10-701
Machine Learning

Classification

Types of classifiers

- We can divide the large variety of classification approaches into roughly two main types

  1. Instance based classifiers
     - Use observation directly (no models)
     - e.g. K nearest neighbors

  2. Generative:
     - build a generative statistical model
     - e.g., Naïve Bayes

  3. Discriminative
     - directly estimate a decision rule/boundary
     - e.g., decision tree

Classification

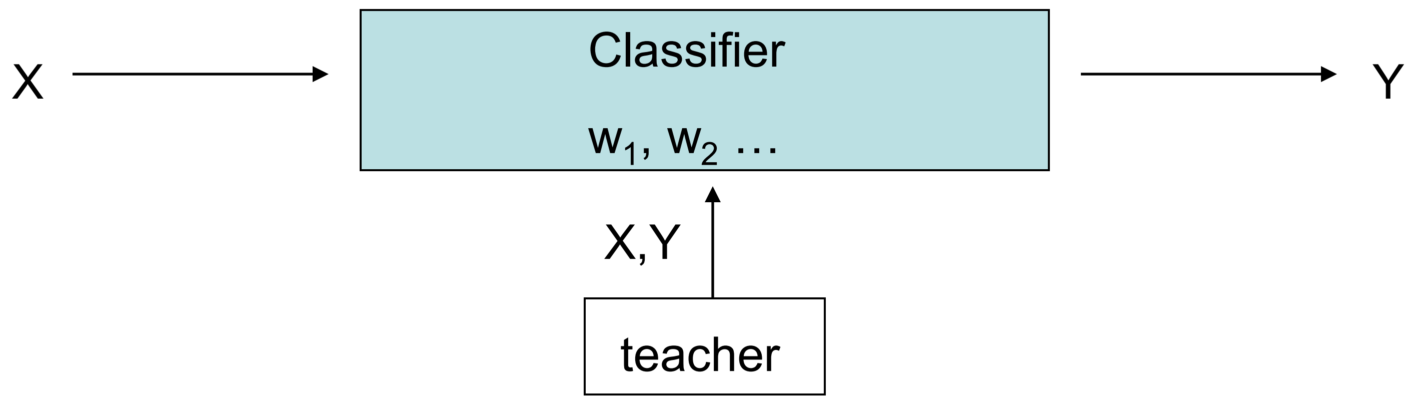
- Assume we want to teach a computer to distinguish between cats and dogs …




Several steps:
1. feature transformation
2. Model / classifier specification
3. Model / classifier estimation (with regularization)
4. feature selection

Supervised learning

- Classification is one of the key components of 'supervised learning'
- Unlike other learning paradigms, in supervised learning the teacher (us) provides the algorithm with the solutions to some of the instances and the goal is to generalize so that a model / method can be used to determine the labels of the unobserved samples



Types of classifiers

- We can divide the large variety of classification approaches into roughly two main types

1. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

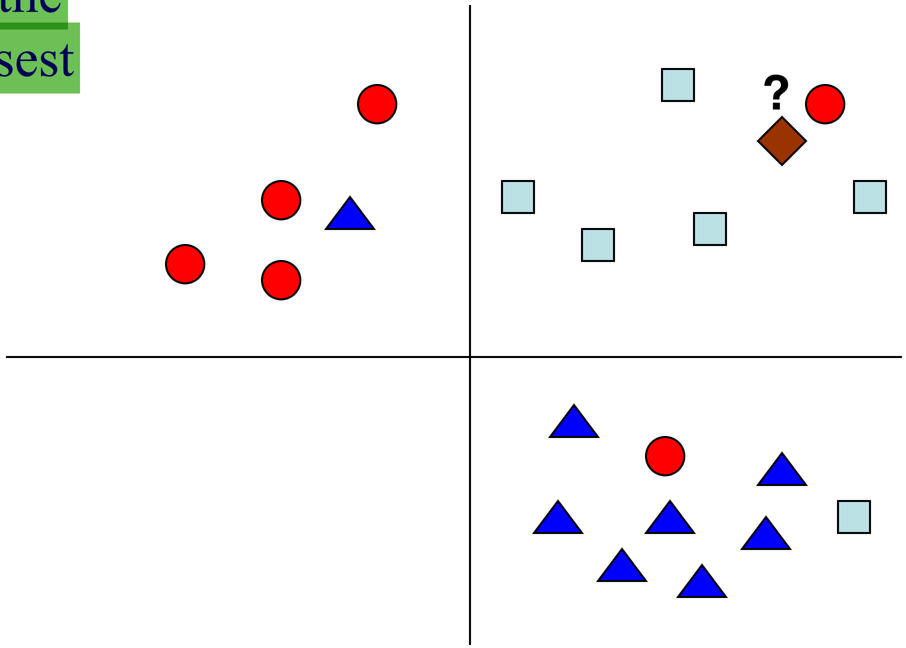
2. Generative:
   - build a generative statistical model
   - e.g., Bayesian networks

3. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., decision tree

K nearest neighbors

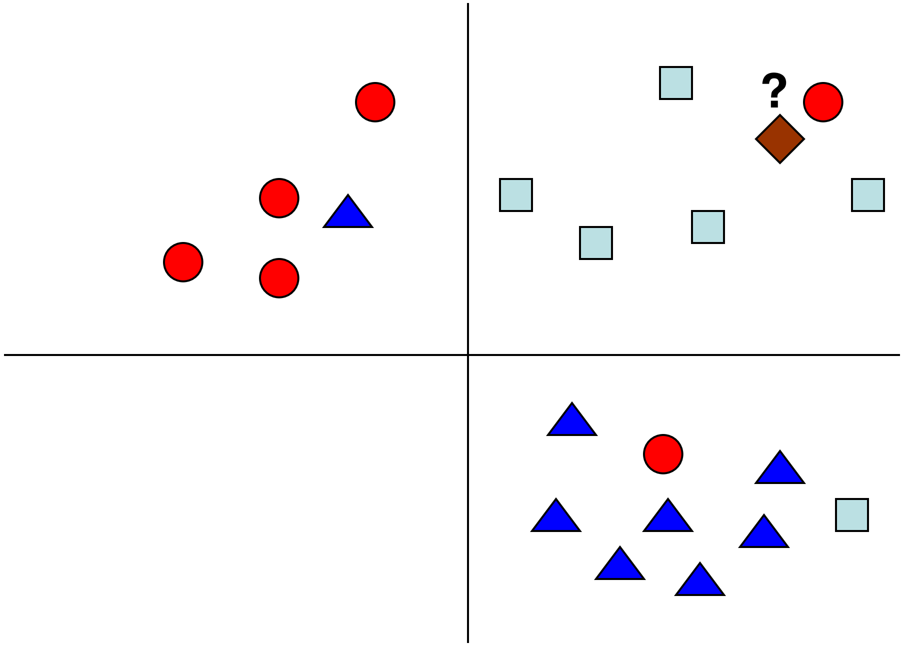
K nearest neighbors (KNN)

- A simple, yet surprisingly efficient algorithm
- Requires the definition of a distance function or similarity measures between samples
- Select the class based on the majority vote in the k closest points



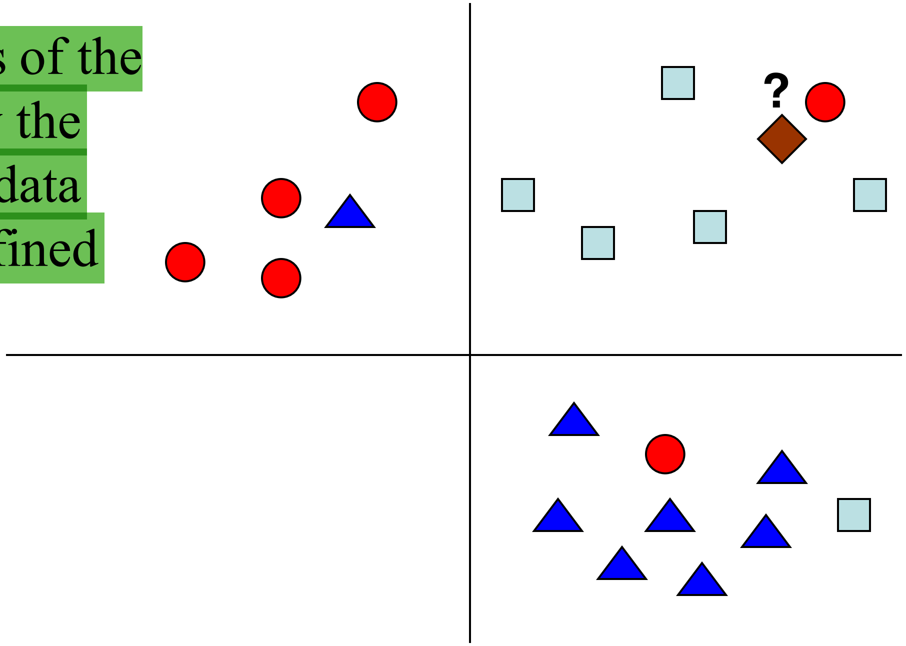
K nearest neighbors (KNN)

- Need to determine an appropriates value for k
- What happens if we chose k=1?
- What if k=3?



