Madrid, February 2019

CONCEPTUAL QUESTIONS

I. Would you rather have too many false positives, or too many false negatives? Explain.

The severity of one or another depends of the problem that you are dealing and the cost associated with each type of error (FP or TP), the idea in general is try to minimize the error that is most costly. In some cases this is easy to estimate, because the cost is measurable (as for example money invested, number of clicks, number of sales, etc), in other cases the cost is more related with ethics or subjective costs, that are not easily measurable.

$$\text{Min } Cost_{FN} \cdot FN + Cost_{FP} \cdot FP$$

In order to explain what I treating to say I will give some examples:

- Send or not a person to a death penalty. Here, you want to avoid FP, because this means that we will sending innocents to prison and then condemn them to death, so the cost o make this kind of mistake is high. So, our loss/cost function should minimize that error.
- Detect a sickness (AIDS, Cancer, etc). Here, a FN is dangerous, because would mean that a person that is sick, is not receiving the treatment that can help her to fight the illness.
- Give o not a job to a candidate at the end of a selection process. In this case a FN would be problematic, because imply the cost associated with miss a better fit for the job.
- Detect or not an e-mail as spam. Again, the worst alternative here is a FP, because would mean classify a good e-mail (from your friends, clients or your boss) as spam.
- Recommendation systems. The FP are not of concern because give a bad recommendation does not affect the number of sales, but a FN would harm the sales of the product that for sure a customer would have buy.

- Classify a household as poor or not. Here, you are interested in minimize the FN, this mean you want to minimize the number of household that are misclassified as not poor. Here, the cost of misclassification is not help people with social programs.
- Churn prediction. Here, it is measurable the cost of losing a customer and the cost of invest in a customer that will not churn. Here, one or another type of error would be minimized according to what is the cost that you want to minimize.

II. Give an example from your work experience of selection bias. Was it important? In general, how can you avoid it?

Selection bias occurs when a part or proportion of the population is not represented in the sample data that you are using for your analysis. This have very big consequences because if your sample not contains all the patterns of behaviours that represents all the possible outcomes of your problem, your conclusions would be incomplete and the results would not be generalized. This is a real problem in surveys where problems as absence of response, deliberate choice of the sample, between others, affects the validity of the results.

Example:

An example of my personal experience is related with a problem in an insurance company. In this problem, the procedure of a call center to sell was offer to the clients different products with different prices (to simplify consider low, medium and premium products). The operators had demographic information and some socioeconomic data about the client, and they have to get the maximum sales possible. The final result was a biased sample with information of the clients, the received offer (low/medium/premium) and a flag that indicates if the client reject or accept the offer.

For the operators of the call center the most simple thing is to start offering the premium products to the clients with higher socioeconomic status, but this is not necessary the right answer, because there are other factors that can affect that a person accept or not a product, and is related with the knowledge about the product, if they have buy a similar product in the past, number of siblings that affect their budget, etc. Not all the people in the group 'high socioeconomic status' make the same purchase decisions.

The same happen for the rest of the clients.

So, in order to have a good dataset to make predictions, the best option was to create three stratified samples of people (one group for each product). In each sample, we will have people from different socioeconomic status and then the resulting sample (after calling the clients) will have information about the behavior of all types of people with different options.

III. What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)? Can you think of an example where Bayesian estimate is the most appropriate method? Another example in which MLE is the most appropriate method?

These two methodologies are used to estimated parameters of a model, as for example the parameters of a distribution (poisson, gamma, etc) or the weights of models like logistic regression or generalized linear models.

These methodologies assume that we know the distribution of the likelihood $p(X|\Theta)$ (where X is a iid r.v. and Θ is the set of parameters that we want to estimate). In the MLE to get Θ we maximize the log-likelihood to get a point value estimation.

In the case of Bayes, the likelihood is an ingredient that we combine with a prior distribution to get the posterior distribution of Θ, this is $p(\Theta|X)$. The prior is some sort of knowledge or expert criteria that allows to include more information to the estimation.

Then, Bayes estimation requires to compute the full posterior distribution of the parameter(s) of interest, opposed to just obtain the value where the posterior gets its maximum value ($p(\Theta|X) \approx p(X|\Theta)p(\Theta)$). Introduce a prior have several benefits, first help to avoid over-fitting when considering extra information and not solely estimates based in data, this is good when the data is scarce and allows to introduce scenarios that help to reduce uncertainty.

