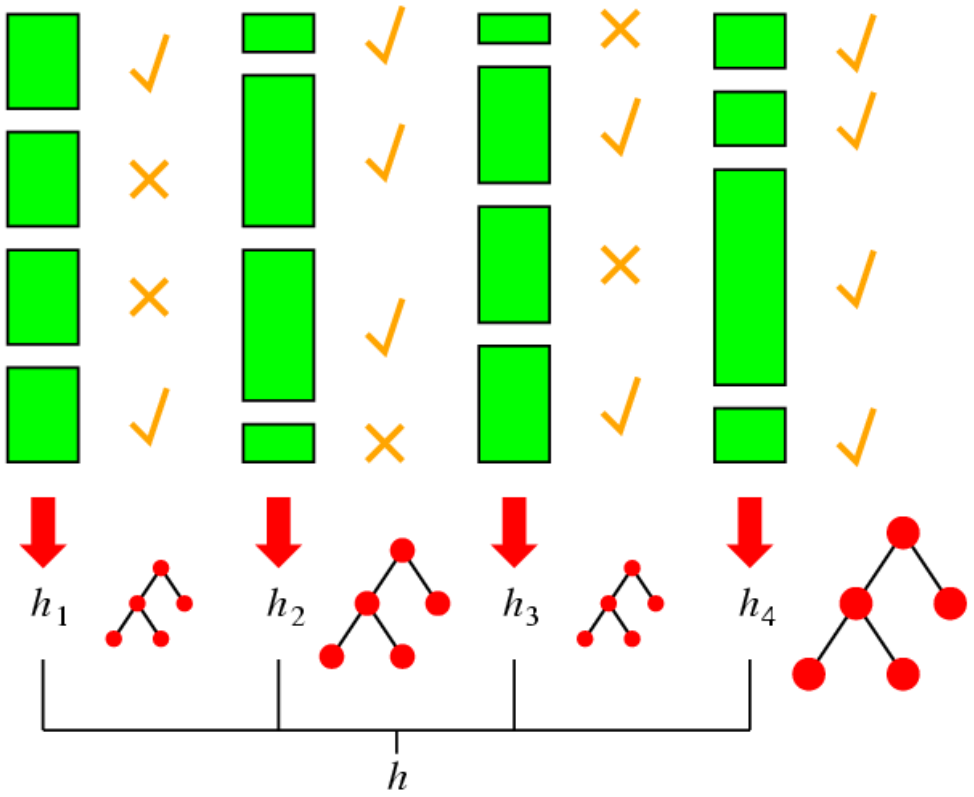
- Generate "bootstrap" replicates of training set by sampling with replacement
- Learn one model on each replicate
- Combine by uniform voting


Sample
Sample 1
Sample 2
Sample m
Base Learner
Base Learner
Base Learner
Model 1
Model 2
Model m
Test Example
Ballot Box
Majority Class

Boosting

- Maintain vector of weights for examples
- Initialize with uniform weights
- Loop:
  - Apply learner to weighted examples (or sample)
  - Increase weights of misclassified examples
- Combine models by weighted voting

Idea of Boosting



Boosting In More Detail
(Pedro Domingos' Algorithm)

1. Set all E weights to 1, and learn H1.
2. Repeat m times: increase the weights of misclassified Es, and learn H2,…Hm.
3. H1..Hm have "weighted majority" vote when classifying each test
   Weight(H)=accuracy of H on the training data

ADABoost

- ADABoost boosts the accuracy of the original learning algorithm.

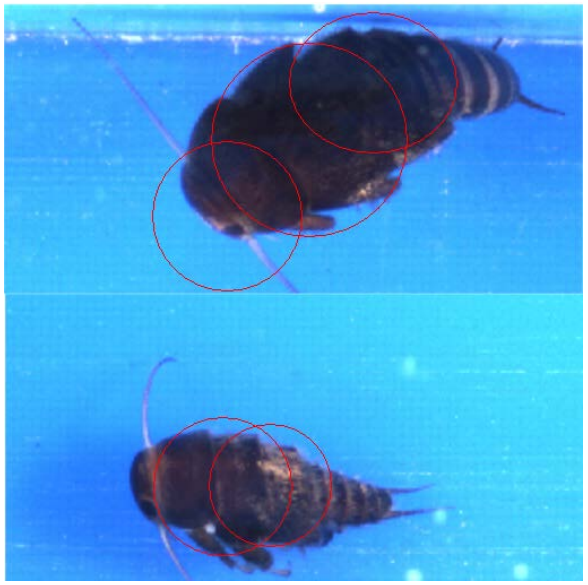
- If the original learning algorithm does slightly better than 50% accuracy, ADABoost with a large enough number of classifiers is guaranteed to classify the training data perfectly.

ADABoost Weight Updating

```
for j = 1 to N do /* go through training samples */
   if h[m](xj) <> yj then error <- error + wj

for j = 1 to N do
   if h[m](xj) = yj then w[j] <- w[j] * error/(1-error)
```

Sample Application: Insect Recognition



Using circular regions of interest selected by an interest operator, train a classifier to recognize the different classes of insects.

Boosting Comparison

- **ADTree classifier only** **(alternating decision tree)**

- Correctly Classified Instances 268 70.1571 %
- Incorrectly Classified Instances 114 29.8429 %
- Mean absolute error 0.3855
- Relative absolute error 77.2229 %

Classified as ->	Hesperperla	Doroneuria
Real Hesperperlas	167	28
Real Doroneuria	51	136

Boosting Comparison

AdaboostM1 with ADTree classifier

- Correctly Classified Instances 303 79.3194 %
- Incorrectly Classified Instances 79 20.6806 %
- Mean absolute error 0.2277
- Relative absolute error 45.6144 %

Classified as ->	Hesperperla	Doroneuria
Real Hesperperlas	167	28
Real Doroneuria	51	136

Boosting Comparison

- **RepTree classifier only** **(reduced error pruning)**

- Correctly Classified Instances 294 75.3846 %
- Incorrectly Classified Instances 96 24.6154 %
- Mean absolute error 0.3012
- Relative absolute error 60.606 %

Classified as ->	Hesperperla	Doroneuria
Real Hesperperlas	169	41
Real Doroneuria	55	125

Boosting Comparison

AdaboostM1 with RepTree classifier

- Correctly Classified Instances 324 83.0769 %
- Incorrectly Classified Instances 66 16.9231 %
- Mean absolute error 0.1978
- Relative absolute error 39.7848 %

