Simple Model Comparison

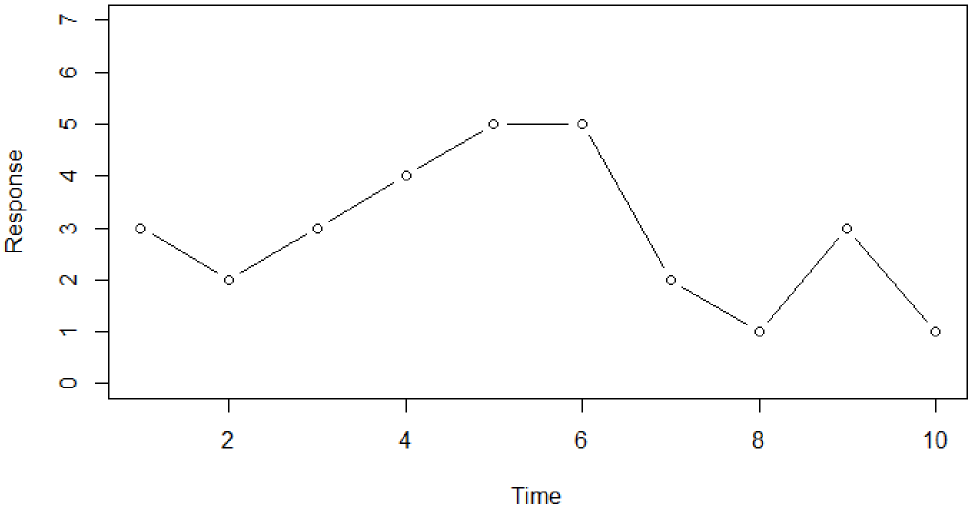
Daskalopoulos Kyriakos

05-02-2023

Suppose we are given the following two dimensional data-set:

$$(4,4),\ (10,1),\ (6,5),\ (1,3),\ (5,5),\ (3,3),\ (7,2),\ (8,1),\ (9,3),\ (2,2)$$

To examine the statistical properties of the data, i.e. the probability distribution behind this experiment, we start by calculating the basic statistical "measures" such as mean value, variance e.t.c.
The data can be looked as time-series data, since the variable X resembles time, so we order in ascending order the X values.
Here we consider x as time, so we order the data in ascending order of x.



Σχήμα 1: Plot of the order data.

There are two models that can be proposed for the random variable X.

$$\begin{cases} M_1 : X_i \sim F, i = 1,\dots,10 \\ M_2 : X_i \sim F_1, i = 1,\dots,6 \\ \quad X_i \sim F_2, i = 7,\dots,10 \end{cases}$$

One Distribution over time

In this case, the sample mean and variance are :

$$\hat{\mathbb{E}}(X) = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{29}{10} = 2.9$$

$$\hat{S}_n^2 = \frac{\sum_{i=1}^n (X_i - \hat{\mathbb{E}}(X))^2}{n-1} = \frac{}{9} = \frac{18.9}{9} = 2.1$$

So, we can conclude that X follows a distribution F with mean 2.9 and variance 2.1(or standard deviation $\sqrt{2.1} = 1.4491$).

The problem now consists of finding a distribution which fit best to the data and has the same mean and variance. This will be accesed later.

Change of Distribution

Here the distribution change at some $t = t_0$. The method for calculating the two distributions F_1, F_2 is the same as above.

Instead of estimating the time of change from the data[1], we can assume that it is equal to 6, which is justified as seen in the graph.

The same procedure for calculating the sample mean and variance as above is followed, so details are skipped:

1. Model 1$(1 \leq i \leq 6)$

$$\hat{\mathbb{E}}(X) = \frac{2 + 2 \times 3 + 4 + 2 \times 5}{6} = \frac{22}{6} = 3,666$$

$$\hat{S}_n^2 = \frac{2 \times (3 - 3.666)^2 + (2 - 3.666)^2 + (4 - 3.666)^2 + 2 \times (5 - 3.666)^2}{5}$$

$$= \frac{2 \times (0.666)^2 + (1.666)^2 + (0.334)^2 + 2 \times (1.334)^2}{5} = \frac{5.55378}{5} = 1.1107$$

