Artificial Intelligence:
Representation and Problem Solving

15-381

April 26, 2007

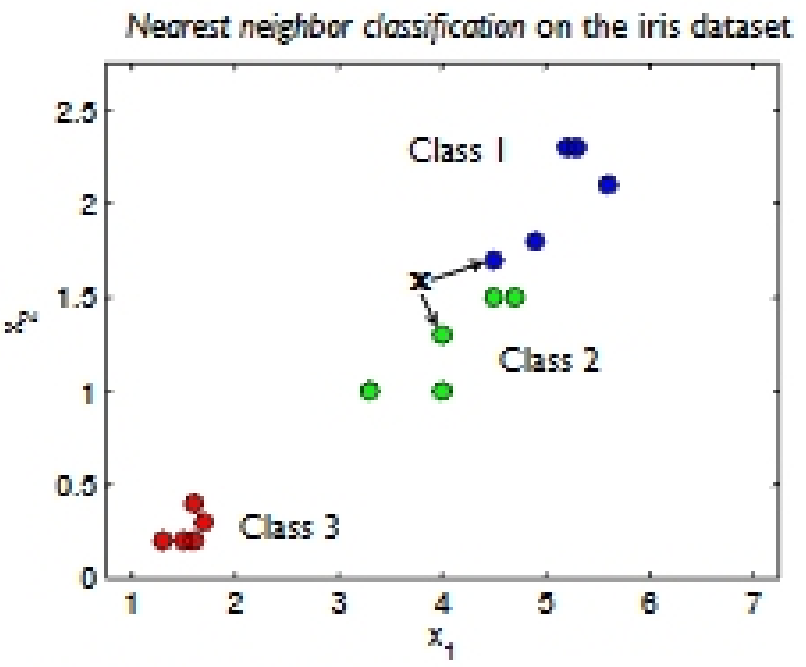
Clustering

(including k-nearest neighbor classification, k-means clustering, cross-validation, and EM, with a brief foray into dimensionality reduction with PCA)

A different approach to classification

- Nearby points are likely to be members of the same class.
- What if we used the points themselves to classify?

  classify $\mathbf{x}$ *in* C_k *if* $\mathbf{x}$ *is "similar" to a point we already know is in* C_k



- Eg: unclassified point $\mathbf{x}$ is more similar Class 2 than Class 1.
- Issue: How to define "similar" ? Simplest is Euclidean distance:

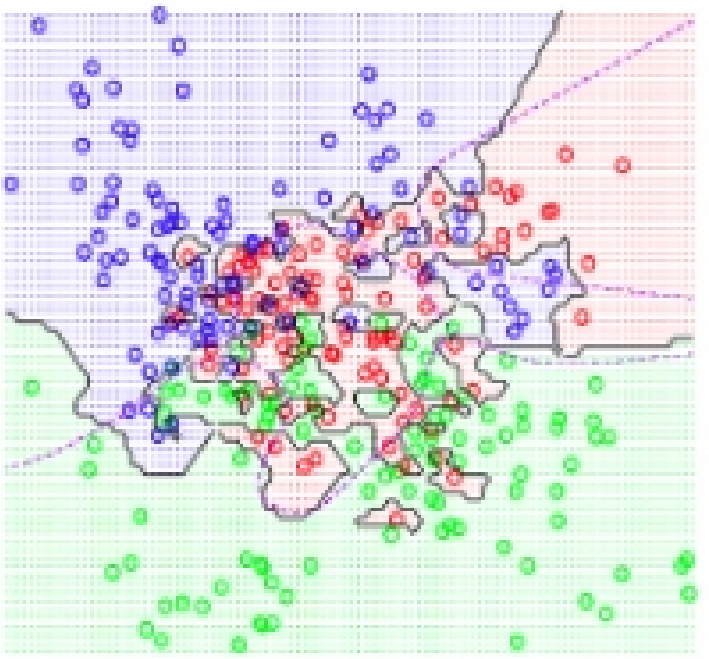
$$d(\mathbf{x}, \mathbf{y}) = \sum_i (x_i - y_i)^2$$

- Could define other metrics depending on application, e.g. text documents, images, etc.

Potential advantages:

- don't need an explicit model
- the more examples the better
- might handle more complex classes
- easy to implement
- "no brain on part of the designer"

A complex, non-parametric decision boundary



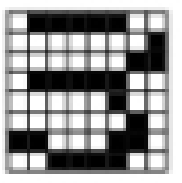
example from Martial Herbert

- How do we control the complexity of this model?
  - difficult
- How many parameters?
  - every data point is a parameter!
- This is an example of a *non-parametric model*, ie where the model is defined by the data. (Also, called, instance based)
- Can get very complex decision boundaries

Example: Handwritten digits

from LeCun et al, 1998
digit data available at:
http://yann.lecun.com/exdb/mnist/

Digits are just represented as a vector.



- Use Euclidean distance to see which known digit is closest to each class.
- But not all neighbors are the same:

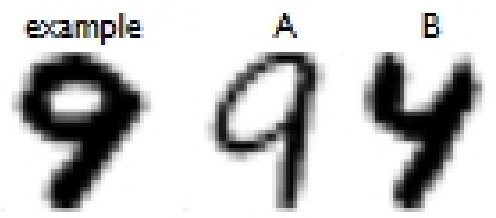
example from Sam Roweis

- "k-nearest neighbors":

  look at k-nearest neighbors and choose most frequent.
- Cautions: can get expensive to find neighbors

The problem of using templates (ie Euclidean distance)

- Which of these is more like the example? A or B?



from Simard et.al, 1998

- Euclidean distance only cares about how many pixels overlap.
- Could try to define a distance metric that is insensitive to small deviations in position, scale, rotation, etc.
- Digit example:
  - 60,000 training images,
  - 10,000 test images
  - no "preprocessing"

performance results of various classifiers (from http://yann.lecun.com/exdb/mnist/)

Classifier	error rate on test data (%)
linear	12.0
k=3 nearest neighbor (Euclidean distance)	5.0
2-layer neural network (300 hidden units)	4.7
nearest neighbor (Euclidean distance)	3.1
k-nearest neighbor (improved distance metric)	1.1
convolutional neural net	0.95
best (the conv. net with elastic distortions)	0.4
humans	0.2 - 2.5

Clustering