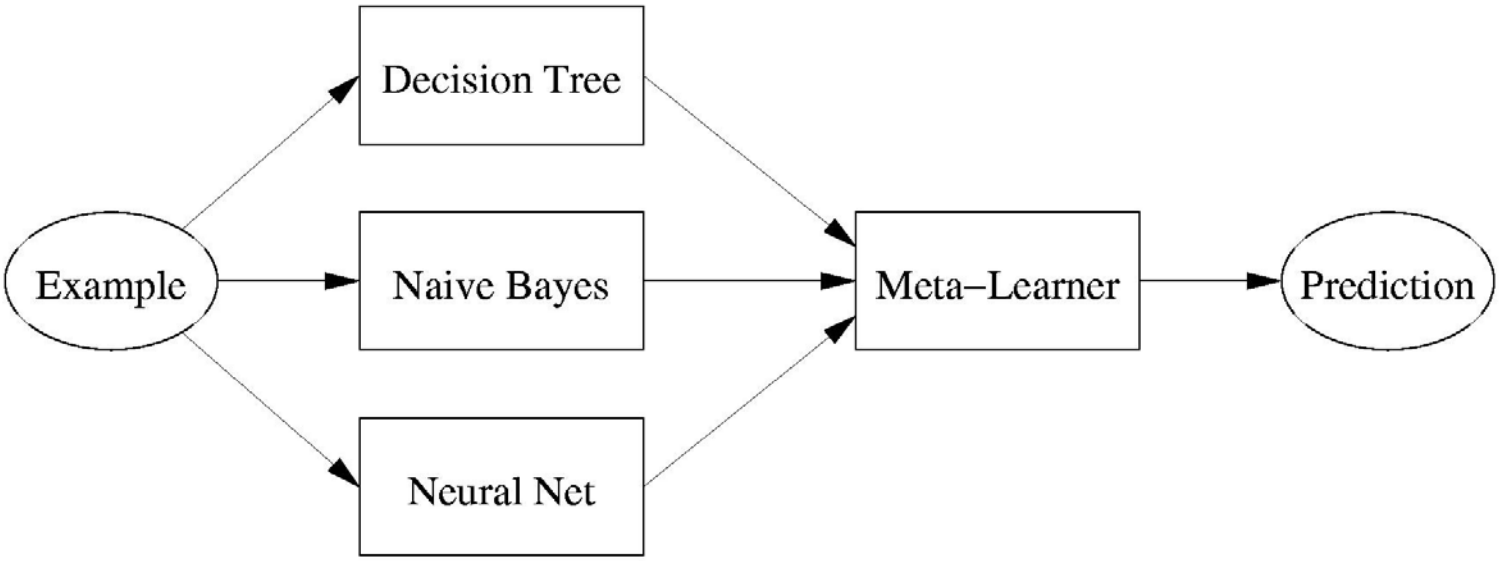
Classified as ->	Hesperperla	Doroneuria
Real Hesperperlas	180	30
Real Doroneuria	36	144

References

- **AdaboostM1**: Yoav Freund and Robert E. Schapire (1996). "Experiments with a new boosting algorithm". Proc International Conference on Machine Learning, pages 148-156, Morgan Kaufmann, San Francisco.
- **ADTree**: Freund, Y., Mason, L.: "The alternating decision tree learning algorithm". Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia, (1999) 124-133.

Stacking

- Apply multiple base learners
  (e.g.: decision trees, naive Bayes, neural nets)
- Meta-learner: Inputs = Base learner predictions
- Training by leave-one-out cross-validation:
  Meta-L. inputs = Predictions on left-out examples

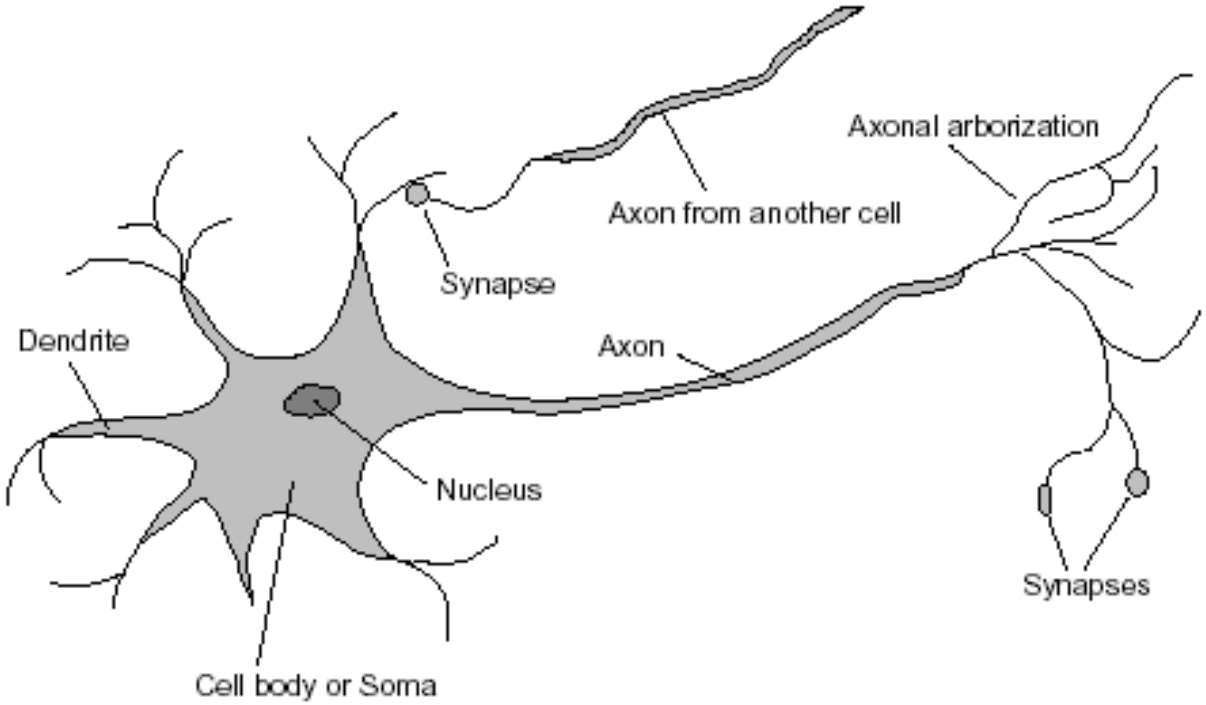


Neural Net Learning

- Motivated by studies of the brain.
- A network of “artificial neurons” that learns a function.
- Doesn’t have clear decision rules like decision trees, but highly successful in many different applications. (e.g. face detection)
- Our hierarchical classifier used neural net classifiers as its components.

Brains

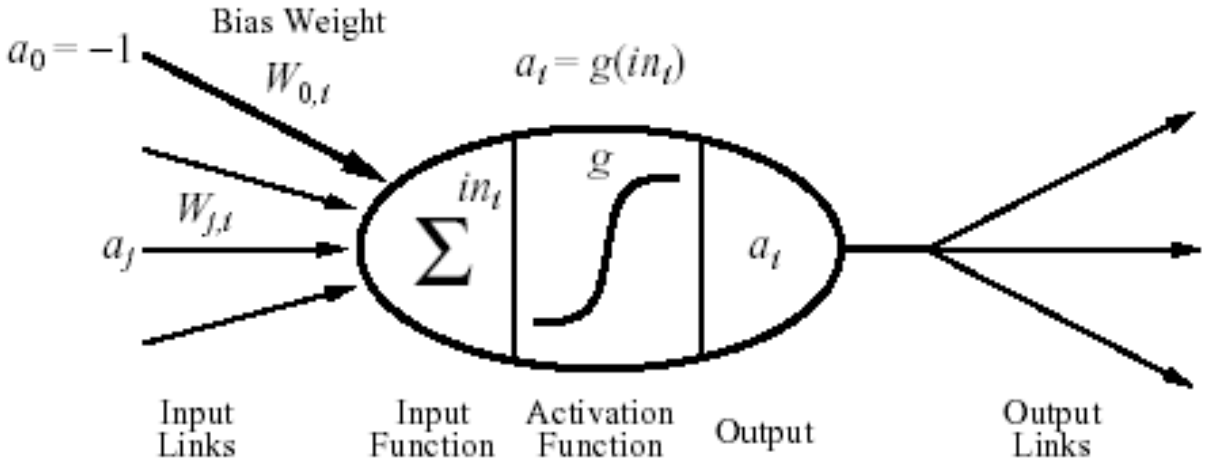
10^{11} neurons of > 20 types, 10^{14} synapses, 1ms–10ms cycle time
Signals are noisy "spike trains" of electrical potential



McCulloch–Pitts "unit"

