Homework 3 - SDS

Laura Mignella, Elisa Pierini and Maria Vittoria Vestini

---

1 Friedman's paper

In the *Multivariate Goodness–of–Fit and Two–Sample Testing* paper, the author Jerome H. Friedman explains how to perform the hypothesis testing in both the problems of *goodness-of-fit* and *two-sample testing*.
The *goodness-of-fit* problem occurs when we have a data set $\{x_i\}_{i=1}^N$ and we have to figure out whether the density distribution $p(x)$ of the data set is the same as the one of the given distribution $p_0(x)$.
The *two-sample testing* problem, instead, occurs when we have two different data sets, $\{x_i\}_{i=1}^N$ drawn from $p(x)$ and $\{z_i\}_{i=1}^M$ drawn from $q(z)$, and we have the goal to test if $p = q$.
Most of the procedures for those problems are implemented for observations with only one attribute. The idea proposed by Friedman was to extend the procedures also to multivariate problems. Machine learning can be used to predict a value y given a sample x. To do so, we will build any binary classifier, produce a model $F(X)$ and assign the value $\hat{y}$ using a decision rule with threshold t^*, chosen to minimize the error rate.

$$\hat{y}(x) = \begin{cases} 1 \; if \; F(x) > t^* \\ -1 \; otherwise \end{cases}$$

While dealing with the *Two-Sample testing* problem, we will collect all of our data in a single data set $\{u_i\}_{i=1}^{N+M} = \{x_i\}_{i=1}^N \cup \{z_i\}_{i=1}^M$ and assign the response value: $y = 1$ to the observations originated from the first sample x, $y = -1$ to the ones form z.
Later on, using the binary classifier, we generate two set of score values: $S_+ = \{s_i\}_{i=1}^N$, $S_- = \{s_i\}_{i=N+1}^{N+M}$, with $\{s_i = F(u_i)\}_{i=1}^{N+M}$
These scores can be viewed as samples from two different univariate distributions $p_-(s)$ and $p_+(s)$, so, we managed to achieve what we wanted: We went from a multivariate distribution to a univariate one, and now we can apply one of the known statistic tests.

$$\hat{t} = T(\{s_i\}_{i=1}^N, \{s_i\}_{i=N+1}^{N+M})$$

Now we want to a actually verify the null hypothesis $H_0 : p = q$. Given a significance level α and the distribution of the null hypothesis $H_0(t)$, we can reject $p = q$ if the value $\hat{t} > quantile(H_0(t), 1 - \alpha)$.
For the *goodness–of–fit* problem, we can just apply the same process with:

- z an artificial "Monte Carlo" sample from p_0, with $q = p_0$.
- $H_0 : p = p_0$.
- $p(x)$ the unknown distribution of x.

2 Statistic Tests

The **Mann–Whitney test** is a nonparametric test. Randomly choosing two samples, of sizes n_1 and n_2, from two different populations: X from the first population, and Y from the second, the test is used to test the null hypothesis $H_0 : \; P(X > Y) = P(Y > X)$, the alternative hypothesis H_1 is that $P(X > Y) \neq P(Y > X)$.
The Mann–Whitney test statistic is calculated by sorting all observations in ascending order, then to the smallest observation is assigned a rank of 1, to the second smallest observation is assigned a rank of 2, and so on. In the end, the test statistic is calculated by summing the ranks for the observations in one of the populations, obtaining the values R_1 and R_2.
To finish the test we need to find the test statistic $U = min(U_1, U_2)$, where $U_i = n_1 * n_2 + \frac{n_i \cdot (n_i+1)}{2} - R_i$.

The **Kolmogorov–Smirnov test** is also a nonparametric test. It is used to determine if two samples come from the same distribution or a single sample follows a particular distribution. So, the test can be used in two ways:

- The first is the one-sample Kolmogorov–Smirnov test, which compares a sample with a reference probability distribution p_0. The null hypothesis in this case is $H_0 : \ p_0 = p(x)$ where $p(x)$ is the distribution of the sample.
- The second is the two-sample Kolmogorov–Smirnov test, which compares the distributions of two different samples. The null hypothesis in this case is $H_0 : \ p(x) = q(x)$ with $p(x)$ the distribution of the first sample and $q(x)$ the distribution of the second sample.

The K-S test statistic is the maximum difference between the empirical CDF of the sample and the CDF of the reference probability distribution (in the case of one-sample K–S test) or the maximum difference between the empirical CDFs of the samples (in the case of two-sample K–S test).

We use the Kolmogorov-Smirnov and Mann-Whitney tests because they both do not require any assumptions about the distribution of the data to be tested.