ITDM - LEZIONE 06

DEL 22/10/2019

DATA MINING

The main objective of data mining is to extract "useful information" from DATA.

The data we want to analyze is structured in a DATASET, which can be represented with a DATA-MATRIX made up of n rows and d colums, in which each column represents a FEATURE and each row represents an INSTANCE.

A possible example of a dataset can be the following one:

	F_1	F_2	F_3
I_1	0.5	0.3	0.4
I_2	0.3	0.7	0.4
I_3	0.4	0.2	0.3
I_4	0.1	0.5	0.3

→ This row contains the characteristic of a single OBJECT

This column contain the values different objects have on the same characteristic.

We can thus think of a feature as a particular characteristic an object may or may not have.

OSS : To make a connection to traditional DB systems we have

FEATURES ⟷ FIELDS OF ENTITIES

INSTANCES ⟷ RECORDS OR TUPLES

We can have the following types of data in a dataset, which are:

- DATA
  - NUMERICAL
  - CATEGORICAL
    - ORDINAL
    - REGULAR

Numerical data comes from QUANTITATIVE MEASURES, like measuring the length of a petal; Categorical data comes from QUALITATIVE MEASURES, like observing the color of a pair of eyes.

To CATEGORICAL data you can apply the = operator; to ORDINAL data you can also apply the ≤ relation to order them.

ESTIMATION

To apply IT methods to datasets in order to solve certain tasks we first have to ESTIMATE the PMF (discrete case) or the PDF (continuous case).

To do this we have two families of estimation methods:

- PARAMETRIC METHODS
- NON-PARAMETRIC METHODS

PARAMETRIC METHODS

In these methods we assume that the data comes from a particular DISTRIBUTION, like a Gaussian $N(\mu, \sigma^2)$, and we try to estimate the various PARAMETERS.

For example, if we want to estimate the mean μ we can use:

- ARITHMETIC MEAN: $\hat{\mu} := \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$
- GEOMETRIC MEAN: $\hat{\mu} := \left(\prod_{i=1}^{N} x_i \right)^{\frac{1}{n}}$

OSS: These means are different from the THEORETICAL DISTRIBUTION MEAN, which is studied in PROBABILITY THEORY and is defined as $\mu = \mathbb{E}[X]$ and is computed using the PMF or PDF.

PARAMETRIC methods are more CONSTRAINED because we have to assume how the initial data is distributed.

NON-PARAMETRIC METHOD

Let us define the INDICATOR FUNCTION,

$$\mathbb{1}(A) := \begin{cases} 1, & \text{if } A \text{ is TRUE} \\ 0, & \text{if } A \text{ is FALSE} \end{cases}$$

We also need the RECTANGULAR WEIGHT FUNCTION, which is defined as

$$I(x) := \begin{cases} 1, & \text{if } |x| \leq 1 \\ 0, & \text{if } |x| > 0 \end{cases}$$

It can then be extended to VECTORS as follows,

$$I(\underline{x}) := \begin{cases} 1, & \text{if } |x_i| \leq 1 \quad \forall i \\ 0, & \text{else} \end{cases}$$

Let us now consider a discrete r.v. X with an unknown PMF $p(x_i)$. We want to ESTIMATE $p(x_i)$ using a SAMPLE.

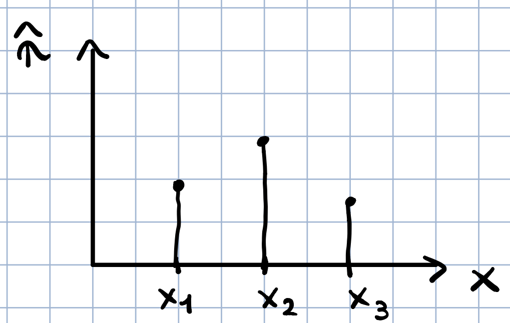
$$\mathcal{X} := \{ x_1, x_2, \dots, x_n \}$$

$$S := \{ s_1, s_2, \dots, s_N \}$$

To do this we can use the EMPIRICAL PMF ESTIMATOR, defined as

$$\hat{p}(X = x_i) := \frac{1}{n} \cdot \sum_{k=1}^{n} \mathbb{1}(s_k = x_i)$$

which comes down to computing the various frequencies in the sample. Graphically we have points



We can also extend this formula to the MULTI-VARIATE case as follows,

$$\hat{p}(\underline{x} = \underline{x}_i) = \frac{1}{n} \cdot \sum_{k=1}^{n} (\underline{s}_k = \underline{x}_i)$$

Consider now a continuous r.v. X with $f(x)$ as PDF. We want to estimate $f(x)$ using a sample $S = \{s_1, s_2, \dots, s_m\}$. To do this we have at least the following two methods:

- HISTOGRAM

The basic idea is to DISCRETIZE the real line $\mathbb{R}$ into a sequence of BINS. All the points that fall inside the same bins are merged together to obtain a single value.