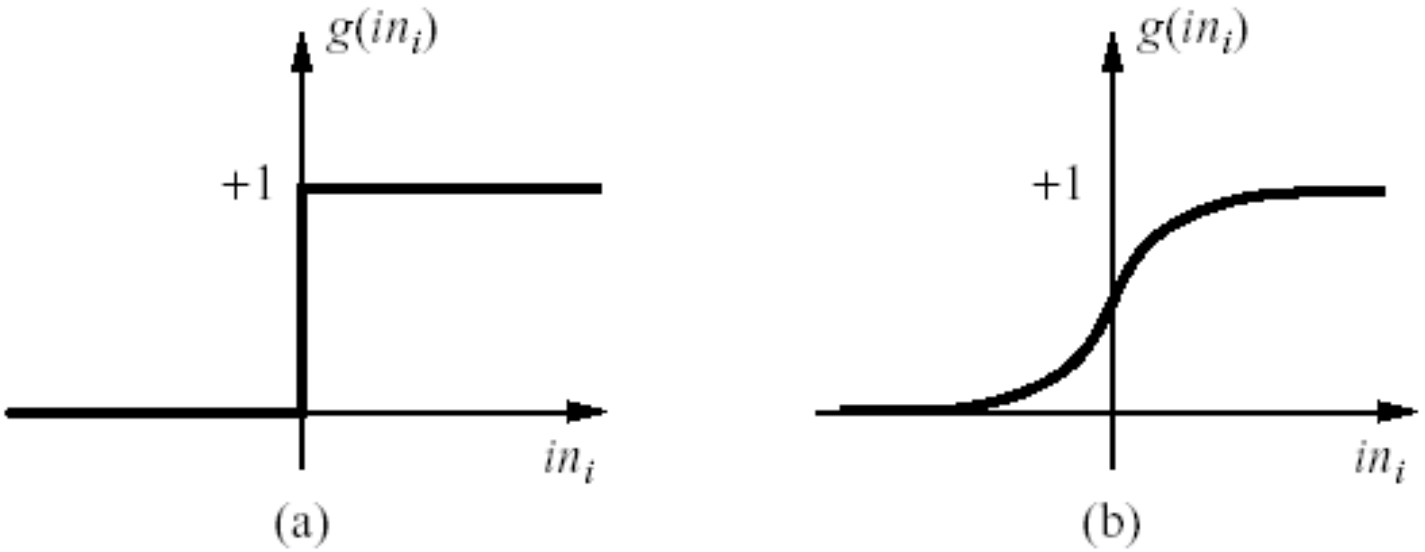
Output is a "squashed" linear function of the inputs:

$$a_i \leftarrow g(in_i) = g\left(\sum_j W_{j,i} a_j\right)$$



A gross oversimplification of real neurons, but its purpose is to develop understanding of what networks of simple units can do

Activation functions

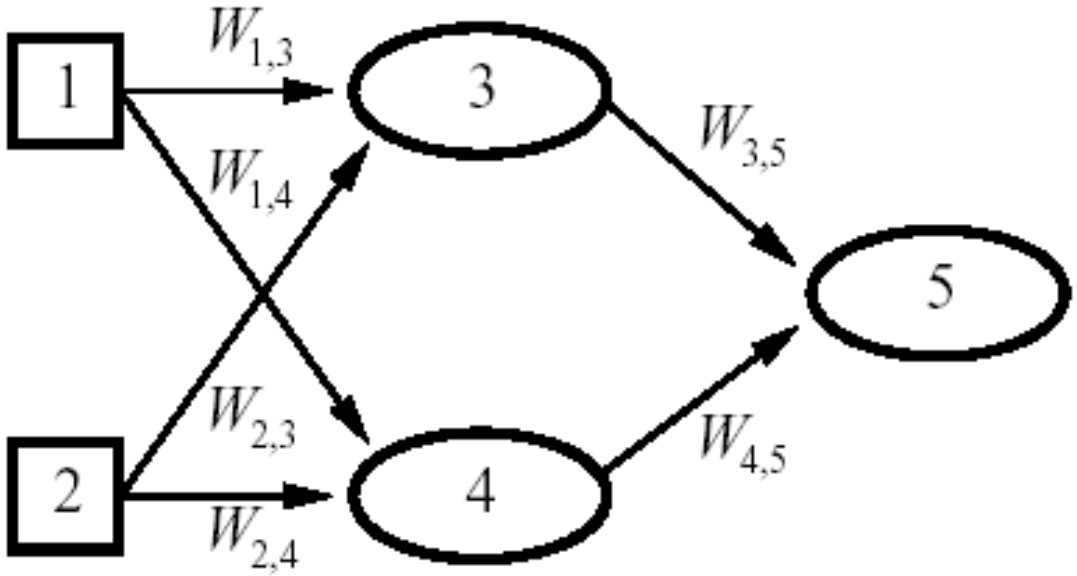


(a) is a step function or threshold function

(b) is a sigmoid function $1/(1+e^{-x})$

Changing the bias weight $W_{0,i}$ moves the threshold location

Feed-forward example



Feed-forward network = a parameterized family of nonlinear functions:

$$\begin{aligned} a_5 &= g(W_{3,5} \cdot a_3 + W_{4,5} \cdot a_4) \\ &= g(W_{3,5} \cdot g(W_{1,3} \cdot a_1 + W_{2,3} \cdot a_2) + W_{4,5} \cdot g(W_{1,4} \cdot a_1 + W_{2,4} \cdot a_2)) \end{aligned}$$

Adjusting weights changes the function: do learning this way!

Perceptron learning

Learn by adjusting weights to reduce error on training set

The squared error for an example with input $\mathbf{x}$ and true output y is

$$E = \frac{1}{2} Err^2 \equiv \frac{1}{2}(y - h_{\mathbf{W}}(\mathbf{x}))^2 \ ,$$

Perform optimization search by gradient descent:

$$\begin{aligned} \frac{\partial E}{\partial W_j} &= Err \times \frac{\partial Err}{\partial W_j} = Err \times \frac{\partial}{\partial W_j}\left(y - g(\Sigma_{j=0}^{n} W_j x_j)\right) \\ &= -Err \times g'(in) \times x_j \end{aligned}$$

Simple weight update rule:

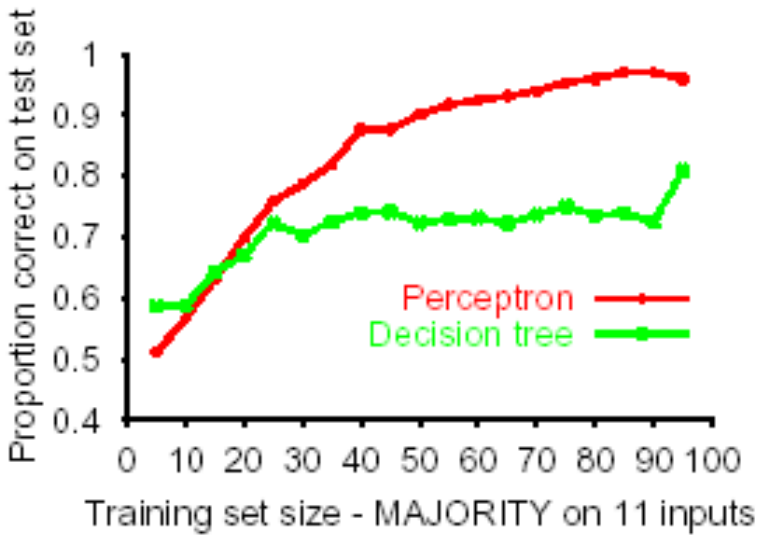
$$W_j \leftarrow W_j + \alpha \times Err \times g'(in) \times x_j$$

