CSC2515 Fall 2008
Introduction to Machine Learning

Lecture 10a
Kernel density estimators and nearest neighbors

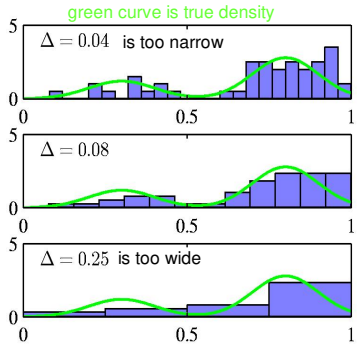
All lecture slides will be available as .ppt, .ps, & .htm at
www.cs.toronto.edu/~hinton

Many of the figures are provided by Chris Bishop
from his textbook: "Pattern Recognition and Machine Learning"

Histograms as density models

- For low dimensional data we can use a histogram as a density model.
  - How wide should the bins be? (width=regulariser)
  - Do we want the same bin-width everywhere?
  - Do we believe the density is zero for empty bins?

$$p_i = \frac{n_i}{N \Delta_i}$$



Some good and bad properties of histograms as density estimators

- There is no need to fit a model to the data.
  - We just compute some very simple statistics (the number of datapoints in each bin) and store them.
- The number of bins is exponential in the dimensionality of the dataspace. So high-dimensional data is tricky:
  - We must either use big bins or get lots of zero counts (or adapt the local bin-width to the density)
- The density has silly discontinuities at the bin boundaries.
  - We must be able to do better by some kind of smoothing.

Local density estimators

- Estimate the density in a small region to be

$$p(x) = \frac{K}{N V}$$

K ← points in region
V ← volume of region
N ↑ total points

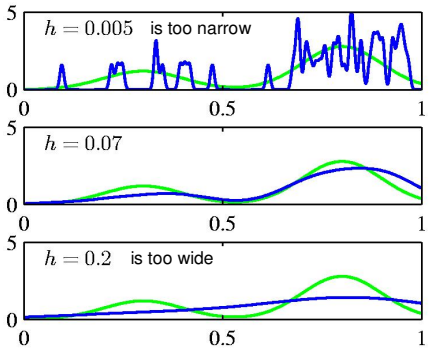
- Problem 1: Variance in estimate if K is small.
- Problem 2: Unmodelled variation across the region if V is big compared with the smoothness of the true density

Kernel density estimators

- Use regions centered on the datapoints
  - Allow the regions to overlap.
  - Let each individual region contribute a total density of 1/N
  - Use regions with soft edges to avoid discontinuities (e.g. isotropic Gaussians)

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2\sigma^2}\right)$$

The density modeled by a kernel density estimator



Nearest neighbor methods for density estimation

$$p(x) = \frac{K}{N\,V}$$

K ← points in region

V ← volume of region

N ↑ total points

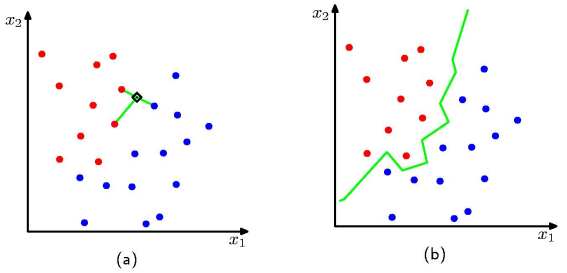
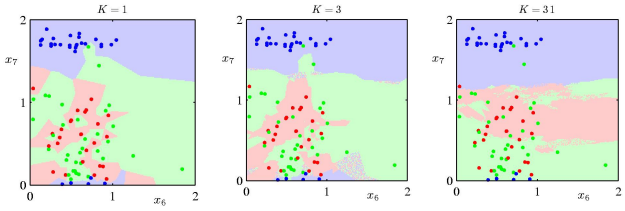
- Vary the size of a hyper-sphere around each test point so that exactly K training datapoints fall inside the hyper-sphere.
  - Does this give a fair estimate of the density?
- Nearest neighbors is usually used for classification or regression:
  - For regression, average the predictions of the K nearest neighbors.
  - For classification, pick the class with the most votes.
    - How should we break ties?

Nearest neighbor methods for classification and regression

- Nearest neighbors is usually used for classification or regression:
- For regression, average the predictions of the K nearest neighbors.
  - How should we pick K?
- For classification, pick the class with the most votes.
  - How should we break ties?
  - Let the k'th nearest neighbor contribute a count that falls off with k. For example, $1+\frac{1}{2^k}$

The decision boundary implemented by 3NN

The boundary is always the perpendicular bisector of the line between two points (Vornoi tesselation)



Regions defined by using various numbers of neighbors