Formally we have the following PARAMETERS

$$\begin{cases} h := \text{INTERVAL WIDTH} \\ x_0 := \text{ORIGIN} \end{cases}$$

The i-th Bin is then the following interval of the real line

$$B_i := [\, x_0 + i \cdot h,\ x_0 + (i+1) \cdot h \,] \ , \ i \in \mathbb{Z}$$

The estimation is then done as follows

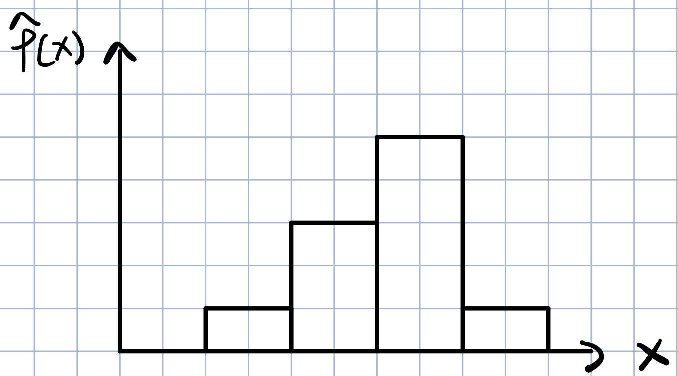
$$\hat{f}(x) := \frac{1}{m \cdot h} \cdot \sum_{k=1}^{m} \sum_{i=-\infty}^{+\infty} \mathbb{1}(x \in B_i) \cdot \mathbb{1}(s_k \in B_i)$$

Notice that $\int_{-\infty}^{+\infty} \hat{f}(x)\, dx = 1$ since:

i) The width of each RECTANGLE is h.

ii) The height of each rectangle is the # of samples in the rectangle divided by n.

Graphically we have the following



- <u>KERNEL FUNCTION METHOD</u> ~ (1:31:00 min)

<u>DEF</u>: K is a KERNEL FUNCTION if it has the following properties:

i) $\int_{-\infty}^{+\infty} K(x)\, dx = 1$

ii) $\int_{-\infty}^{+\infty} x \cdot K(x)\, dx = 0$

Notice that if X is a r.v. with $f(x)$ as PDF, and $\mathbb{E}[X] = 0$, then $f(x)$ is a KERNEL FUNCTION.

Let $S = \{s_1, s_2, \dots, s_n\}$, we can now define the KERNEL DENSITY ESTIMATOR as follows,

$$\hat{p}(x) := \frac{1}{n \cdot h} \sum_{K=i}^{n} K\left(\frac{x - s_K}{h}\right)$$

where h is called the BANDWIDTH.

Notice that $\hat{p}(x)$ is a SMOOTHER estimation than the one obtained by the HISTOGRAM METHOD. Also, the smaller the value of h is, the smoother the curve will be. However, if h is too small we can lose BIMODIAL DISTRIBUTIONS.

EXAMPLE (HISTOGRAM) :

Consider the following SAMPLE SET

$$S = \{ 0.15,\ 0.31,\ 0.28,\ 0.57,\ 0.27,\ 0.21,\ 0.83,\ 0.33,\ 0.42,\ 0.24,\ 0.31 \}$$

Let $h = 0.1$, $n = 11$.

B_i	# OF SAMPLE IN BIN i	$P(Y \in B_i)$
$[0, 0.1)$	0	0
$[0.1, 0.2)$	1	$\frac{1}{11 \cdot 0.1}$
$[0.2, 0.3)$	4	$\frac{4}{11 \cdot 0.1}$
$[0.3, 0.4)$	3	$\frac{3}{11 \cdot 0.1}$
$[0.4, 0.5)$	1	$\frac{1}{11 \cdot 0.1}$
$[0.5, 0.6)$	1	$\frac{1}{11 \cdot 0.1}$
$[0.6, 0.7)$	0	0
$[0.7, 0.8)$	0	0
$[0.8, 0.9)$	1	$\frac{1}{11 \cdot 0.1}$

Notice that: $0.1 \cdot \frac{1}{11 \cdot 0.1} + \dots + \ = 1$