THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

COSC 522 – Machine Learning

Lecture 5 – Nonparametric Learning

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

Racap - Bayes Decision Rule

AICIP RESEARCH

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

Maximum Posterior Probability

For a given x, if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$, then x belongs to class 1, otherwise, 2.

Discriminant Function

The classifier will assign a feature vector x to class ω_i if $g_i(x) > g_j(x)$

Case 1: Minimum Euclidean Distance (Linear Machine), $\Sigma_i = \sigma^2 I$

Case 2: Minimum Mahalanobis Distance (Linear Machine), $\Sigma_i = \Sigma$

Case 3: Quadratic classifier , Σ_i = arbitrary

All assuming Gaussian pdf

Estimate Gaussian parameters using MLE

Questions

AICIP RESEARCH

- In general, what is non-parametric learning?
- Under what conditions that non-parametric learning would be preferred?
- What is parzen window and what are the potential issues?
- What is kNN intuitively?
- Is kNN optimal in Baysian sense?
- We know the three cases of discriminant functions essentially follow the MPP decision rule. Does kNN also follow the MPP decision rule?
- What is the decision boundary of kNN?
- When k is fixed, is the radius of neighborhood fixed?
- Is 1NN the same as minimum distance classifier?
- What is the cost function of kNN? What is the optimization approach used?
- What are the potential issues with kNN?

intuitive explanation

kNN and MPP?

issues

Motivation

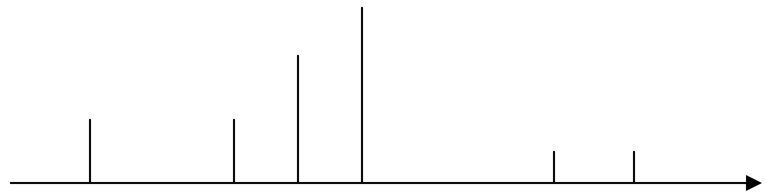
- Estimate the density functions without the assumption that the pdf has a particular form

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)}$$

Start from Histogram

AICIP RESEARCH

- In order to generate a reasonable representation for the density, we'd like to first "smooth" the data over cells



- The probability that a vector *x* will fall into a region *R* is

$$P = \int_R p(x')dx'$$

- If *p*(*x*) does not vary significantly within *R*, then
  - *V* is the volume enclosed by *R*

$$P = p(x)V$$

- For a training set of *n* samples, *k* of them fall into the hypervolume *V*, we can then estimate *p*(*x*) by

$$p(x) \approx p_n(x) = \frac{k_n / n}{V_n}$$

Questions

- In general, what is non-parametric learning?
- Under what conditions that non-parametric learning would be preferred?
- What is parzen window and what are the potential issues?
- What is kNN intuitively?
- Is kNN optimal in Baysian sense?
- We know the three cases of discriminant functions essentially follow the MPP decision rule. Does kNN also follow the MPP decision rule?
- What is the decision boundary of kNN?
- When k is fixed, is the radius of neighborhood fixed?
- Is 1NN the same as minimum distance classifier?
- What is the cost function of kNN? What is the optimization approach used?
- What are the potential issues with kNN?

intuitive explanation

kNN and MPP?

issues

Parzen Windows

$$p_n(x) = \frac{k_n / n}{V_n}$$

- The density estimation at *x* is calculated by counting the number of samples fall within a hypercube of volume V_n centered at *x*
- Let *R* be a *d*-dimensional hypercube, whose edges are h_n units long. Its volume is then $V_n = h_n^d$
- The window function

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \le 0.5, \quad j = 1, \cdots, d \\ 0 & \text{otherwise} \end{cases}$$

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

- Therefore

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)}{V_n}$$

Problem

- Hypercube – why should a point just inside the hypercube contribute the same as a point very near to **x**, while a point just outside the hypercube contributes nothing?
- Use a continuous window function

Continuous Window Function

AICIP RESEARCH

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)}{V_n}$$

- Univariate $\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$
- Multi-variate

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n^d}\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)^T\Sigma^{-1}\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right]$$

