ASDS Statistics, YSU, Fall 2020

Lecture 06

Michael Poghosyan

26 Sep 2020

Contents

Last Lecture Recap

- Give the Definition of the KDE.

KDE

It is easy to see that $KDE(x)$ will give a function satisfying the properties of the PDF, i.e., will be nonnegative and will integrate to 1:

$$\int_{-\infty}^{+\infty} KDE(x)dx =$$

KDE

It is easy to see that $KDE(x)$ will give a function satisfying the properties of the PDF, i.e., will be nonnegative and will integrate to 1:

$$\int_{-\infty}^{+\infty} KDE(x)dx = \frac{1}{nh} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_i}{h}\right) dx =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_i}{h}\right) d\frac{x - x_i}{h} \overset{u=\frac{x-x_i}{h}}{=}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K(u)du = \frac{1}{n} \cdot \sum_{i=1}^{n} 1 = 1.$$

KDE

Note: Like in the case of the Density histogram, where that histogram was depending on the bins choice, the KDE depends on the choice of $h > 0$. h is called the **bandwidth**, and its estimation is another story.

KDE

Note: Like in the case of the Density histogram, where that histogram was depending on the bins choice, the KDE depends on the choice of $h > 0$. h is called the **bandwidth**, and its estimation is another story.

Note: One can prove that under some conditions, KDE is approximating well the unknown PDF behind the data.

KDE

Note: Like in the case of the Density histogram, where that histogram was depending on the bins choice, the KDE depends on the choice of $h > 0$. h is called the **bandwidth**, and its estimation is another story.

Note: One can prove that under some conditions, KDE is approximating well the unknown PDF behind the data. In fact,

Theorem: Assume we are constructing the $KDE = KDE(\cdot, h_n)$ for the IID r.v $X_1, X_2, ..., X_n$, coming from an unknown PDF f, and with the bandwidth h_n.

KDE

Note: Like in the case of the Density histogram, where that histogram was depending on the bins choice, the KDE depends on the choice of $h > 0$. h is called the **bandwidth**, and its estimation is another story.

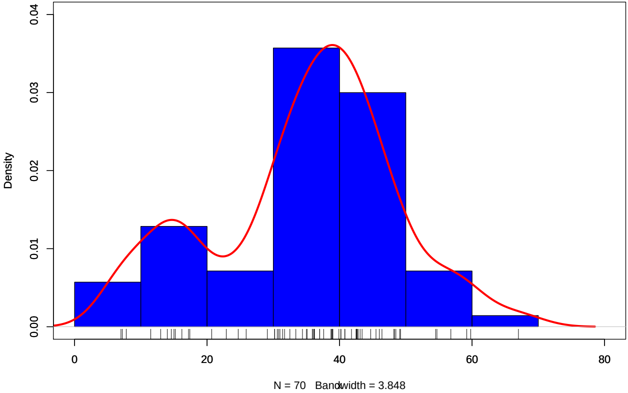
Note: One can prove that under some conditions, KDE is approximating well the unknown PDF behind the data. In fact,

Theorem: Assume we are constructing the $KDE = KDE(\cdot, h_n)$ for the IID r.v $X_1, X_2, ..., X_n$, coming from an unknown PDF f, and with the bandwidth h_n. If the PDF f is continuous at the point x, and if $h_n \to 0$ and $n \cdot h_n \to \infty$, then

$$KDE(x, h_n) \to f(x) \qquad in \quad \mathbb{P}.$$

KDE Example

```
x <- precip; d <- density(x)
hist(x, freq = FALSE, xlim = c(0, 80), ylim = c(0,0.04),
     col = "blue", main = "")
rug(x); par(new = TRUE)
plot(d, lwd = 3, col = "red", xlim = c(0,80), ylim = c(0,0.04),
     main = "")
```



KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$.

KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$. We assume that $\hat{f}$ depends on some parameter h (smoothing parameter), and we want to find some *best value* for h.

KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$. We assume that $\hat{f}$ depends on some parameter h (smoothing parameter), and we want to find some *best value* for h. We define the **Risk** of Estimating f through $\hat{f}$, $Risk(\hat{f}, f)$.

KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$. We assume that $\hat{f}$ depends on some parameter h (smoothing parameter), and we want to find some *best value* for h. We define the **Risk** of Estimating f through $\hat{f}$, $Risk(\hat{f}, f)$. One of the standard ways to define the Risk is to choose the Mean Integrated Squared Error:

$$Risk(\hat{f}, f) = MISE(\hat{f}, f) = \mathbb{E} \int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx.$$

KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$. We assume that $\hat{f}$ depends on some parameter h (smoothing parameter), and we want to find some *best value* for h. We define the **Risk** of Estimating f through $\hat{f}$, $Risk(\hat{f}, f)$. One of the standard ways to define the Risk is to choose the Mean Integrated Squared Error:

$$Risk(\hat{f}, f) = MISE(\hat{f}, f) = \mathbb{E} \int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx.$$

Now, the idea is to choose h in such a way that the Risk of $\hat{f}$ will be the minimal.

