MATH 829: Introduction to Data Mining and Analysis
Kernel density estimation and classification

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

March 23, 2016

Using Bayes theorem for classification

- Suppose we have observations $X \in \mathbb{R}^{n \times p}$ and $Y \in \{1, \ldots, K\}^n$ obtained at random.

Using Bayes theorem for classification

- Suppose we have observations $X \in \mathbb{R}^{n \times p}$ and $Y \in \{1, \ldots, K\}^n$ obtained at random.
- Before, we built classification models based on $P(Y = i | X = x)$, i.e., based on the conditional probability of $Y = i$ given $X = x$.

Using Bayes theorem for classification

- Suppose we have observations $X \in \mathbb{R}^{n\times p}$ and $Y \in \{1, \ldots, K\}^n$ obtained at random.
- Before, we built classification models based on $P(Y = i|X = x)$, i.e., based on the conditional probability of $Y = i$ given $X = x$.
- Using Bayes' rule, we can obtain $P(Y = i|X = x)$ from $P(X = x|Y = i)$ and $P(Y = i)$:

$$\begin{aligned} P(Y = i|X = x) &= \frac{P(X = x|Y = i)P(Y = i)}{P(X = x)} \\ &= \frac{P(X = x|Y = i)P(Y = i)}{\sum_{j=1}^{K} P(X = x|Y = j)P(Y = j)} \\ &\approx \frac{P(X = x|Y = i)\hat{\pi}_i}{\sum_{j=1}^{K} P(X = x|Y = j)\hat{\pi}_j}, \end{aligned}$$

where $\hat{\pi}_j =$ proportion of observations in category j.

Using Bayes theorem for classification

- Suppose we have observations $X \in \mathbb{R}^{n \times p}$ and $Y \in \{1, \dots, K\}^n$ obtained at random.
- Before, we built classification models based on $P(Y = i | X = x)$, i.e., based on the conditional probability of $Y = i$ given $X = x$.
- Using Bayes' rule, we can obtain $P(Y = i | X = x)$ from $P(X = x | Y = i)$ and $P(Y = i)$:

$$
\begin{aligned}
P(Y = i | X = x) &= \frac{P(X = x | Y = i) P(Y = i)}{P(X = x)} \\
&= \frac{P(X = x | Y = i) P(Y = i)}{\sum_{j=1}^{K} P(X = x | Y = j) P(Y = j)} \\
&\approx \frac{P(X = x | Y = i) \hat{\pi}_i}{\sum_{j=1}^{K} P(X = x | Y = j) \hat{\pi}_j},
\end{aligned}
$$

where $\hat{\pi}_j$ = proportion of observations in category j.

Question: How can we estimate the density of a distribution? (e.g. $P(X = x | Y = j) \dots$)

Density estimation

- More generally, suppose $x_1, \ldots, x_n$ is a random sample drawn from a probability density $f_X(x)$.

Density estimation

- More generally, suppose $x_1, \ldots, x_n$ is a random sample drawn from a probability density $f_X(x)$.
- The *nonparametric density estimation* (NPDE) problem is to estimate f_X without specifying a formal parametric structure.

Density estimation

- More generally, suppose $x_1, \ldots, x_n$ is a random sample drawn from a probability density $f_X(x)$.

- The *nonparametric density estimation* (NPDE) problem is to estimate f_X without specifying a formal parametric structure.
- A *bona fide* estimator of the density of a continuous random vector $X \in \mathbb{R}^p$ is a function $f : \mathbb{R}^p \to [0, \infty)$ such that

$$\int_{\mathbb{R}^p} f(x)\, dx = 1.$$

Density estimation

- More generally, suppose $x_1, \ldots, x_n$ is a random sample drawn from a probability density $f_X(x)$.

- The *nonparametric density estimation* (NPDE) problem is to estimate f_X without specifying a formal parametric structure.
- A *bona fide* estimator of the density of a continuous random vector $X \in \mathbb{R}^p$ is a function $f : \mathbb{R}^p \to [0, \infty)$ such that

$$\int_{\mathbb{R}^p} f(x)\, dx = 1.$$

- Example: Histogram estimation of the density

$$\hat{f}_X(x_0) = \frac{\#\{i : x_i \in N_\lambda(x_0)\}}{n\lambda},$$

where $N_\lambda(x_0)$ denotes a neighborhood of x_0 of width λ.