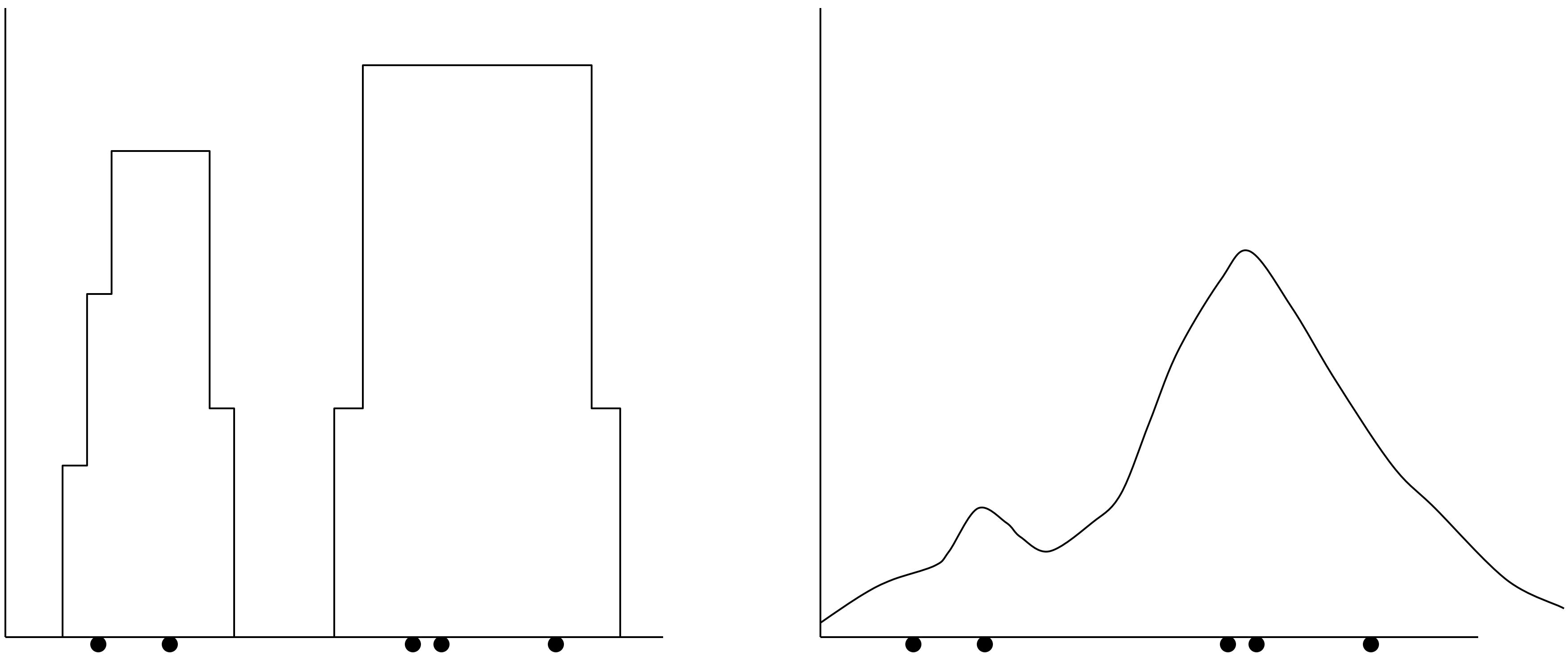
- Making Σ an identity matrix

$$p(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1 h_2\cdots h_d}\frac{1}{(2\pi)^{d/2}}\prod_{j=1}^{d}\exp\left[-\frac{1}{2}\left(\frac{x_j - x_{ij}}{h_j}\right)^2\right]$$

- h_j reflects the variance (spread) of the smoothing kernel (window function) in the *j*th coordinate direction. If we assume the spread is equal in all directions

$$p(x) = \frac{1}{nh^d(2\pi)^{d/2}}\sum_{i=1}^{n}\prod_{j=1}^{d}\exp\left[-\frac{1}{2}\left(\frac{x_j - x_{ij}}{h}\right)^2\right]$$

Comparison



Another Problem

- How to choose *h*?
- A large *h* will result in a great deal of smoothing and loss of resolution
- A very small *h* will tend to degenerate the estimator into a collection of *n* sharp peaks, each centered at a sampling point
- Solution: *h* should depend on the number of samples. If only a few samples are available, we require a large *h* and considerable smoothing, whereas if many points are available, we can use a smaller *h* without the danger of degenerating into separate peaks.

The Choice of h

AICIP RESEARCH

- We make *h* a function of *n*

$$h = \frac{1}{\sqrt{n}}$$

Problem with Parzen Windows

- Discontinuous window function -> Gaussian
- The choice of h
- Still another one: fixed volume

Questions

- In general, what is non-parametric learning?
- Under what conditions that non-parametric learning would be preferred?
- What is parzen window and what are the potential issues?
- What is kNN intuitively?
- Is kNN optimal in Baysian sense?
- We know the three cases of discriminant functions essentially follow the MPP decision rule. Does kNN also follow the MPP decision rule?
- What is the decision boundary of kNN?
- When k is fixed, is the radius of neighborhood fixed?
- Is 1NN the same as minimum distance classifier?
- What is the cost function of kNN? What is the optimization approach used?
- What are the potential issues with kNN?

intuitive explanation

kNN and MPP?

issues

The k-nearest neighbor (kNN) Decision Rule - Intuitively

AICIP RESEARCH

- The decision rule tells us to look in a neighborhood of the unknown test sample for *k* samples. If within that neighborhood, more training samples lie in class *i* than any other class, we assign the unknown as belonging to class *i*.

