Advanced Machine Learning: HW 1

Note: this text is dummy translation of task formulated in Russian: https://docs.google.com/document/d/1snU4dXicuPmFz9XjTF8c7nSw0qGdJAgdJeZVrg-NakI

The first HW consists of two parts: the first part is about Bayes' theorem and general probabilistic reasoning, the second part is about linear regression. As a solution, a link to the jupyter notebook on your github (either public or with access for the snikolenko user) is expected. The decision must be submitted on the Academy portal.

Part I: On Bayes' Theorem

The first part consists of two questions and one task. Its meaning is to carry out probabilistic reasoning in situations where the model itself is very simple, but it still needs to be constructed correctly, to correctly reflect the life situation. Therefore, I recommend giving detailed answers to the first two questions; it is better to write these answers directly in a notebook, LaTeX in jupyter works at a level that is quite sufficient for us.

There was a murder. At the scene of the murder, blood was found, which clearly (we assume that with probability 1) belongs to the real killer. Blood belongs to a rare group that is present in only 1% of the population. And it so happened that the defendant has just this rare blood type!

1. The prosecutor comes out first and says: “The chance that the defendant would have just such a blood type if he was innocent is only 1%; so, with a probability of 99% he is guilty, I propose to condemn ”. Where is the prosecutor wrong? Indicate which probabilities he estimated and which he should have estimated.
2. Then a lawyer comes out, explains the prosecutor's mistakes and takes the floor himself: “A million people live in the city. This means that about 10,000 of them have this blood type. This means that all this blood tells us is that the defendant committed murder with a 0.01% probability. It turns out that this testimony is not just not a strict proof, but even gives us a negligible chance that my client is guilty, so I propose to exclude him from the case materials ”. What is the lawyer wrong? What probabilities did he estimate and what should he estimate?
3. Let's go back to the test for a terrible disease from the first lecture. Let me remind you that according to the assumptions of the problem, 1% of the population is sick, and the test has a probability of error of 5% in both directions. We have seen that as a result, a person who tests positive has a posterior probability of being sick of only about 16%.

   The doctors decided that this could not be tolerated; they are ready to conduct urgent research and improve the quality of the test. But there will be enough energy and money only to reduce one type of mistakes, i.e. decrease either only the number of false positives (when the test gives a positive result in a healthy person), or only the number of false negatives (when the test gives a negative result in a patient).

   Help the doctors: build the dependences of the posterior probability of illness after a positive and negative test on both types of errors (for example, in the form of graphs), draw conclusions and give recommendations to doctors - what is better to focus on?

Part II: About Linear Regression

This part of the first homework assignment deals with analyzing a specific dataset. Let's take a dataset that is pretty relevant to real life; it is available at the following link:

https://ourworldindata.org/coronavirus-source-data

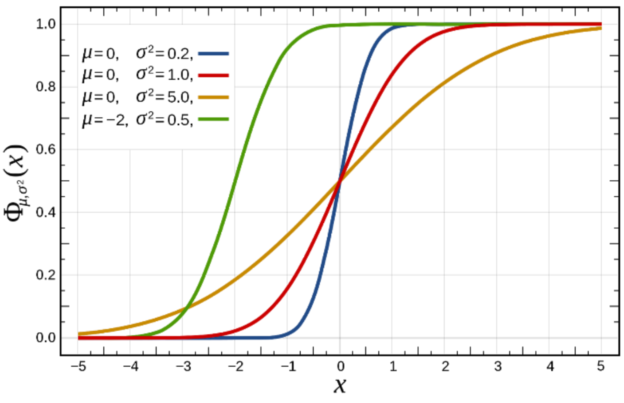
1. Download data in csv format, select data for Russia from the table, starting from March 3, 2020 (at that moment there were more than 2 cases for the first time). Take the number of cases (columns total_cases and new_cases) as the target variable; to simplify processing, you can replace all zeros with ones in the new_cases column. For consistency, let's capture the training set as the first 50 counts (days) starting March 3rd; the rest of the data can be used as a test case (and it will even grow as the task progresses). In other words, we will "play" for the Russian authorities, who are trying to figure out what to do by looking at the data on the epidemic in May 2020.

2. Plot the target variables. You will see that the number of cases is growing very quickly, at first glance, exponentially. For the first approach to the projectile, let's use this.

   a. Using linear regression, train the model with exponential growth in the number of cases: y ~ exp (linear function of x), where x is the number of the current day.
   b. Find the posterior distribution of the parameters of this model for a sufficiently wide prior distribution. Estimate the required value of the variance of the noise in the data based on your own maximum posterior model (this is actually the first step of empirical Bayes).
   c. Sample many different exponents, build graphs. How many coronavirus cases are predicted in Russia based on these samples by May 1, 2020? by June 1? by September 1? Plot predictive distributions (you can empirically, based on sampling data).

3. The predictions of the exponential model must have turned out to be sad. But this, of course, is too pessimistic - exponential growth in nature can never last forever. The curve of the total number of cases during an epidemic in reality has a sigmoidal appearance: after the initial phase of exponential growth, saturation inevitably occurs. As a specific form of such a sigmoid, let's take the form of the distribution function for the Gaussian:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$



Naturally, in our case, the sigmoid does not tend to unity, i.e. the constant before the integral can be arbitrary (and it can be introduced into the exponent), and in the exponent under the integral there can be an arbitrary quadratic function of t.

a. Suggest a way to train the parameters of such a sigmoidal function using linear regression.
b. Train these parameters on a dataset of coronavirus cases in Russia. Find the posterior distribution of the parameters of this model for a sufficiently wide prior distribution. Estimate the required value of the variance of the noise in the data based on your own maximum posterior model.
c. Sample many different sigmoids from the posterior distribution, plot graphs. How many cases of coronavirus will be in Russia based on these samples? Build an empirical predictive distribution, draw graphs. What is your projection for the number of coronavirus cases in the pessimistic scenario (90th percentile in the sample of cases)? In the optimistic scenario (10th percentile)?

Bonus: conduct the same analysis for other countries (here you will have to pick up the days of the start of modeling with your hands - the coronavirus came to different countries at different times). How different are the parameters? Is it possible to divide countries into clusters (at least visually) depending on these parameters?

[This part of the assignment is not graded, there are no right or wrong answers, but I will be glad to know what you think]

What have you learned from this exercise? What can you say about the coronavirus based on the results of such a simulation? How to make a decision, for example, whether to enter quarantine?