Data Science Test

Vlad-Marius Griguta

Q1. You measured two variables 100 times and found that their correlation was 0.68. Later, you realised there was a systematic constant offset when reading one of the two variables. How would you adjust the measured correlation?

A1: There is no need to adjust the measured correlation. The reason is that the correlation is conserved by the addition of a constant offset to any of the two variables.

Q2. Consider a roulette wheel game. It has 37 spots: 18 red, 18 black and 1 green, each has probability 1/37 of being selected. A player can choose only red or black, and the player wins the same amount he/she puts in if the wheel stops on the same colour chosen by the player (if the wheel stops on green, the player loses).

- **What is the expected payoff?**
- **If you were to try your luck by placing £100 worth of bet (putting £1 on the table 100 times or £100 on the table once count as placing £100 worth of bet. Putting £1 on the table 150 times counts as placing £150 worth of bet), what would you do? Explain your decision.**

A2:

1. Let *X* be the amount bet. The expected payoff is $\left(\frac{18}{37} * 2 + \frac{18}{37} * 0 + \frac{1}{37} * 0\right) X - X = -\frac{1}{37} X$
2. There is a negative expected payoff for playing the roulette wheel game. Therefore, the rational decision would be to not play the game at all.
   If I were to try my luck, I would focus on reducing the potential loss by eliminating the risk. In this circumstance, the risk can be mitigated by diversification, which involves betting £1 for 100 times. This way, the risk of losing the whole bet at once would be reduced from $\frac{19}{37} = 51.35\%$ to ${\frac{19}{37}}^{100} = 1.1 * 10^{-27}\%$

Q3. If you observe from a sample the value of a ratio (e.g. default rate or click through rate) to be R', when the expected value is R, how would you determine whether the difference between R and R' is statistically significant? Is there anything outside of pure statistical considerations that you would also consider?

A2:

To test if the new result R' deviates significantly from R we would need to know the standard deviations of the measurements of both variables, σ_R and $\sigma_{R'}$. These can be estimated from the variance of the list of values measured. Then we can compute if the measured values are consistent with each other through the following ratio $r = \frac{|R-R'|}{\sqrt{\sigma_R^2+\sigma_{R'}^2}}$. In the academic literature it is generically accepted that if $r > 3$, or in other words, if the probability of the outcome of measuring R' is less than 0.3%, then R' is statistically different from R.

Besides the expected statistical variance of R', we should also consider if there are any sources of systematic uncertainty in the measurements of both R and R'.

Q4. Given a sample of 1000100 events, of which 100 were 'signals'. A model was built to pick out those signal events. In one setting, the model predicted 200 signals, of which 90 were true positives (i.e. genuine signal events).

- **Estimate the area under the ROC curve. Would you consider this as a good model?**
- **Would you use this metric, or other metrics such as precision-recall? Explain your choice.**

From the data provided, we can estimate the true positive rate as $\frac{90}{90+10} = 0.9$ and the false positive rate as $\frac{110}{110+10^6-110} = 1.1 * 10^{-4}$. Using the trapezoidal approximation to estimate the area under the ROC curve, AUC yields 0.95. The model used was extremely efficient in extracting the signals from the noise.

For the classification task presented I would NOT use an accuracy metric that averages over multiple thresholds such as the area under the ROC curve. The reason is that AUC does not offer the flexibility of customizing the cost of misclassifying across different classes. An alternative metric would be to consider the ROC curve in its entirety and select a threshold that yields the best compromise between signal and noise. However, for that we would need to analyze the model outputs before they are converted to binary outcomes.

Still another alternative metric is the f_k score, which tests the accuracy by considering both the precision and the recall metrics. If the cost of 'losing' signal would be higher than the benefit of having more noise, I would advise weighting the recall more by selecting $k > 1$.