kNN in Classification

AICIP RESEARCH

$$p_n(x) = \frac{k_n / n}{V_n}$$

- Given *c* training sets from *c* classes, the total number of samples is

$$n = \sum_{m=1}^{c} n_m$$

- Given a point **x** at which we wish to determine the statistics, we find the hypersphere of volume ***V*** which just encloses *k* points from the combined set. If within that volume, k_m of those points belong to class *m*, then we estimate the density for class *m* by

$$p(x \mid \omega_m) = \frac{k_m}{n_m V} \qquad P(\omega_m) = \frac{n_m}{n} \qquad p(x) = \frac{k}{nV}$$

kNN Classification Rule

$$P\left(\omega_m \mid x\right) = \frac{p\left(x \mid \omega_m\right) P\left(\omega_m\right)}{p\left(x\right)} = \frac{\frac{k_m}{n_m V} \frac{n_m}{n}}{\frac{k}{nV}} = \frac{k_m}{k}$$

- The decision rule tells us to look in a neighborhood of the unknown feature vector for *k* samples. If within that neighborhood, more samples lie in class *i* than any other class, we assign the unknown as belonging to class *i*.

kNN Decision Boundary

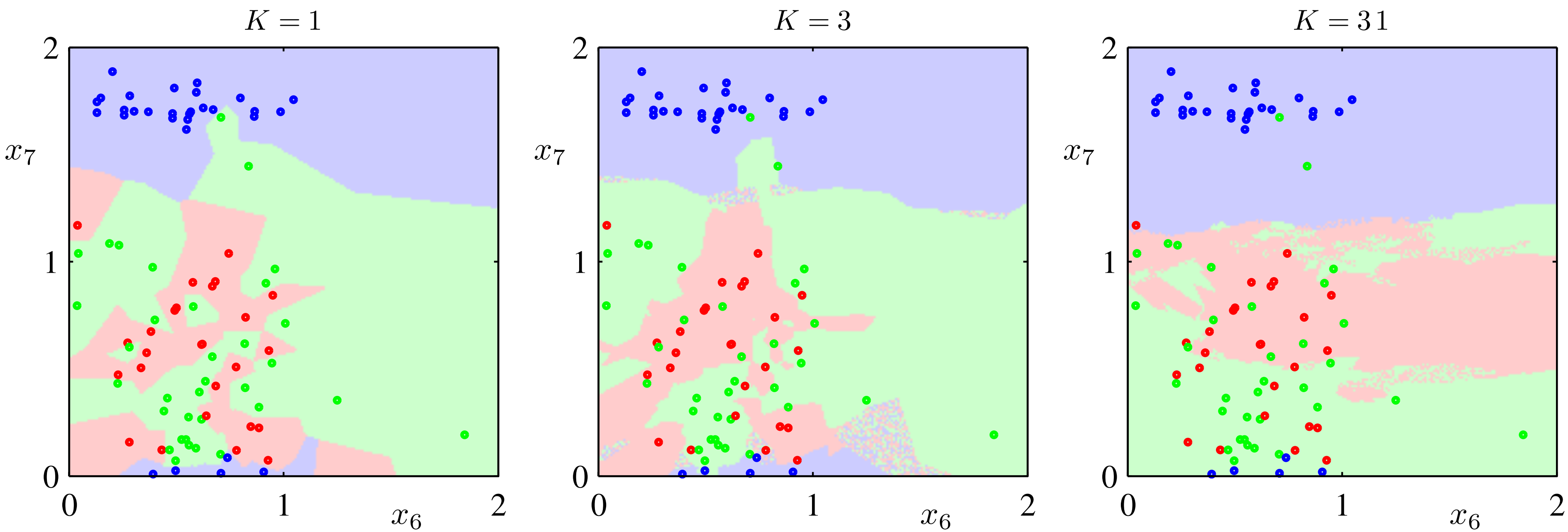


Figure 2.28 Plot of 200 data points from the oil data set showing values of x_6 plotted against x_7, where the red, green, and blue points correspond to the 'laminar', 'annular', and 'homogeneous' classes, respectively. Also shown are the classifications of the input space given by the K-nearest-neighbour algorithm for various values of K.

From [Bishop 2006]

Questions

AICIP RESEARCH

- In general, what is non-parametric learning?
- Under what conditions that non-parametric learning would be preferred?
- What is kNN?
- Is kNN optimal in Baysian sense?
- We know the three cases of discriminant functions essentially follow the MPP decision rule. Does kNN also follow the MPP decision rule?
- What is the decision boundary of kNN?
- When k is fixed, is the radius of neighborhood fixed?
- Is 1NN the same as minimum distance classifier?
- What is the cost function of kNN? What is the optimization approach used?
- What are the potential issues with kNN?

intuitive explanation

kNN and MPP?

issues

Potential Issues

AICIP RESEARCH

- What is a good value of “k”?
- What kind of distance should be used to measure “nearest”
  - Euclidean metric is a reasonable measurement
- Computation burden
  - Massive storage burden
  - Need to compute the distance from the unknown to all the neighbors

kNN (k-Nearest Neighbor)

- To estimate $p(x)$ from n samples, we can center a cell at x and let it grow until it contains k_n samples, and k_n can be some function of n
- Normally, we let

$$k_n = \sqrt{n}$$