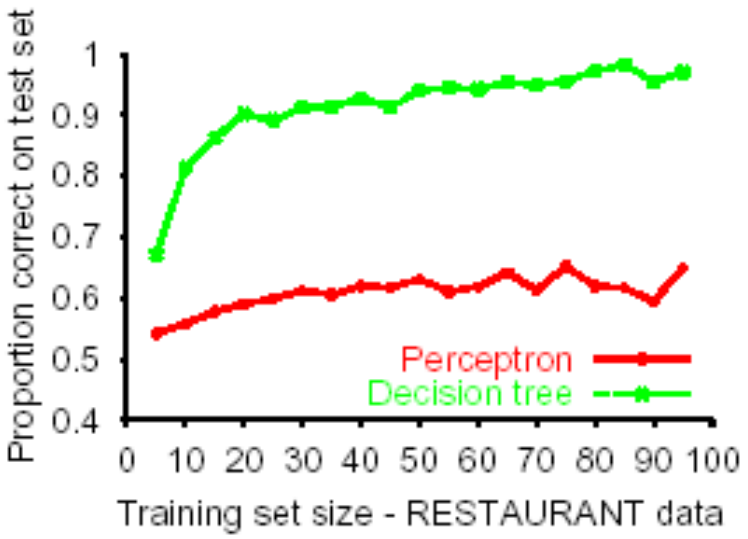
E.g., +ve error $\Rightarrow$ increase network output

$\Rightarrow$ increase weights on +ve inputs, decrease on -ve inputs

Perceptron learning contd.

Perceptron learning rule converges to a consistent function
for any linearly separable data set



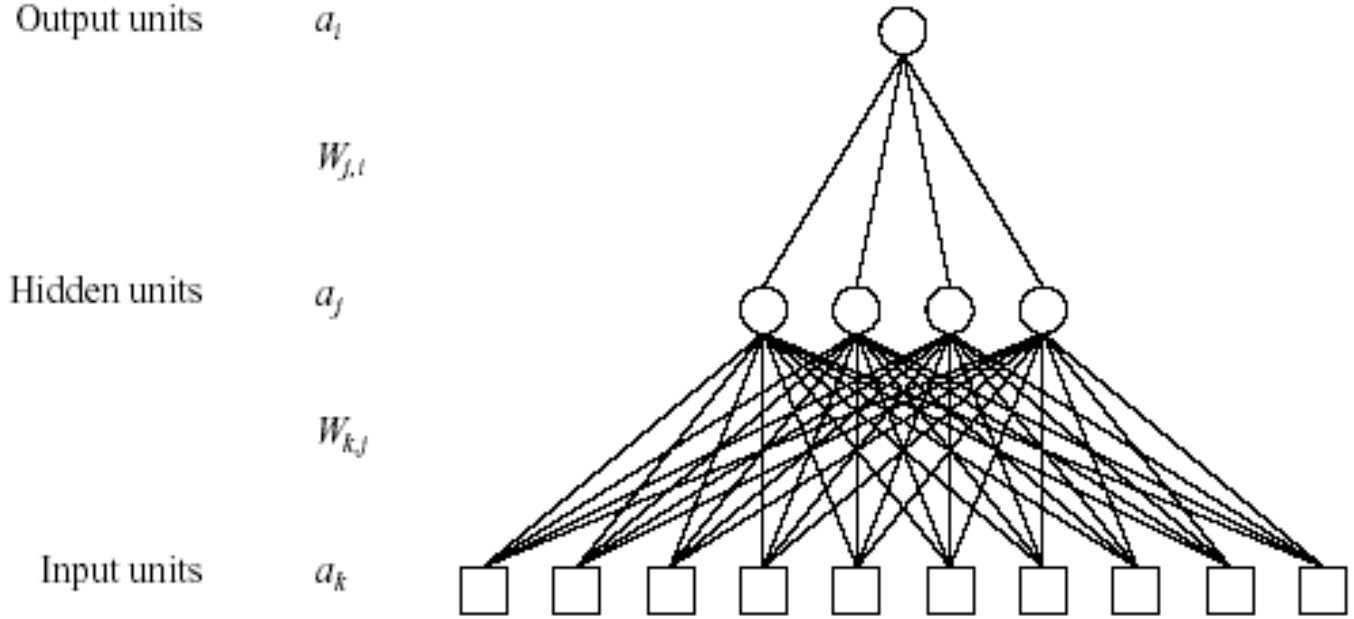


Perceptron learns majority function easily, DTL is hopeless

DTL learns restaurant function easily, perceptron cannot represent it

Multilayer perceptrons

Layers are usually fully connected;
numbers of hidden units typically chosen by hand



Back-propagation learning

Output layer: same as for single-layer perceptron,

$$W_{j,i} \leftarrow W_{j,i} + \alpha \times a_j \times \Delta_i$$

where $\Delta_i = Err_i \times g'(in_i)$

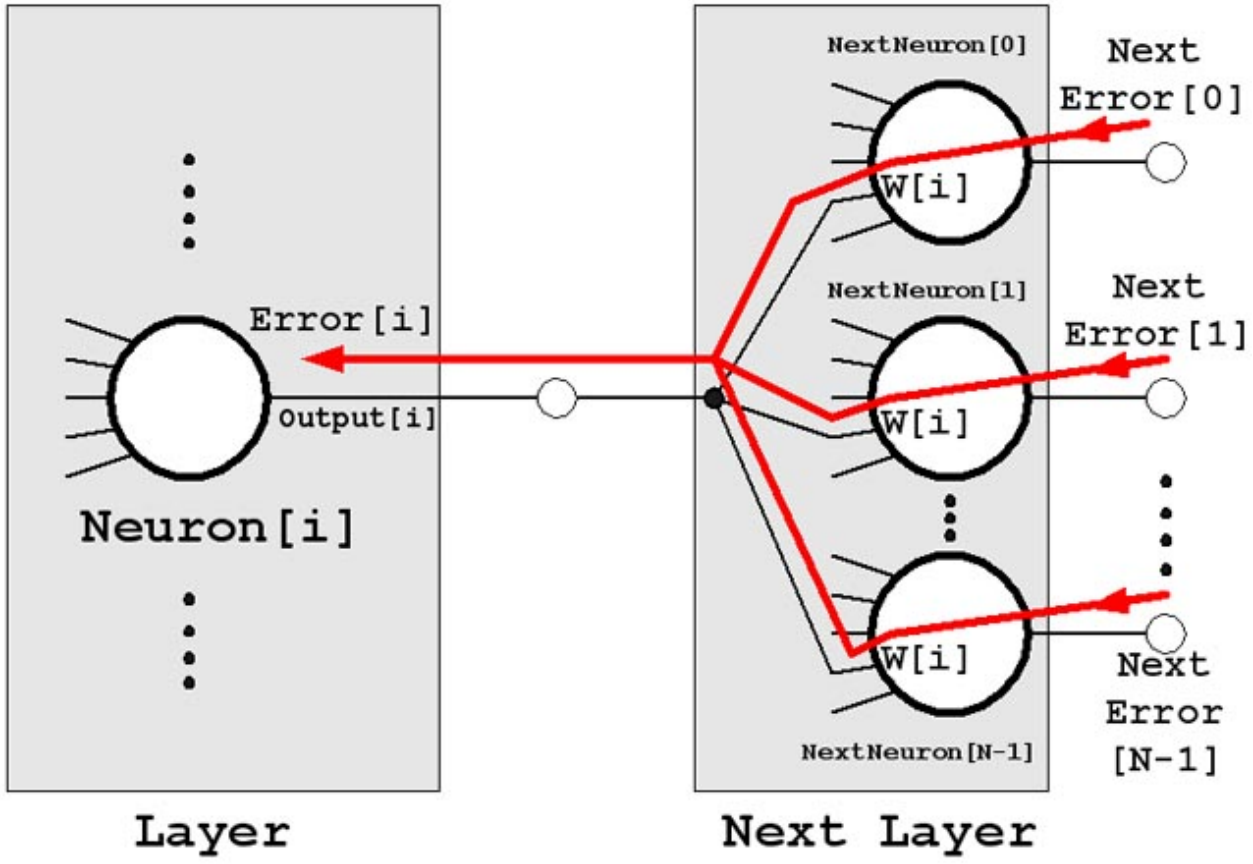
Hidden layer: **back-propagate** the error from the output layer:

