CS542
Neural Computation With
Artificial Neural Networks

Lecture 15

Nonparametric Methods:
Density Estimation and Classification
Nonparametric Methods: Density Estimation and Classification
  - Kernel Methods
  - k-Nearest Neighbors
  - Nonparametric Classification
  - kd-Trees

Handout:
  - Class Notes
  - Ch.4 of Duda & Hart 1977
  - Read Bishop Ch. 2.5 and 2.6

Reading Assignment for Next Class
  - Ch.4 of Duda & Hart 1977
Nonparametric Methods

- **Working Definition**
  - The name “nonparametric” is to indicate that the data to be modeled stem from very large families of distributions which cannot be indexed by a finite dimensional parameter vector in a natural way.

- **Remarks**
  - This does not mean that nonparametric methods have no parameters!
  - Most nonparametric methods avoid to make assumptions about the parametric form of the underlying distributions (except some smoothness properties).
  - Nonparametric methods are often memory-based (but not necessarily)
  - Sometimes called “lazy learning”
  - There is a difference between nearest neighbor methods and approaches in the dual representation

- **Can be applied to**
  - density estimation
  - classification
  - regression
Nonparametric Density Estimation: Histograms

- Histograms, one of the simplest nonparametric density estimator
Motivating Nonparametric Density Estimation

- Probability that a data point \( x \) comes from a region \( R \)
  \[
P = \int_R p(x) \, dx
  \]
- Probability that \( K \) out of \( N \) data points from \( p(x) \) will fall into \( R \) is binomial
  \[
  \Pr(K) = \frac{N!}{K!(N-K)!} P^K (1 - P)^{N-K}
  \]
- The expected value for \( K \) is:
  \[
  E\{K\} = NP
  \]
  \[
  Var\{K\} = NP[1 - P]
  \]
  or
  \[
  E\left\{ \frac{K}{N} \right\} = P
  \]
  \[
  Var\left\{ \frac{K}{N} \right\} = \frac{P}{N}[1 - P]
  \]
Motivating Nonparametric Density Estimation (cont’d)

- If \( p(\mathbf{x}) \) does not change significantly in the region \( R \), we can approximate:

\[
P = \int_{R} p(\mathbf{x}) d\mathbf{x} = p(\mathbf{x}) \int_{R} d\mathbf{x} = p(\mathbf{x}) V
\]

- Thus, the probability density can be finally approximated by:

\[
p(\mathbf{x}) \approx \frac{K}{NV}
\]

- Thus, two assumptions entered the derivation:
  - \( R \) is large enough to estimate \( P \) properly
  - \( R \) is small enough such that \( p(\mathbf{x}) \) is approx. constant

- Obviously, the motivates two approaches to density estimation
  - \( K=\text{const} \Rightarrow \text{find} V: \text{k-nearest neighbors} \)
  - \( V=\text{const} \Rightarrow \text{find} K: \text{kernel-based methods} \)
Kernel-based Methods

- E.g., define volume $V$ as a hypercube, centered at a point $x$
  \[
  V = h^d
  \]
- Define a kernel function $H(u)$:
  \[
  H(u) = \begin{cases} 
  1 & \text{if } |u_j| < 0.5 \text{ where } j = 1, \ldots, d \\
  0 & \text{otherwise}
  \end{cases}
  \]
  “Parzen Window”
- Such that the number of points in $V$ becomes:
  \[
  K = \sum_{n=1}^{N} H(u) = \sum_{n=1}^{N} H\left(\frac{x^n - x}{h}\right)
  \]
- The density estimate becomes:
  \[
  p(x) = \frac{K}{NV} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^d} H\left(\frac{x^n - x}{h}\right)
  \]
Example: Box-Cart Kernel
What is the right kernel?

- Criteria
  - spatially localized
  - smoothness (in particular when going to zero)
  - symmetry
  - efficiency (convergence to the true \( p(x) \))
  - “width” of the kernel

For convenience: \( \int H(u) du = 1 \)
Example: Gaussian Kernel

$h$ acts like a smoothing parameter

$$p(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{d/2}} \exp \left( -\frac{(x - x^n)^T (x - x^n)}{2h^2} \right)$$
Nearest-Neighbor Methods

- Kernel-based methods with fixed $h$ are not ideal since in areas with low density, the estimate becomes quite noisy.
  => make $h$ a function of $x$, i.e., $h=h(x)$
- $k$-nearest neighbor methods prescribe the number of points $K$ in the kernel and then determine the volume which is covered by this data.

$$p(x) \approx \frac{K}{NV}$$
Example: k-NN with Box-Cart Kernel
Example: k-NN with Gaussian Kernel
The Major Open Parameter is the Distance Metric

- in 1D data, it is only a smoothing parameter
- in multi-dimensional data, it becomes a matrix (diagonal or full)
- more about distance metrics in next class
Nonparametric Methods for Classification

- Assume data is labeled with $i=1,\ldots,c$, then a joint probability can be defined as:
  \[
p(x,i) = \frac{k_i}{NV}
  \]

- Thus, the posterior probability is:
  \[
P(i \mid x) = \frac{p(x,i)}{\sum_{j=1}^{c} p(x,j)} = \frac{K_i}{K}
  \]

Note that for sufficiently many data points, the nearest neighbor classifier is almost as good as a Bayes optimal classifier, and it is very easy to obtain!
Example

Note that this is a classifier that can generate very complex, nonlinear decision boundaries!
Example
KD-Trees

- kd-trees are a data structure built from a recursive partitioning of the data
- the kd-tree provides fast access to data and allows to develop learning algorithms on top of it
Summary

- Nonparametric methods based on neighborhood relationships offer a very simple set of methods for density estimation and classification (and regression).
- Most methods are memory based
- The distance metric is the key parameter that controls the quality of result (cross validation?)
- For large amounts of data, need special data structures
- Local learning:
  - Pros: low interference, very simple
  - Cons: usually more variance in estimation, potential global sub-optimality