Second Computer Homework of the course engineering probability and statistics

Dr. Mohammad Karbasi

Feel free to ask any question in telegram from the teaching assistant:
Mohammad Amin Mirzaii(@Mam1381)

1 Distribution estimation

1.1 Part 1

In this part we are going to estimate the distribution function of a random variable from a sample dataset which has been generated from that distribution. We have a dataset with 1000 samples which are in a excel file that has been given to you.
To estimate the distribution, we are going to use a method which is called parzen window. In this method, we will replace each sample by a guassian distribution with a known variance and the mean value of that sample data. At the end ,we will add these distributions and divide the whole distribution by the number of the samples. Plot the estimated distribution with values of $\sigma = 0.5$, $\sigma = 1$,$\sigma = 1.5$ and $\sigma = 2$ for the gaussian distribution function and calculate the expected values and variances for the estimated distributions.
How does the expected values and variances change when you change the σ?

1.2 Part 2

In this part we are going to calculate the variance of the estimated distribution for very small values of σ's. Calculate the variance for $\sigma = 0.1$, $\sigma = 0.05$ and $\sigma = 0.01$.
To what number does the variance converge when you decrease σ?

1.3 Part 3

Calculate the variance of the estimated distribution theorically as a function of σ and our sample datas and explain why the variance converges to an specific number when you decrease the σ?

2 Transforming the distributions

2.1 Part 1

Consider that we have a random variable X with a distribution function $f_X(x)$. We will define another random variable Y which is a function of random variable X and we have $y = F_X(x)$ and $F_X(x)$ is CDF of random variable X. Prove that the distribution function of Y is Uni(0,1).

2.2 Part 2

Consider a random variable with Rayleigh distribution function:

$f_X(x) = \frac{x}{2\sigma^2} e^{\frac{-x^2}{2\sigma^2}}$ for $x > 0$ and $\sigma = 2$.

Generate 1000 sample from this distribution and calculate $y_{sample} = F_X(x_{sample})$ and estimate the distribution function of random variable Y from y_{sample}'s with parzen window method that you learned in question 1 with $\sigma = 0.1$ and plot the distribution function.
Calculate the mean and variance of the y_{sample}'s and validate that they are same with the expected value and variance of the uniform distribution.
Point : generate the random samples with numpy library and set the numpy.random.seed equal to your Student number.

2.3 Part 3

In this part, consider that we have a random variable Z with a distribution function $f_Z(z) = Uni(0, 1)$. Consider another random variable U which is a function of random variable Z, $u = g(z)$ and $g(z)$ is the inverse function of CDF of a specific distribution function:

$g(u) = F_U^{-1}(u)$

Prove that the random variable Y has a distribution function which is equal to $f_U(u)$.

2.4 Part 4

Use the y_{sample}'s that you generated in part 2 and calculate the $u_{sample} = F_U^{-1}(y_{sample})$. Consider that random variable U has a exponential distribution function ($f_U(u) = \lambda e^{-\lambda u}$ for $u > 0$) and $\lambda = \frac{1}{5}$. Estimate the distribution function of u_{sample}'s with parzen window method with $\sigma = 1$ and plot the distribution.

Calculate the mean and variance of the u_{sample}'s and validate that they are same with the expected value and variance of the exponential distribution function.

3 Classification problem

Medical reaserchs has shown that the blood pressure of people has a gaussian distribution. In this question, we consider two classes of people. Class 1 are the persons who don't have any diseases and Class 2 are the persons with heart disease. The blood pressure of the first class has a gaussian distribution with parameters $\mu = 8$ and $\sigma = 2$ and the blood pressure of the second class has a gausssian distribution with parameters $\mu = 11$ and $\sigma = 1$.

3.1 Part 1

Consider a community that 80 percent of them don't have any diseases but 20 percent of them have heart disease. Generate 1000 sample of blood pressure from this community.
Hint: Define a bernoulli random variable with parameter $p = 0.2$. Generate 1000 sample from that distribution. For each of these samples, if the sample data was equal to 0, generate a random number with gaussian distribution with parameters of the first class. Else generate a random number with gaussian distribution with parameters of the second class.
Point: Generate the random numbers with numpy library and before genearting, set the numpy.random.seed equal to your Student number.

3.2 Part 2

Now consider that the sample dataset that you generated in the previous part has been given to you and you don't have any idea that each sample belongs to which class. We are going to classify our sample datas using the Bayes' theorem. Consider that w_1 is class 1 and w_2 is class 2. So we have :

$p(w_i|x) = \frac{p(w_i)p(x|w_i)}{p(x)}$

And we have $p(w_1) = 0.8$ and $p(w_2) = 0.2$. The $p(x|w_1)$ is equal to gaussian distribution of class 1 and $p(x|w_2)$ is equal to gaussian distribution of class 2 and from the law of total probability $p(x) = p(w_1)p(x|w_1) + p(w_2)p(x|w_2)$.

Now we will classify each sample in this way :

If $p(w_1|x_n) > p(w_2|x_n) \Rightarrow x_n \in w_1$

If $p(w_1|x_n) < p(w_2|x_n) \Rightarrow x_n \in w_2$

This is equivalent to checking the condition:

$p(w_1)p(x_n|w_1) > p(w_2)p(x_n|w_2)$.

So classify each sample with the Bayes' theorem and make a confusion matrix which we call it C with two rows and two columns. The columns are for the true classes of the samples and the rows are for the classes that you have classified with the Bayes' theorem. For example the element in the first row and second column of the matrix is the number of samples that in fact belong to the second class but we have classified them to the first class.

3.3 Part 3

We will define two parameters for our classification based of matrix C elements. First one is accuracy which is defined in this way:
accuracy $= \frac{C_{1,1}+C_{2,2}}{N_{sample}}$
N_{sample}: total number of samples
And the second one is sensitivity:
sensitivity $= \frac{C_{2,2}}{C_{1,2}+C_{2,2}}$
Point: accuracy checks the correctness of the whole classification but the sensitivity checks the correctness of classification for the datas that belongs to the class 2. When we have unbalanced datas for two classes, the accuracy is not enough to check the correctness of the classifier.

Now find these two parameters for your classification.

3.4 Part 4

For a classifier to detect the heart disease, having high accuracy is not enough, because most of the people don't have this disease. So for a good classifier we need high sensitivity. Because if a person has heart disease and we predict that there is no disease, it is a very bad mistake but if he doesn't have any problem and you predict that he has disease, this will be clarified in the next experiments.

To avoid this issue, we will define a risk parameter r. Now we will change our classifier in this way:

If $p(w_1|x_n) > r * p(w_2|x_n) \Rightarrow x_n \in w_1$

If $p(w_1|x_n) < r * p(w_2|x_n) \Rightarrow x_n \in w_2$

Do the classification for r=2,5,10 and find accuracy and sensitivity and explain how each parameters changes when you increase the risk parameter.