$$\Delta_j = g'(in_j) \sum_i W_{j,i} \Delta_i \ .$$

Update rule for weights in hidden layer:

$$W_{k,j} \leftarrow W_{k,j} + \alpha \times a_k \times \Delta_j \ .$$

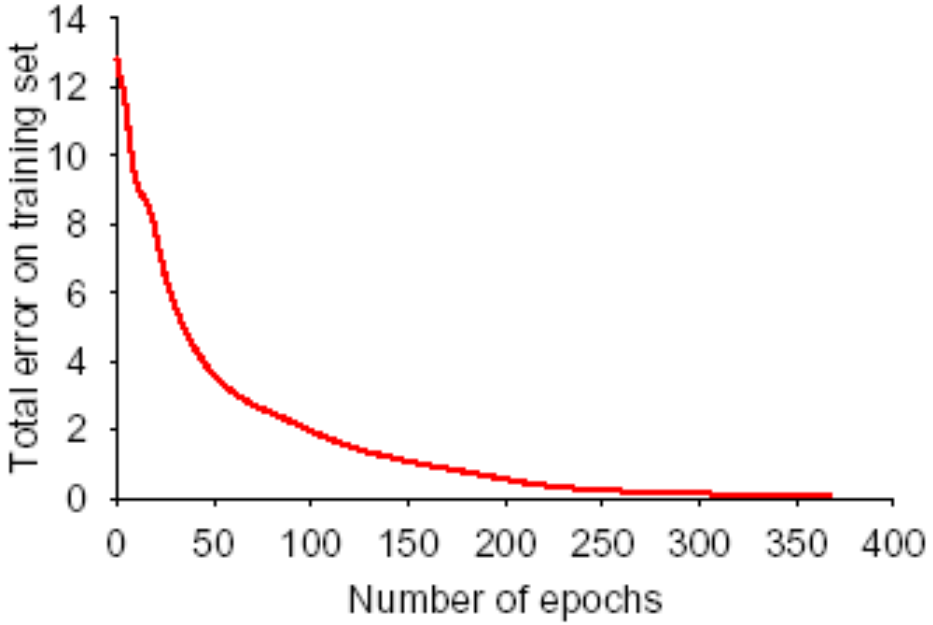
(Most neuroscientists deny that back-propagation occurs in the brain)


NextNeuron[0]
Next
Error[0]
W[i]
NextNeuron[1]
Next
Error[1]
Error[i]
output[i]
W[i]
Neuron[i]
W[i]
Next
Error
[N-1]
NextNeuron[N-1]
Layer
Next Layer

Back-propagation learning contd.

At each epoch, sum gradient updates for all examples and apply

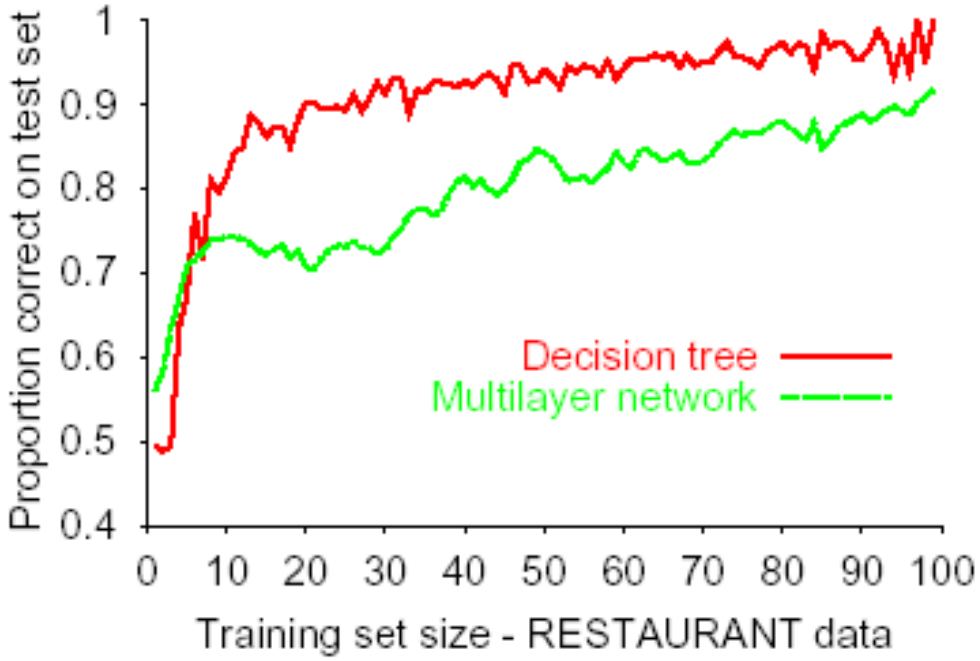
Training curve for 100 restaurant examples: finds exact fit



Typical problems: slow convergence, local minima

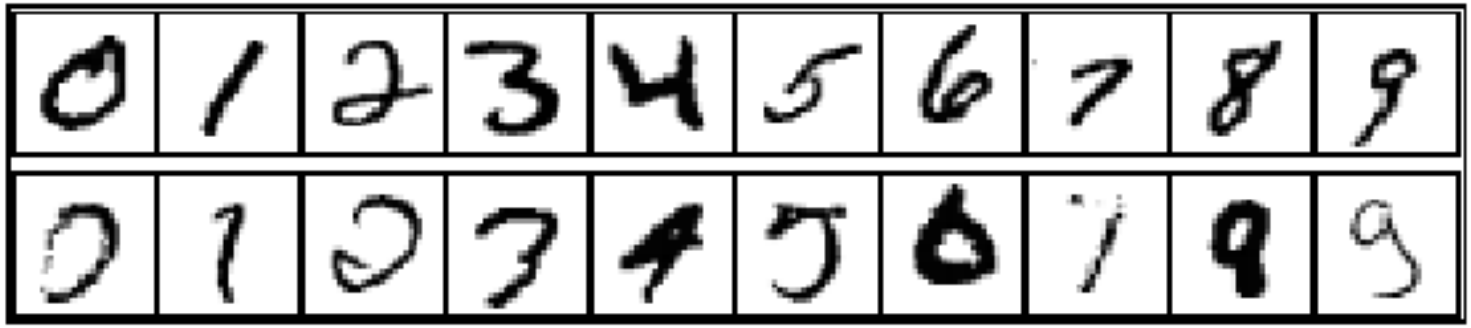
Back-propagation learning contd.

Learning curve for MLP with 4 hidden units:



MLPs are quite good for complex pattern recognition tasks, but resulting hypotheses cannot be understood easily

Handwritten digit recognition



3-nearest-neighbor = 2.4% error
400–300–10 unit MLP = 1.6% error
LeNet: 768–192–30–10 unit MLP = 0.9% error

Current best (kernel machines, vision algorithms) $\approx$ 0.6% error

Summary

Most brains have lots of neurons; each neuron $\approx$ linear–threshold unit (?)