- More generally, suppose $x_1, \ldots, x_n$ is a random sample drawn from a probability density $f_X(x)$.

- The *nonparametric density estimation* (NPDE) problem is to estimate f_X without specifying a formal parametric structure.
- A *bona fide* estimator of the density of a continuous random vector $X \in \mathbb{R}^p$ is a function $f : \mathbb{R}^p \to [0, \infty)$ such that

$$\int_{\mathbb{R}^p} f(x)\, dx = 1.$$

- Example: Histogram estimation of the density

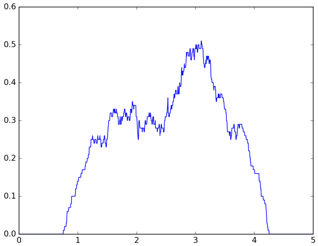
$$\hat{f}_X(x_0) = \frac{\#\{i : x_i \in N_\lambda(x_0)\}}{n\lambda},$$

where $N_\lambda(x_0)$ denotes a neighborhood of x_0 of width λ.

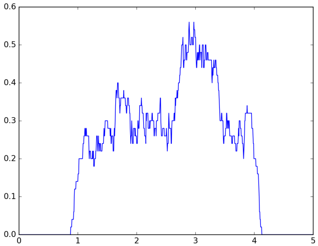
Exercise: Verify that $\hat{f}_X(x_0)$ is a *bona fide* estimator.

Example

```
import numpy as np

N = 200

X = 1+3*np.random.rand(N)

def density_hist(x, l, X):
    nb = ((X >= x-l/2.0)
         & (X <= x+l/2.0)).sum()
    n = X.shape[0]
    y = nb/(n*l)
    return y

nb_pts = 1000
x = np.linspace(0,5,nb_pts)
y = np.zeros(nb_pts)

l = 0.25

for i in range(nb_pts):
    y[i] = density_hist(x[i],l,X)

import matplotlib.pyplot as plt

plt.plot(x,y)
plt.show()
```



$\lambda = 0.5$



$\lambda = 0.25$

Kernel density

- We generally prefer to use a *smooth* estimate of the density:

$$\hat{f}_X(x_0) = \frac{1}{C}\sum_{i=1}^{n} K_\lambda(x_0, x_i),$$

where $K_\lambda(\cdot, \cdot)$ is some kernel, and C is a normalization constant.

Kernel density

- We generally prefer to use a *smooth* estimate of the density:

$$\hat{f}_X(x_0) = \frac{1}{C} \sum_{i=1}^{n} K_\lambda(x_0, x_i),$$

where $K_\lambda(\cdot, \cdot)$ is some kernel, and C is a normalization constant.

- A popular choice for K_λ is the *Gaussian kernel*:

$$K_\lambda(x_0, x) = \phi\left(\frac{|x - x_0|}{\lambda}\right) \qquad (\lambda > 0),$$

where $\phi(x) = \frac{1}{2\pi} e^{-x^2/2}$ is the $N(0, 1)$ density. In that case,

$$\hat{f}_X(x_0) = \frac{1}{n\lambda} \sum_{i=1}^{n} K_\lambda(x_0, x_i).$$

Common kernels

Kernel Function	$K(x)$
Rectangular	$\frac{1}{2}I_{[\|x\|\leq 1]}$
Triangular	$(1-\|x\|)I_{[\|x\|\leq 1]}$
Bartlett–Epanechnikov	$\frac{3}{4}(1-x^2)I_{[\|x\|\leq 1]}$
Biweight	$\frac{15}{16}(1-x^2)^2I_{[\|x\|\leq 1]}$
Triweight	$\frac{35}{32}(1-x^2)^3I_{[\|x\|\leq 1]}$
Cosine	$\frac{\pi}{4}\cos(\frac{\pi}{2}x)I_{[\|x\|\leq 1]}$

Source: Izenman, *Modern multivariate statistical techniques*.

