10: Vector Data: Density Estimation

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

November 1, 2017
## Methods Learnt: Last Lecture

<table>
<thead>
<tr>
<th></th>
<th>Vector Data</th>
<th>Set Data</th>
<th>Sequence Data</th>
<th>Text Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td>Logistic Regression; Decision Tree; KNN SVM; NN</td>
<td></td>
<td></td>
<td>Naïve Bayes for Text</td>
</tr>
<tr>
<td><strong>Clustering</strong></td>
<td>K-means; hierarchical clustering; DBSCAN; DBSCAN; Mixture Models</td>
<td></td>
<td></td>
<td>PLSA</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>Linear Regression GLM*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frequent Pattern Mining</strong></td>
<td>Apriori; FP growth</td>
<td>GSP; PrefixSpan</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Search</strong></td>
<td></td>
<td></td>
<td></td>
<td>DTW</td>
</tr>
</tbody>
</table>
Vector Data: Density Estimation

• Introduction

• Nonparametric Density Estimation

• Parametric Density Estimation

• Summary
Vector Data: Density Estimation

- Introduction

- Nonparametric Density Estimation

- Parametric Density Estimation

- Summary
Density Estimation from Data

- **Goal**
  - Estimate density function for a random variable from data
- **Can be considered as an extension of histogram**
  - Smoothed version
Recall

- Density-based clustering can be viewed as identifying connected dense areas of a distribution
- Critical for many other mining functions
  - Classification
  - Outlier detection
Nonparametric vs. parametric methods

- **Nonparametric methods**
  - No assumptions of the forms of the underlying densities
  - Can be used with arbitrary distributions

- **Parametric methods**
  - Have assumptions of the forms of the underlying densities
  - The densities are determined by fixed but unknown parameters
Vector Data: Density Estimation

• Introduction

• Nonparametric Density Estimation

• Parametric Density Estimation

• Summary
Kernel Density Estimation

• Given a dataset \( D = (x_1, x_2, ..., x_n) \), estimate its density function \( f(x) \)

• Kernel density estimator:

\[
\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)
\]

• \( h \): bandwidth, controlling the smoothness of \( f \)

• \( K \): a non-negative real-valued integrable function, serving as weighting function

• \( \int_{-\infty}^{+\infty} K(u)du = 1 \) (normalization)

• \( K(u) = K(-u) \) for all \( u \) (symmetric)
Examples of Kernels

- Uniform
- Triangle
- Epanechnikov
- Quartic
- Triweight
- Gaussian
- Cosine
Gaussian Kernel in 1-D case

- Example: Gaussian kernel

\[ K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \]

- Scaled kernel

\[ K_h(u) = \frac{1}{h} K \left( \frac{u}{h} \right) \]

- In the Gaussian kernel case: \( K_h(u) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{u^2}{2h^2}} \)
Influence from one data point

- The influence of $x_i$ to $x$ can be considered as a weighting function centered at $x_i$

$$K_h(x - x_i) = \frac{1}{h \sqrt{2\pi}} e^{\frac{(x-x_i)^2}{2h^2}}$$

Recall: $\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$
Influence from multiple data points

- Aggregate influence from multiple data points to $x$

![Histogram](image1.png)  
**histogram**

![Density Functions](image2.png)  
Density: Each red curve indicates $\frac{1}{n}K_h(x - x_i)$
Is it a density function?

\[ \hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) \]

• A density function has to integrate to 1

\[ K_h(x - x_i) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}} \] integrates to 1

• Therefore, its average does so!
Impact of bandwidth

higher h, smoother density function
DENCLUE: Using Statistical Density Functions for Clustering

- DENsity-based CLUstEring by Hinneburg & Keim (KDD’98)
- Using statistical density functions:

\[ f_{\text{Gaussian}}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}} \]

\[ f^D_{\text{Gaussian}}(x) = \sum_{i=1}^{N} e^{-\frac{d(x, x_i)^2}{2\sigma^2}} \]

\[ \nabla f^D_{\text{Gaussian}}(x, x_i) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}} \]

- Major features
  - Solid mathematical foundation
  - Good for data sets with large amounts of noise
  - Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
  - Significant faster than existing algorithm (e.g., DBSCAN)
  - But needs a large number of parameters
Overall density of the data space can be calculated as the sum of the influence function of all data points.