Perceptrons (one-layer networks) insufficiently expressive

Multi-layer networks are sufficiently expressive; can be trained by gradient descent, i.e., error back-propagation

Many applications: speech, driving, handwriting, fraud detection, etc.

Engineering, cognitive modelling, and neural system modelling
subfields have largely diverged

Kernel Machines

- A relatively new learning methodology (1992) derived from statistical learning theory.
- Became famous when it gave accuracy comparable to neural nets in a handwriting recognition class.
- Was introduced to computer vision researchers by Tomaso Poggio at MIT who started using it for face detection and got better results than neural nets.
- Has become very popular and widely used with packages available.

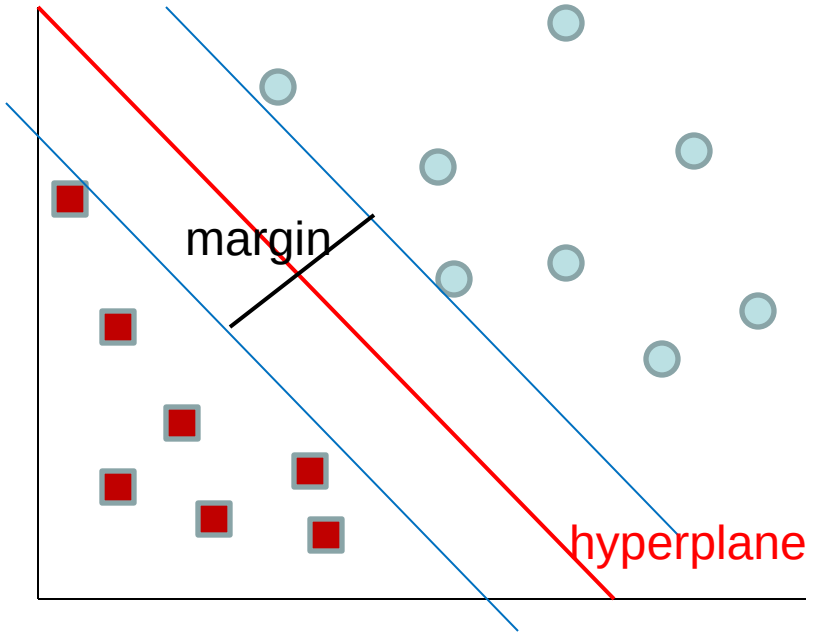
Support Vector Machines (SVM)

- Support vector machines are learning algorithms that try to find a hyperplane that separates the different classes of data the most.
- They are a specific kind of kernel machines based on two key ideas:
  - maximum margin hyperplanes
  - a kernel ‘trick’

Maximal Margin (2 class problem)

In 2D space,
a hyperplane is
a line.

In 3D space,
it is a plane.



Find the hyperplane with maximal margin for all the points. This originates an optimization problem which has a unique solution.

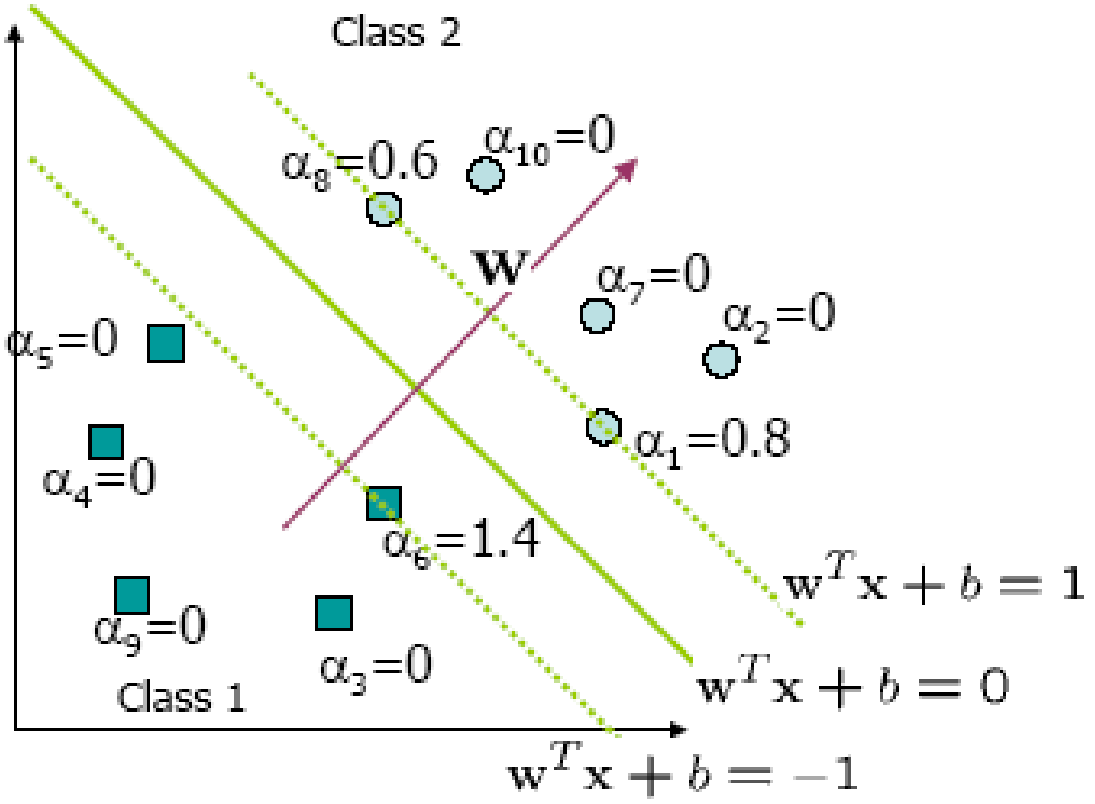
Support Vectors

- The weights α_i associated with data points are zero, except for those points closest to the separator.

- The points with nonzero weights are called the support vectors (because they hold up the separating plane).

- Because there are many fewer support vectors than total data points, the number of parameters defining the optimal separator is small.

A Geometric Interpretation



Kernels

- A kernel is just a similarity function. It takes 2 inputs and decides how similar they are.