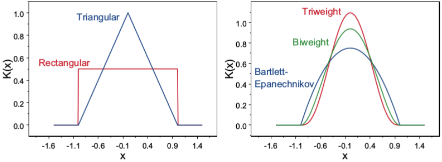
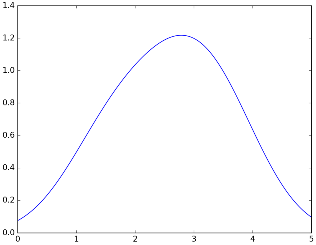


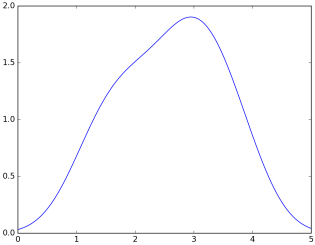
FIGURE 4.5. *Univariate kernel functions with compact support. Left panel: rectangular and triangular kernels. Right panel: Bartlett–Epanechnikov, biweight, and triweight kernels.*

Example

```
def density_gauss(x, l, X):
    n = np.double(X.shape)
    y = np.exp(-1*(x-X)**2/(2*l)).sum()/(n*l)
    return y
```



$\lambda = 0.5$



$\lambda = 0.25$

Example

Application: comparing data from two independent samples (Izenman, 2013).

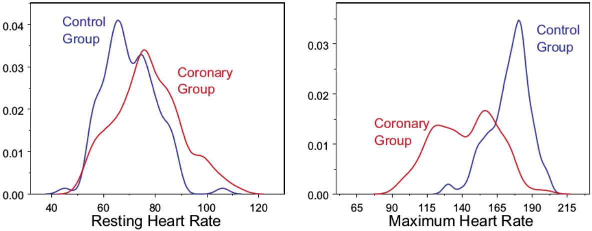
- 117 coronary heart disease patients (the *coronary group*).
- 117 age-matched healthy men (the *control group*).
- Heart rates recorded at rest and at their maximum after a series of exercises.
- A statistic used to monitor activity of the heart is the change in heart rate from a resting state to that after exercise.

Example

Application: comparing data from two independent samples (Izenman, 2013).

- 117 coronary heart disease patients (the *coronary group*).
- 117 age-matched healthy men (the *control group*).
- Heart rates recorded at rest and at their maximum after a series of exercises.
- A statistic used to monitor activity of the heart is the change in heart rate from a resting state to that after exercise.

Kernel density estimate:



Example: 1872 Hidalgo Postage Stamps of Mexico

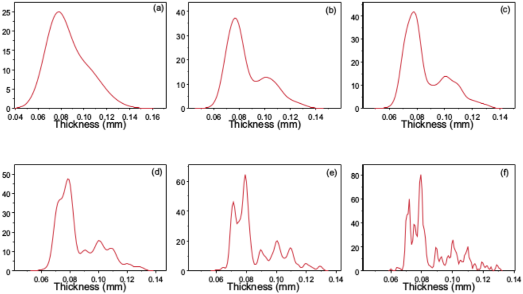
Example: (Izenman, 2013).

- 485 measurements of the thickness of the paper on which the 1872 Hidalgo Issue postage stamps of Mexico were printed.
- Stamps were deliberately printed on a mixture of paper types, each having its own thickness characteristics due to poor quality control in paper manufacture.
- Today, the thickness of the paper on which this particular stamp image is printed is a primary factor in determining its price.

Example: 1872 Hidalgo Postage Stamps of Mexico

Example: (Izenman, 2013).