- **Influence function**: describes the impact of a data point within its neighborhood.

Clusters can be determined mathematically by identifying density attractors.

- **Density attractors** are local maximal of the overall density function.
- **Center defined clusters**: assign to each density attractor the points density attracted to it.
- **Arbitrary shaped cluster**: merge density attractors that are connected through paths of high density (> threshold).
Density Attractor

Can be detected by hill-climbing procedure of finding local maximums
*Noise Threshold*

- Noise Threshold $\xi$
  - Avoid trivial local maximum points
  - A point can be a density attractor only if $\hat{f}(x) \geq \xi$
*Center-Defined and Arbitrary

Figure 3: Example of Center-Defined Clusters for different $\sigma$

Figure 4: Example of Arbitray-Shape Clusters for different $\xi$
Vector Data: Density Estimation

• Introduction

• Nonparametric Density Estimation

• Parametric Density Estimation

• Summary
Maximum-Likelihood Estimation

- Data: \( D = (x_1, x_2, ..., x_n) \)
- Parameters: \( \theta \)
- Model: \( p(x|\theta) \)
- Likelihood of \( \theta \) with respective to a set of data samples

\[
L(\theta; D) = p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta)
\]

- Maximum likelihood principle: find \( \hat{\theta} \) that maximizes \( L \)
  - Agrees the most with the observation of current dataset
Log-likelihood function

- log-likelihood function

\[
l(\theta) \equiv \ln L(\theta) = \ln p(D|\theta) = \sum_i \ln p(x_i|\theta)
\]

- Maximize likelihood function is equivalent to maximize log-likelihood function

\[
\hat{\theta} = \arg \max_\theta l(\theta)
\]

\[
\Rightarrow \nabla_\theta l(\theta) = 0
\]

\[
\nabla \theta \equiv \begin{bmatrix}
\frac{\partial}{\partial \theta_1} \\
\vdots \\
\frac{\partial}{\partial \theta_p}
\end{bmatrix}
\]
The Gaussian Case: Unknown Mean

• Consider 1-d Gaussian Distribution

\[ x_i \sim N(\mu, \sigma^2) \]

where \( \sigma^2 \) is known, i.e., \( \theta = \mu \)

\[ p(x_i | \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \]

• The log-likelihood is then

\[ l(\mu) = \sum_i \ln p(x_i | \mu) = \sum_i \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \]

• The MLE estimator for \( \mu \) is then

\[ \nabla_{\mu} l(\mu) = 0 \Rightarrow \sum_i (x_i - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_i x_i \]
The Gaussian Case: Unknown Mean and Variance

- Consider 1-d Gaussian Distribution
  \[ x_i \sim N(\mu, \sigma^2) \]
  where both \( \mu \) and \( \sigma^2 \) are unknown, i.e., \( \Theta = (\mu, \sigma^2) \)

  \[ p(x_i | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- The log-likelihood is then

  \[ l(\mu, \sigma^2) = \sum_i \ln p(x_i | \mu, \sigma^2) = \sum_i \left( -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \]

- The MLE estimators for \( \mu \) and \( \sigma^2 \) are then

  - \[ \frac{\partial l(\mu, \sigma^2)}{\partial \mu} = 0 \Rightarrow \sum_i (x_i - \hat{\mu})/\sigma^2 = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_i x_i \]
  - \[ \frac{\partial l(\mu, \sigma^2)}{\partial \sigma^2} = 0 \Rightarrow \sum_i \left( -\frac{1}{2\sigma^2} + \frac{(x_i - \hat{\mu})^2}{2(\sigma^2)^2} \right) = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2 \]

  Note it is biased
Vector Data: Density Estimation

• Introduction

• Nonparametric Density Estimation

• Parametric Density Estimation

• Summary
Summary

- Nonparametric Density Estimation
  - Kernel density estimation
- Parametric Density Estimation
  - Maximum likelihood estimation