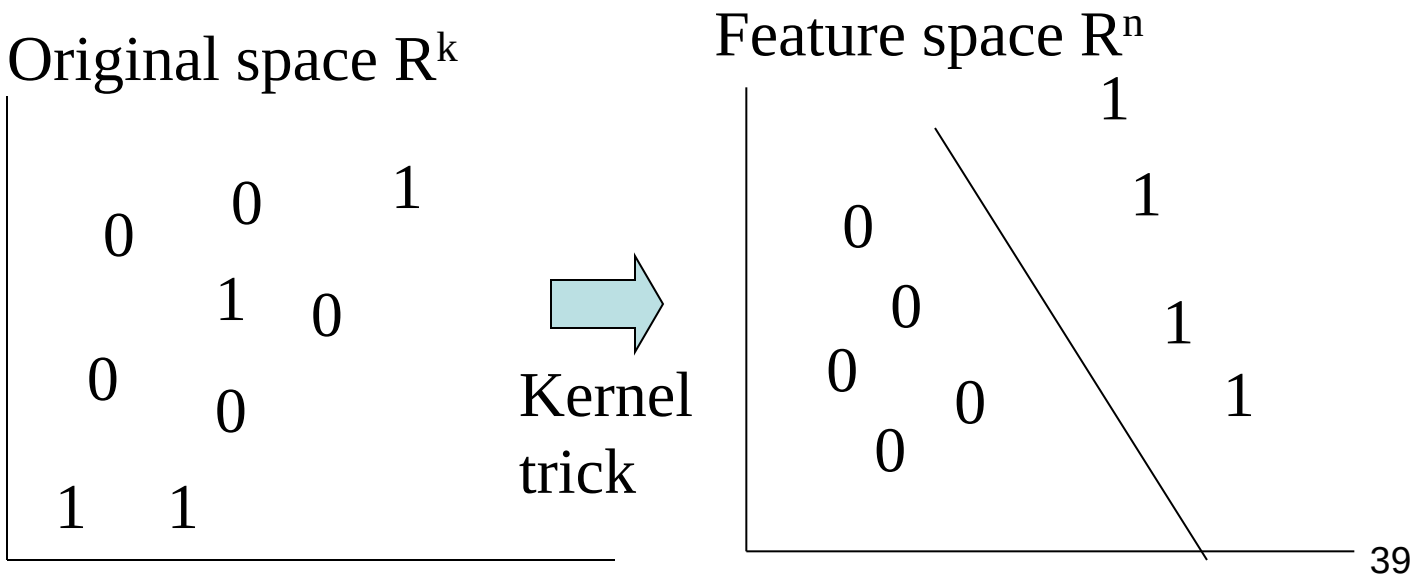
- Kernels offer an alternative to standard feature vectors. Instead of using a bunch of features, you define a single kernel to decide the similarity between two objects.

Kernels and SVMs

- Under some conditions, every kernel function can be expressed as a dot product in a (possibly infinite dimensional) feature space (Mercer's theorem)
- SVM machine learning can be expressed in terms of dot products.
- So SVM machines can use kernels instead of feature vectors.

The Kernel Trick

The SVM algorithm implicitly maps the original data to a feature space of possibly infinite dimension in which data (which is not separable in the original space) becomes separable in the feature space.

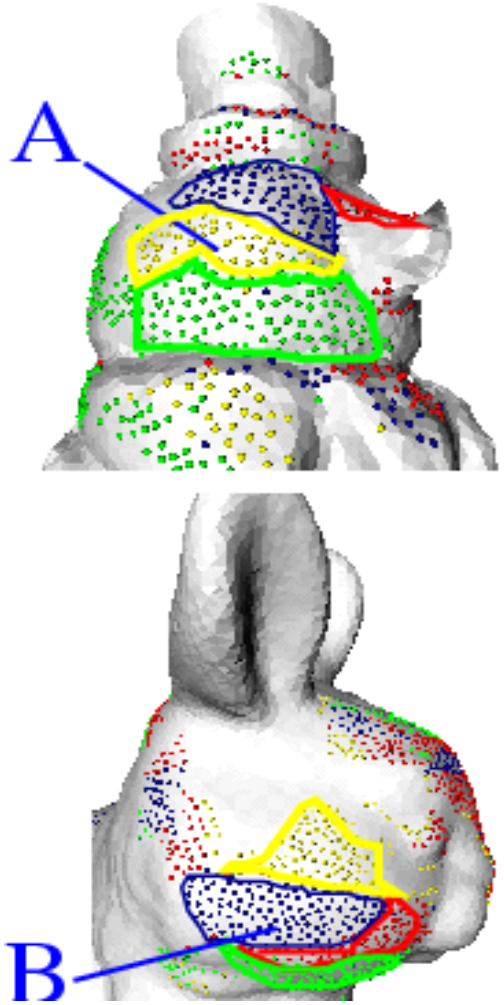


Kernel Functions

- The kernel function is designed by the developer of the SVM.
- It is applied to pairs of input data to evaluate dot products in some corresponding feature space.
- Kernels can be all sorts of functions including polynomials and exponentials.

Kernel Function used in our 3D Computer Vision Work

- $k(A,B) = \exp(-\theta^2_{AB}/\sigma^2)$
- A and B are shape descriptors (big vectors).
- θ is the angle between these vectors.
- σ^2 is the "width" of the kernel.



What do SVMs solve?

- The SVM is looking for the best separating plane in its alternate space.
- It solves a quadratic programming optimization problem

$$\underset{\alpha}{\textit{argmax}} \sum_j \alpha_j - 1/2 \sum_{j,k} \alpha_j \alpha_k y_j y_k (\mathbf{x}_j \bullet \mathbf{x}_k)$$

subject to $\alpha_j > 0$ and $\sum_j \alpha_j y_j = 0$.

- The equation for the separator for these optimal α_j is

$$h(\mathbf{x}) = \text{sign}(\sum_j \alpha_j y_j (x \bullet x_j) - b)$$

Unsupervised Learning

- Find patterns in the data.
- Group the data into clusters.
- Many clustering algorithms.
  - K means clustering
  - EM clustering
  - Graph-Theoretic Clustering
  - Clustering by Graph Cuts
  - etc

Clustering by K-means Algorithm

Form K-means clusters from a set of n-dimensional feature vectors

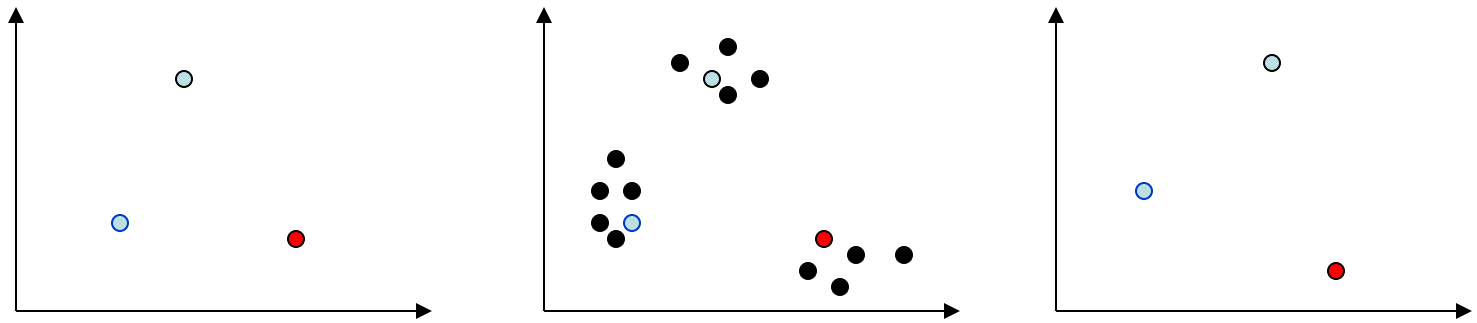
1. Set ic (iteration count) to 1

2. Choose randomly a set of K means $m_1(1), \ldots, m_K(1)$.

3. For each vector x_i, compute $D(x_i, m_k(ic))$, $k=1,\ldots K$
   and assign x_i to the cluster C_j with nearest mean.

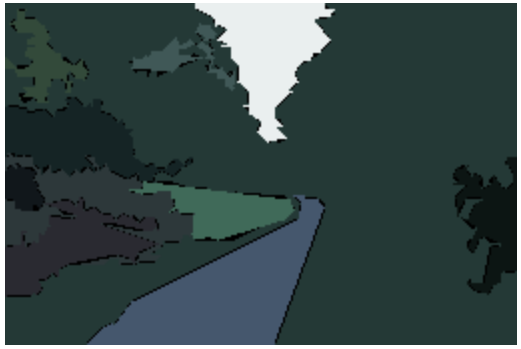
4. Increment ic by 1, update the means to get $m_1(ic),\ldots,m_K(ic)$.