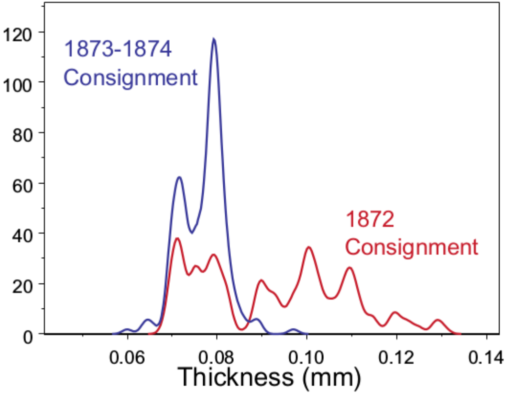
- 485 measurements of the thickness of the paper on which the 1872 Hidalgo Issue postage stamps of Mexico were printed.
- Stamps were deliberately printed on a mixture of paper types, each having its own thickness characteristics due to poor quality control in paper manufacture.
- Today, the thickness of the paper on which this particular stamp image is printed is a primary factor in determining its price.



Gaussian density estimate for different window sizes. (Source: Izenman, 2013)

Example: 1872 Hidalgo Postage Stamps of Mexico (cont.)

Every stamp from the 1872 Hidalgo Issue was overprinted with year-of-consignment information: there was an 1872 consignment (289 stamps) and an 1873-1874 consignment (196 stamps).



Gaussian density estimate for each consignment, window size = 0.0015. (Source: Izenman, 2013)

Multivariate generalization

- The previous ideas naturally generalize to multivariate data.

- The previous ideas naturally generalize to multivariate data.
- Given $x_1, \dots, x_n \in \mathbb{R}^p$, $x_0 \in \mathbb{R}^p$, and an invertible matrix H, we can use

$$\hat{f}_H(x_0) = \frac{1}{n \cdot \det H} \sum_{i=1}^{n} K(H^{-1}(x_0 - x_i))$$

Multivariate generalization

- The previous ideas naturally generalize to multivariate data.
- Given $x_1, \ldots, x_n \in \mathbb{R}^p$, $x_0 \in \mathbb{R}^p$, and an invertible matrix H, we can use

$$\hat{f}_H(x_0) = \frac{1}{n \cdot \det H} \sum_{i=1}^{n} K(H^{-1}(x_0 - x_i))$$

- Multiplicative kernels:

$$K(x) \propto f(x_1) f(x_2) \ldots f(x_p)$$

- Spherical kernels:

$$K(x) \propto f(\|x\|).$$

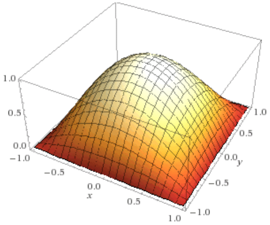
Recall that the *Epanechnikov* kernel is given by

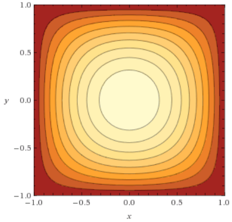
$$K_\lambda(x, x') = D\left(\frac{|x - x'|}{\lambda}\right),$$

where

$$D(t) := \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Multiplicative 2D version:





Examples

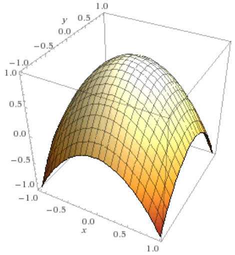
Recall that the *Epanechnikov* kernel is given by

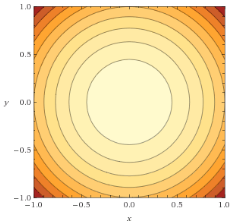
$$K_\lambda(x, x') = D\left(\frac{|x - x'|}{\lambda}\right),$$

where

$$D(t) := \begin{cases} \frac{3}{4}(1 - t^2) & \text{if } |t| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Spherical 2D version:



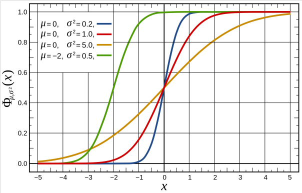


Statistical properties of density estimator

The empirical cdf: Let X be a (one-dimensional) random variable.

Recall that the cumulative distribution function (cdf) of X is

$$F_X(x) = P(X \leq x).$$

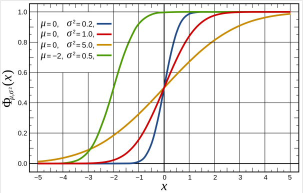


Normal cdf (source: wikipedia).