K nearest neighbors (KNN)

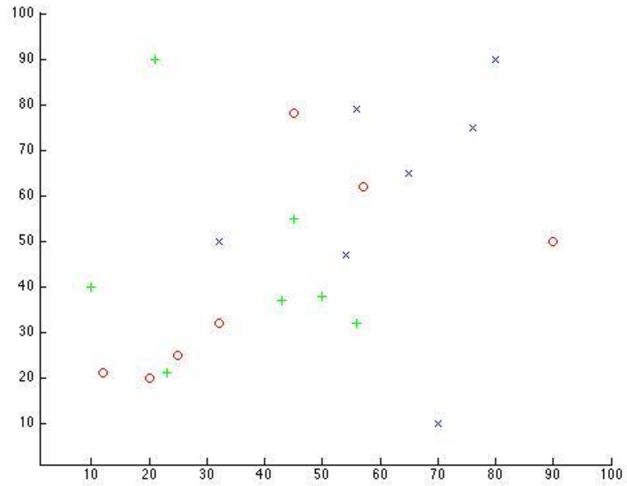
- Choice of k influences the ‘smoothness’ of the resulting classifier
- In that sense it is similar to a kernel methods (discussed later in the course)
- However, the smoothness of the function is determined by the actual distribution of the data (p(x)) and not by a predefined parameter.



The effect of increasing k

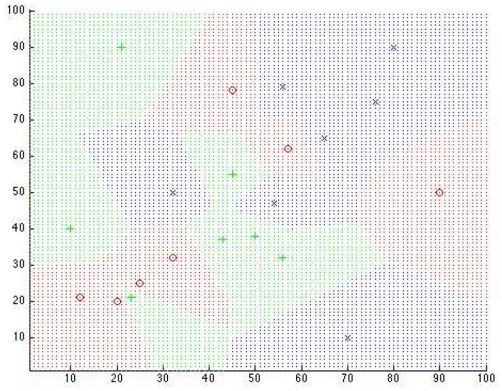
We will be using Euclidian distance to determine what are the k nearest neighbors:

$$d(x,x') = \sqrt{\sum_i (x_i - x_i')^2}$$

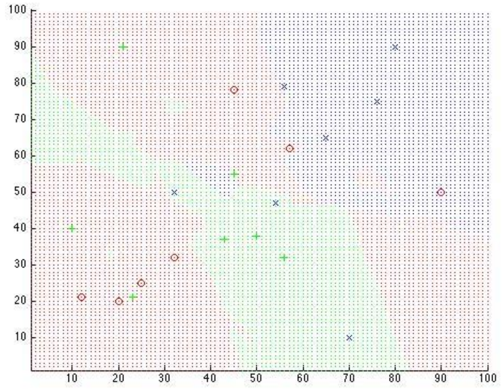


Comparisons of different k's

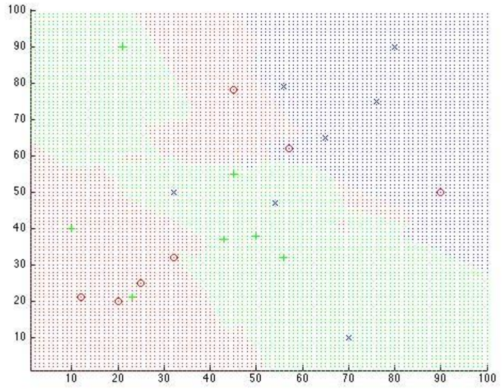
K = 1



K = 3



K = 5



A probabilistic interpretation of KNN

- The decision rule of KNN can be viewed using a probabilistic interpretation
- What KNN is trying to do is approximate the Bayes decision rule on a subset of the data
- To do that we need to compute certain properties including the conditional probability of the data given the class ($p(x|y)$), the prior probability of each class ($p(y)$) and the marginal probability of the data ($p(x)$)
- These properties would be computed for some small region around our sample and the size of that region will be *dependent on the distribution of the test samples**

* Remember this idea. We will return to it when discussing kernel functions

Computing probabilities for KNN

- Let V be the volume of the m dimensional ball around z containing the k nearest neighbors for z (where m is the number of features).
- Then we can write

$$p(x)V = P = \frac{K}{N} \qquad p(x) = \frac{K}{NV} \qquad p(x \mid y=1) = \frac{K_1}{N_1 V} \qquad p(y=1) = \frac{N_1}{N}$$

- Using Bayes rule we get:

$$p(y=1 \mid z) = \frac{p(z \mid y=1)\,p(y=1)}{p(z)} = \frac{K_1}{K}$$

z – new data point to classify

V - selected ball

P – probability that a random point is in V

N - total number of samples

K - number of nearest neighbors

N_1 - total number of samples from class 1

K_1 - number of samples from class 1 in K

Computing probabilities for KNN

N - total number of samples

V - Volume of selected ball

K - number of nearest neighbors

N_1 - total number of samples from class 1

K_1 - number of samples from class 1 in K

- Using Bayes rule we get:

$$p(y=1 \mid z) = \frac{p(z \mid y=1)\,p(y=1)}{p(z)} = \frac{K_1}{K}$$

Using Bayes decision rule we will chose the class with the highest probability, which in this case is the class with the highest number of samples in K

Important points

- Optimal decision using Bayes rule
- Types of classifiers
- Effect of values of k on knn classifiers
- Probabilistic interpretation of knn