2. Model 2 $(7 \leq i \leq 10)$

$$\hat{\mathbb{E}}(X) = \frac{1 \times 2 + 2 + 3}{4} = 1.75$$

$$\hat{S}_n^2 = \frac{(1 - 1.75)^2 + (2 - 1.75)^2 + (3 - 1.75)^2}{3} = \frac{2.75}{3} = 0.9166$$

Model Comparison

To calculate the models posterior probability, we consider that the two models have equal a-priori probability. So $\mathbb{P}(M_1) = \mathbb{P}(M_2) = \frac{1}{2}$.

According to Bayes,

[1] Which would be an interesting problem. A prososal is

$$\hat{t} = \sup_j \left(|X_j - X_{j-1}|^2 + |X_j - X_{j+1}|^2\right),\ j = 1, \ldots, n$$

This is the 1-neighbor mean distance. It can be generalized to k-neighbor by averaging over the k nearest data points.

$$\mathbb{P}(M_1|x) = \frac{\mathbb{P}(x|M_1)\mathbb{P}(M_1)}{\int_A \mathbb{P}(x|M)\mathbb{P}(M)} = \frac{\mathbb{P}(x|M_1)\mathbb{P}(M_1)}{\mathbb{P}(x|M_1)\mathbb{P}(M_1) + \mathbb{P}(x|M_2)\mathbb{P}(M_2)} = \frac{\mathbb{P}(x|M_1)}{\mathbb{P}(x|M_1) + \mathbb{P}(x|M_2)}$$

and similarly,

$$\mathbb{P}(M_2|x) = \frac{\mathbb{P}(x|M_2)}{\mathbb{P}(x|M_1) + \mathbb{P}(x|M_2)}$$

Now it is relatively easy to make the calculations.

1. $$\mathbb{P}(x|M_1) := \prod_{i=1}^{n} \mathbb{P}_{\mathbb{F}}(X = x_i)$$

2. $$\mathbb{P}(x|M_2) = \prod_{i=1}^{6} \mathbb{P}_{F_1}(X = x_i) \prod_{i=7}^{10} \mathbb{P}_{F_2}(X = x_i)$$

, where generally $\mathbb{P}_{\mathbb{D}}$ is the pdf function of the distribution D.

So when F, F_1, F_2 are known or estimated, the posteriori probabilities of the two models can be calculated and compared.

Example

We assume a normal distribution under Model 1 and two different discrete distributions for model 2. So $F \sim \mathcal{N}(2.9, 2.1)$, $F_1 \sim dU(2, 3, 4, 5)$ and $F_2 \sim dU(1, 2, 3)$.

1. $$\mathbb{P}(x|M_1) := \prod_{i=1}^{n} \mathbb{P}_{\mathbb{F}}(X = x_i) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sqrt{2.1}} \exp\left(\frac{(x_i - 2.9)^2}{2 \times 2.1}\right) = 0.0047$$

2. $$\mathbb{P}(x|M_2) = \prod_{i=1}^{6} \mathbb{P}_{F_1}(X = x_i) \prod_{i=7}^{10} \mathbb{P}_{F_2}(X = x_i) = \frac{1}{6^6}\frac{1}{3^4} = 2.64 \times 10^{-7}$$

So we have for the posterior probabilities:

1. $$\mathbb{P}(M_1|x) = \frac{\mathbb{P}(x|M_1)}{\mathbb{P}(x|M_1) + \mathbb{P}(x|M_2)} = \frac{0.0047}{0.0047 + 2.64 \times 10^{-7}} = 0.999$$

2. $$\mathbb{P}(M_2|x) = \frac{\mathbb{P}(x|M_2)}{\mathbb{P}(x|M_1) + \mathbb{P}(x|M_2)} = \frac{2.64 \times 10^{-7}}{0.0047 + 2.64 \times 10^{-7}} = 5.616 \times 10^{-5}$$

As another example, if we compare the models $M_1 : X \sim \mathcal{N}(2.9, 2.1)$ and $M_2 : X \sim Pois(2.9)$, we find $\mathbb{P}(M_1|x) = 0.609$.