The empirical cdf: Let X be a (one-dimensional) random variable.

Recall that the cumulative distribution function (cdf) of X is

$$F_X(x) = P(X \leq x).$$



Normal cdf (source: wikipedia).

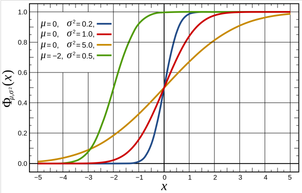
The empirical cdf of a sample $x_1, \ldots, x_n$ drawn from X is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}(x_i).$$

The empirical cdf: Let X be a (one-dimensional) random variable.

Recall that the cumulative distribution function (cdf) of X is

$$F_X(x) = P(X \leq x).$$



Normal cdf (source: wikipedia).

The empirical cdf of a sample $x_1, \ldots, x_n$ drawn from X is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(x_i).$$

Theorem: (Glivenko–Cantelli) Let $X_1, \ldots, X_n$ be iid random variables with cdf F. Let

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(X_i).$$

Then

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \qquad \text{almost surely.}$$

Statistical properties of density estimator (cont.)

- The Glivenko–Cantelli theorem shows that cdfs can be recovered consistently using the empirical cdf.

Statistical properties of density estimator (cont.)

- The Glivenko–Cantelli theorem shows that cdfs can be recovered consistently using the empirical cdf.
- Unfortunately, the empirical cdf does not provide a good estimate of the pdf (puts a probability 0 between two observations).

Statistical properties of density estimator (cont.)

- The Glivenko–Cantelli theorem shows that cdfs can be recovered consistently using the empirical cdf.
- Unfortunately, the empirical cdf does not provide a good estimate of the pdf (puts a probability 0 between two observations).

Desirable properties of a density estimator: Let $\hat{f}_n(x)$ be an estimator obtained from an iid sample with density $f(x)$, $x \in \mathbb{R}^p$. (Note: $\hat{f}_n(x)$ is a random variable.)

1. Unbiasedness: $E(\hat{f}_n(x)) = f(x)$ for all $x \in \mathbb{R}^p$.

Statistical properties of density estimator (cont.)

- The Glivenko–Cantelli theorem shows that cdfs can be recovered consistently using the empirical cdf.
- Unfortunately, the empirical cdf does not provide a good estimate of the pdf (puts a probability 0 between two observations).

Desirable properties of a density estimator: Let $\hat{f}_n(x)$ be an estimator obtained from an iid sample with density $f(x)$, $x \in \mathbb{R}^p$. (Note: $\hat{f}_n(x)$ is a random variable.)

1. Unbiasedness: $E(\hat{f}_n(x)) = f(x)$ for all $x \in \mathbb{R}^p$.
   It is known that **no** bona fide density estimator based upon a finite data set that is unbiased for all continuous densities can exist (Rosenblatt, 1956). As a result, people look at *asymptotic unbiasedness*.

Statistical properties of density estimator (cont.)

- The Glivenko–Cantelli theorem shows that cdfs can be recovered consistently using the empirical cdf.
- Unfortunately, the empirical cdf does not provide a good estimate of the pdf (puts a probability 0 between two observations).

Desirable properties of a density estimator: Let $\hat{f}_n(x)$ be an estimator obtained from an iid sample with density $f(x)$, $x \in \mathbb{R}^p$. (Note: $\hat{f}_n(x)$ is a random variable.)

1. Unbiasedness: $E(\hat{f}_n(x)) = f(x)$ for all $x \in \mathbb{R}^p$.
   It is known that **no** bona fide density estimator based upon a finite data set that is unbiased for all continuous densities can exist (Rosenblatt, 1956). As a result, people look at *asymptotic unbiasedness*.
2. Consistency: Ability to recover f as $n \to \infty$. How to measure "closeness" between the densities?