5. Repeat steps 3 and 4 until $C_k(ic) = C_k(ic+1)$ for all k.



K-Means Classifier (shown on RGB color data)

original data
one RGB per pixel



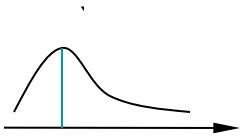
color clusters

K-Means → EM

The clusters are usually Gaussian distributions.

- Boot Step:
  - Initialize K clusters: $C_1, \ldots, C_K$

    (μ_j, Σ_j) and $P(C_j)$ for each cluster j.



- Iteration Step:
  - Estimate the cluster of each datum

    $p(C_j \mid x_i)$ ➡ Expectation
  - Re-estimate the cluster parameters

    $(\mu_j, \Sigma_j), p(C_j)$ For each cluster j ➡ Maximization

The resultant set of clusters is called a **mixture model**; if the distributions are Gaussian, it's a Gaussian mixture.

EM Algorithm Summary

- Boot Step:
  - Initialize K clusters: $C_1, \ldots, C_K$

    (μ_j, Σ_j) and $p(C_j)$ for each cluster j.
- Iteration Step:
  - Expectation Step

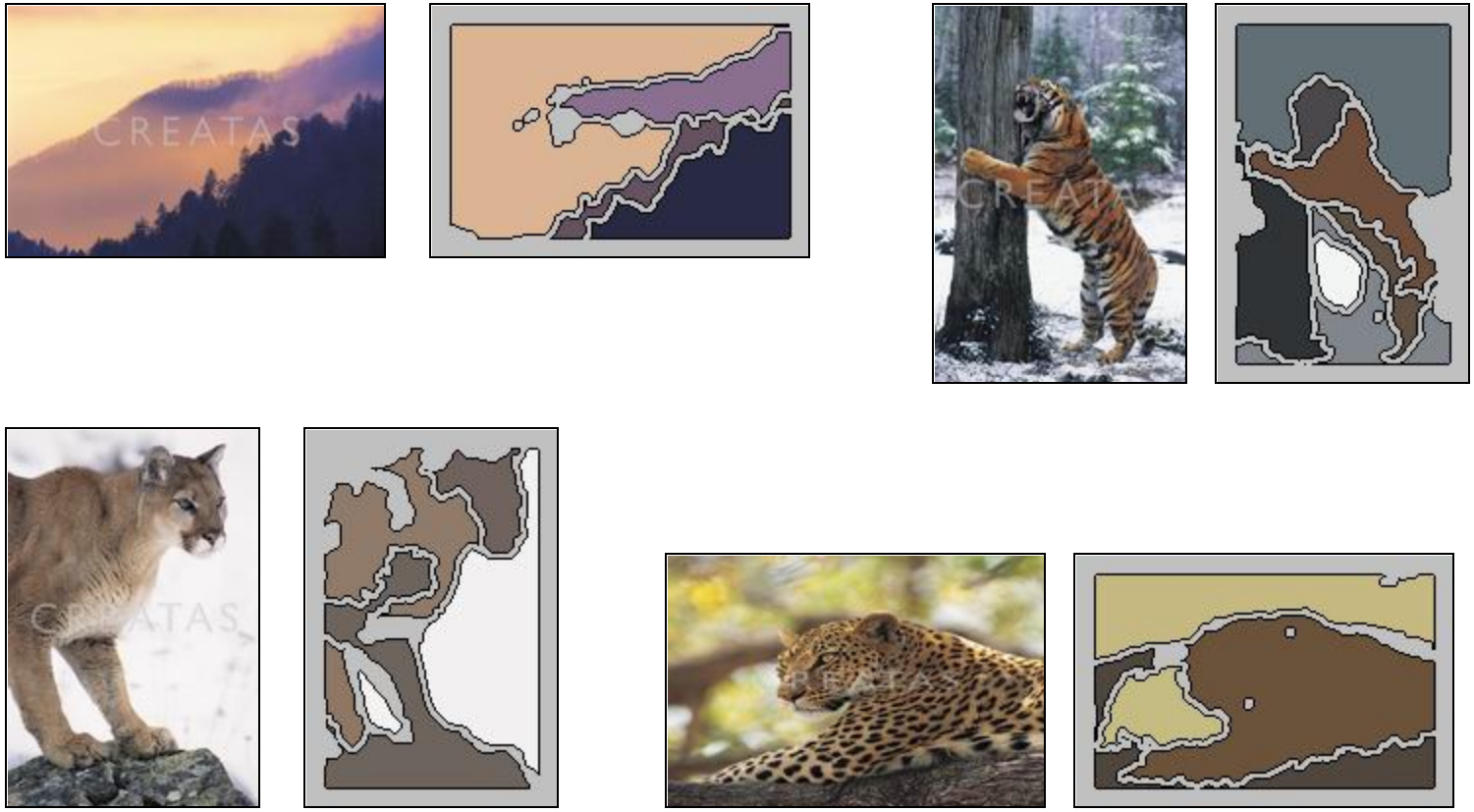
$$p(C_j \mid x_i) = \frac{p(x_i \mid C_j) \cdot p(C_j)}{p(x_i)} = \frac{p(x_i \mid C_j) \cdot p(C_j)}{\sum_j p(x_i \mid C_j) \cdot p(C_j)}$$

  - Maximization Step

$$\mu_j = \frac{\sum_i p(C_j \mid x_i) \cdot x_i}{\sum_i p(C_j \mid x_i)} \qquad \Sigma_j = \frac{\sum_i p(C_j \mid x_i) \cdot (x_i - \mu_j) \cdot (x_i - \mu_j)^T}{\sum_i p(C_j \mid x_i)} \qquad p(C_j) = \frac{\sum_i p(C_j \mid x_i)}{N}$$

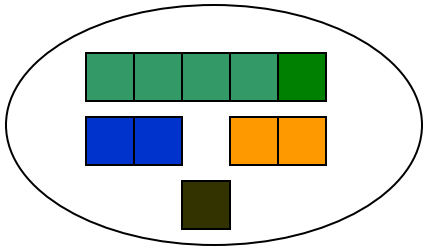
EM Clustering using color and texture information at each pixel

(from Blobworld)

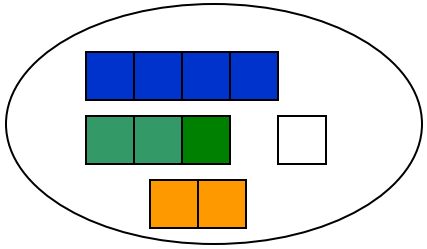


EM for Classification of Images in Terms of their Color Regions

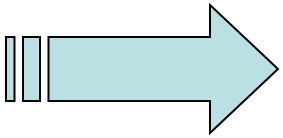
Initial Model for “trees”



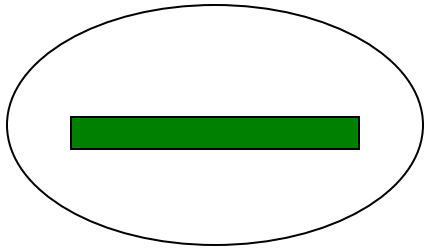
Initial Model for “sky”



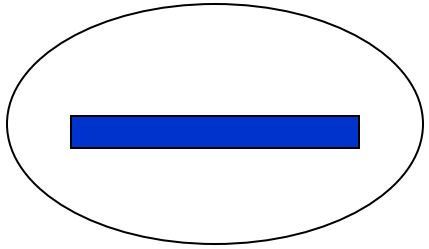
EM



Final Model for “trees”



Final Model for “sky”



Sample Results

cheetah

Sample Results (Cont.)

grass

Sample Results (Cont.)

lion