The MLE is a simple model that works well in general, although it can fail with few data. Bayes estimate is a generalization of the MLE. But most of the time MLE is chosen for its simplicity.

Part II - Conceptual Problem: The table shows data for two retail stores. It includes

the average weekly sales, the standard deviation of the data and the number of data points. You are in charge of giving a bonus to the store managers based on performance. What would you do?

With the information that we have it is necessary to make some assumptions about the problem, because there are elements that we do not know. For example, we do not know the distribution function of this two samples, and our conclusions would be not controlled by external variables that may explain most of the variation between this two retail stores. Variables like: location, socioeconomic status, number of competitors close to the store, number of employees in the stores, between many other factors determine the sales potential of this two stores, for these reasons, it is normal to think that other variables should be considered to understand correctly the differences and make a good decision based on data.

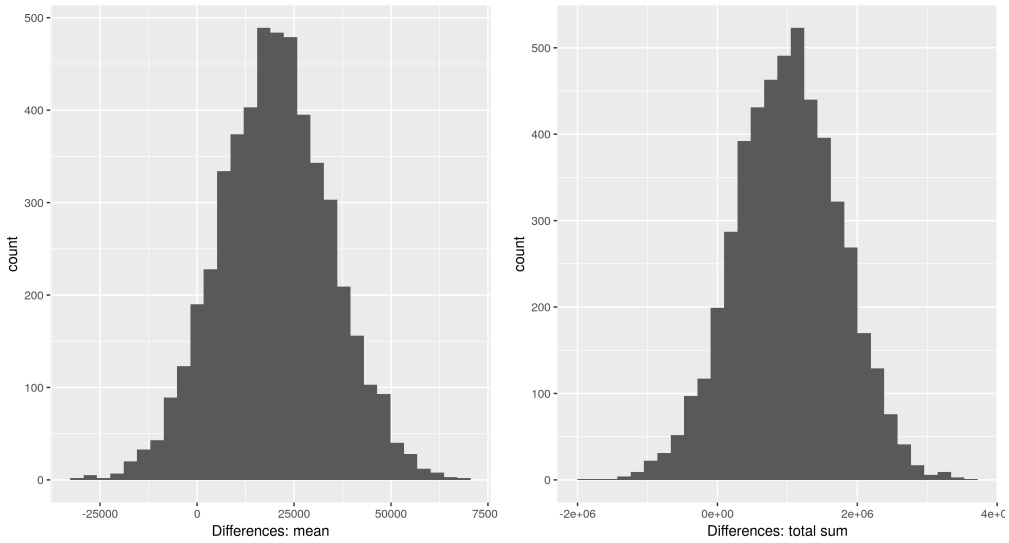
With the given information we can do some analysis that would give us an idea of what is happening.

- Hypothesis testing of differences of means. With this we want to see if there is or not random differences between the two distributions. If there are not differences, then there are not bonuses for anyone.
  - Here we take the central limit theorem idea of normality of the means. Here we will use z-test of two tails.
  - Bayes theorem. In order to work with a not frequentist alternative.

    In this case for time limitations we consider only the frequentist alternative.
- Monte Carlo simulation: simulate the 50 weeks with the given parameters, but different distributions and see which is the best under the two conditions (measure by the total of money that they make in the 50 weeks). This help us to understand better the results of the Hypothesis testing analysis.
  - Using the mean per each week
  - Using the total per week.

Applying hypothesis testing it seems that there is not differences between the means of the two retail stores. It seems that a mean difference of 20.000 does not mean too much in comparison with average values between 780.000 and 800.0000 per week for the given st. and sample size. So, in order to

understand better the differences, we simulate the sales of the 50 weeks to get the mean and the total sum of the differences in sales. We generate a sample of size 1000. We will use the normal distribution to get the values.



In the figures we can see the differences between Downtown store and Suburban store, these graphs seems to indicate that most of the time 'Downtown store' can improve the results of 'Suburban store', clearly when we took the total sum the difference seems important. But, what if the size of the difference that makes one better than the other? In this case we propose to create a **Key Condition** based in this simulation and give a bonus to the 'Downtown store' those times that exceeds this condition, considering the total of sales instead of the mean.

Although as we indicate before we need more information to give a better answer.