Exam - Probability and Statistics

8th November 2020

Instructions

The complete solution must be sent as a unique PDF file with name "Exam_PS_nom_prenom.pdf" to the address giancarlo.fissore@inria.fr.

Deadline: Sunday, November 15th, midnight.

The PDF must contain the full code thoroughly commented; numerical results and plots must be visible on the file. If you use Jupyter Notebooks, you can easily export all the code and plots as a PDF file; if you use other tools, you should take care of exporting the results in the same way. Pen and paper exercises can be scanned and included in the same PDF file (many PDF merging utilities are available to download or for use online) or directly typed in digital form.

The Bonus exercises are not compulsory; they give extra points.

1 Computing Pi with Montecarlo

Montecarlo algorithms employ sampling of random variables to compute quantities of interest.

Starting from the observation that the ratio of the area of the unit circle to the area of the square enclosing it is $\frac{\pi}{4}$, we can write a montecarlo algorithm to compute an approximate value for π.

- Area of the unit circle (circle of radius 1, centered at the origin):

$$A_c = \pi r^2 \tag{1}$$

- Area of the square enclosing the unit circle
(square with vertices $(1,1), (-1,1), (-1,-1), (1,-1)$):

$$A_s = 4r^2 \tag{2}$$

- Ratio:

$$\frac{A_c}{A_s} = \frac{\pi}{4} \tag{3}$$

Algorithm:

- sample N values of x between -1 and 1, uniformly
- sample N values of y between -1 and 1, uniformly
- compute the ratio of the areas as the number of points (x, y) falling into the unit circle divided by the total N points:

$$\frac{\text{\# of points inside the circle}}{N} \sim \frac{\pi}{4} \tag{4}$$

Exercise. Implement the above algorithm and print an approximate value for Pi.

Remark. Points belonging to the unit circle satisfy the following equation:

$$x^2 + y^2 \leq 1 \tag{5}$$

Bonus exercise: plot the square, the circle, and the sampled points (different colors for points inside and outside the circle).

2 Continuous distributions

2.1 Cumulative Distribution Function (CDF)

The CDF of a continuous probability distribution is defined as

$$C(z) = \int_{-\infty}^{z} p(x)dx \tag{6}$$

Useful properties of the CDF:

- $C(+\infty) = 1$ (follows from normalization)
- the function is non-decreasing

More on the CDF: `https://en.wikipedia.org/wiki/Cumulative_distribution_function`

2.2 Inverse Transform Sampling

Given the properties of the CDF, we can define an algorithm to sample from arbitrary probability distributions starting from uniform samples.

Algorithm. Sampling from $p(x)$:

- Compute the CDF of $p(x)$: $C(z)$

- Compute the inverse of the CDF: $F(x) = C^{-1}(x)$
- Sample a uniform random number u in the interval $[0, 1]$
- Compute a sample x distributed according to $p(x)$ using the formula: $x = F(u)$

More details: `https://en.wikipedia.org/wiki/Inverse_transform_sampling`

2.3 The exponential distribution

Let's consider the exponential distribution with parameter λ:

$$p(x) = \frac{1}{\lambda} e^{-x/\lambda}, \qquad x \geq 0 \tag{7}$$

Pen and paper exercises.

1. Compute the mean μ and the variance σ^2 of the distribution.
2. Compute the corresponding Cumulative Distribution Function (CDF); we call it $C(z)$.

   Remember: $p(x)$ is defined for positive values of x.
3. Compute the inverse of the CDF: $F(x) = C^{-1}(x)$
4. Provide a concise explanation of how the Inverse Transform Sampling algorithm works.

Coding problems. Set $\lambda = 3$.

5. Implement a function that computes samples from the exponential distribution using the Inverse Transform Sampling algorithm. Plot a histogram of the distribution.
6. Compare your sampling function to the *numpy* function for sampling exponential distributions. Build 2 histograms and superimpose the theoretical curve of $p(x)$.
7. Consider the function $t(x) = x^2$. Supposing that the variable x is distributed according to an exponential distribution with parameter $\lambda = 3$, compute the mean m of t over $N_s = 10000$ samples. Repeat the process N_t times and give an estimate of the error. What is the minimum value of N_t needed to obtain an error smaller then $\epsilon = 3 \cdot 10^{-3}$?

3 SVD and MNIST

We will use the MNIST dataset.

1. Compute the SVD of the full dataset X
2. Visualize the first 10 components as images.
3. Calculate the projections of all the digits (0 to 9) on the first 3 components.
4. Produce 2 scatter plots along the directions 1-2, 2-3. Choose a different color for each digit.

Bonus. Produce a 3D scatter plot along the directions 1-2-3.