Consistency

Important notions of consistency:

1. Strong pointwise consistency: $\hat{f}_n(x) \to f(x)$ almost surely $\forall x \in \mathbb{R}^p$ as $n \to \infty$.
2. Pointwise consistency of f in *quadratic mean*:

$$\text{MSE}(x) = E\left((\hat{f}_n(x) - f(x)^2)\right) \to 0 \quad \forall x \in \mathbb{R}^p \text{ as } n \to \infty.$$

3. Consistency of f in *mean integrated squared error* (MISE):

$$\text{MISE} = E\left(\int_{\mathbb{R}^p} (\hat{f}_n(x) - f(x))^2 \, dx\right) \to 0 \quad \text{ as } n \to \infty.$$

4. Consistency of f in *mean integrated absolute error* (MIAE):

$$\text{MIAE} = E\left(\int_{\mathbb{R}^p} |\hat{f}_n(x) - f(x)| \, dx\right) \to 0 \quad \text{ as } n \to \infty.$$

Consistency

Important notions of consistency:

1. Strong pointwise consistency: $\hat{f}_n(x) \to f(x)$ almost surely $\forall x \in \mathbb{R}^p$ as $n \to \infty$.
2. Pointwise consistency of f in *quadratic mean*:

$$\mathrm{MSE}(x) = E\left((\hat{f}_n(x) - f(x)^2)\right) \to 0 \quad \forall x \in \mathbb{R}^p \text{ as } n \to \infty.$$

3. Consistency of f in *mean integrated squared error* (MISE):

$$\mathrm{MISE} = E\left(\int_{\mathbb{R}^p} (\hat{f}_n(x) - f(x))^2 \, dx\right) \to 0 \quad \text{as } n \to \infty.$$

4. Consistency of f in *mean integrated absolute error* (MIAE):

$$\mathrm{MIAE} = E\left(\int_{\mathbb{R}^p} |\hat{f}_n(x) - f(x)| \, dx\right) \to 0 \quad \text{as } n \to \infty.$$

- Many other norms are used (e.g. Hellinger distance, etc.).

Asymptotic results for kernels

Suppose we use a kernel coming for a multivariate probability density function $K : \mathbb{R}^p \to [0, \infty)$:

$$\int_{\mathbb{R}^p} K(x)\ dx = 1.$$

Asymptotic results for kernels

Suppose we use a kernel coming for a multivariate probability density function $K : \mathbb{R}^p \to [0, \infty)$:

$$\int_{\mathbb{R}^p} K(x)\ dx = 1.$$

In other words, we define:

$$K_\lambda(x, y) := K\left(\frac{x-y}{\lambda}\right), \qquad x, y \in \mathbb{R}^p, \lambda > 0.$$

(e.g. Gaussian kernel).

Asymptotic results for kernels

Suppose we use a kernel coming for a multivariate probability density function $K : \mathbb{R}^p \to [0, \infty)$:

$$\int_{\mathbb{R}^p} K(x)\ dx = 1.$$

In other words, we define:

$$K_\lambda(x, y) := K\left(\frac{x-y}{\lambda}\right), \qquad x, y \in \mathbb{R}^p, \lambda > 0.$$

(e.g. Gaussian kernel).
A remarkable result in density estimation is that the density estimator from this class of kernels is **always** consistent.

Asymptotic results for kernels

Suppose we use a kernel coming for a multivariate probability density function $K : \mathbb{R}^p \to [0, \infty)$:

$$\int_{\mathbb{R}^p} K(x)\ dx = 1.$$

In other words, we define:

$$K_\lambda(x, y) := K\left(\frac{x-y}{\lambda}\right), \qquad x, y \in \mathbb{R}^p, \lambda > 0.$$

(e.g. Gaussian kernel).

A remarkable result in density estimation is that the density estimator from this class of kernels is **always** consistent.

Theorem:(Devroye, 1983; Devroye and Penrod, 1984)

Let $\hat{f}_n$ be a kernel estimator as above with window size λ_n, obtained from an iid sample of size n. Suppose $\lambda_n \to 0$ and $n\lambda_n \to \infty$. Then

1. $\hat{f}_n$ is pointwise strongly consistent.
2. Moreover, in the univariate case, $\mathrm{MIAE} = O(n^{-2/5})$.

Explicit formulas for the asymptotically optimal window size λ_n are also known.