Visualizing 2D Data

In case we have a 2D numerical Dataset

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n),$$

we usually do the ScatterPlot - the plot of all points (x_i, y_i), $i = 1, ...n$.

Example

Example: Graph the ScatterPlot for the following data:

Person ID	Age	Weight
1	20	69
2	22	57
3	40	65
4	20	70

Example

Say, consider again the *cars* Dataset:

```r
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```r
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

Example

Say, consider again the *cars* Dataset:

```r
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

```r
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations.

Example

Say, consider again the *cars* Dataset:

```
head(cars, 3)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

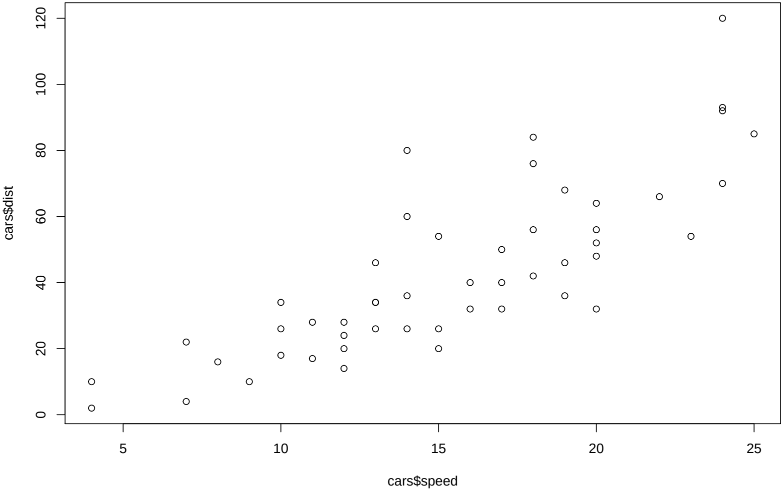
```
str(cars)
```

```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

It has 2 Variables: *Speed* and *Distance*, and 50 Observations. Let us do the ScatterPlot of Observations:

ScatterPlot

```
plot(cars$speed, cars$dist)
```



Notes

- In this graph you can see that there is some relationship between the *Speed* and *Distance*, there is a *trend*: if the speed gets larger, the (stopping) distance is tending to increase.

Notes

- In this graph you can see that there is some relationship between the *Speed* and *Distance*, there is a *trend*: if the speed gets larger, the (stopping) distance is tending to increase.
- Here we have visualized 2D Dataset of the (N, N) type, i.e., both Variables were N=Numerical.

Notes

- In this graph you can see that there is some relationship between the *Speed* and *Distance*, there is a *trend*: if the speed gets larger, the (stopping) distance is tending to increase.
- Here we have visualized 2D Dataset of the (N, N) type, i.e., both Variables were N=Numerical. Think about how we can visualize a 2D Dataset of the type, say (C=Categorical)

$$(N, C),\ (C, N),\ (C, C),\ (N, N, C),\ \ldots$$

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

- One can draw the 2D (N, N) in 3D, by constructing 2D Histograms and KDEs

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

- One can draw the 2D (N, N) in 3D, by constructing 2D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (if that 3rd Variable is Categorical or Discrete Numerical)

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

- One can draw the 2D (N, N) in 3D, by constructing 2D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (if that 3rd Variable is Categorical or Discrete Numerical)
- One can add the 4th Dimension by using the Size of Points

Additions: Multidimensional Graphs

The topic of Data Visualization is a very rich and interesting one. And there are various types of graphs to help visualize our data, see, e.g., https://datavizcatalogue.com/.

And some ideas for multidimensional Visualizations:

- One can draw the 2D (N, N) in 3D, by constructing 2D Histograms and KDEs
- One can draw 3D in 2D, using the 3rd variable as the Color (if that 3rd Variable is Categorical or Discrete Numerical)
- One can add the 4th Dimension by using the Size of Points
- And add the 5-th one by using the Shape of Points, . . .

Examples

See, for example, beautiful visualizations by **Hans Rosling**.

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Examples

See, for example, beautiful visualizations by **Hans Rosling**. Say, this short one: Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four

Or, the following one: Gender Gap in Earnings per University

Additions: Multidimensional Graphs

- One can do the Pairs Plot

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the Correlation Matrix HeatMap

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the Correlation Matrix HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data

Additions: Multidimensional Graphs

- One can do the Pairs Plot
- One can draw the Correlation Matrix HeatMap
- One can use a Dimensionality Reduction Methods to Visualize some high dimensional Data
- etc ...