Name: Barba Paolo, Candi Matteo

Assay of Friedman's paper
Homework 3

Statistics for Data Science

Summary

Friedman's paper proposes a methodology for solving two-sample testing in the case when each observation consists of many measured attributes $x_i = x_{i1}, x_{i2} \dots x_{in_1}$ and $z_i = z_{i1}, z_{i2} \dots z_{in_2}$. The goal to achieve is to test whether two (Multivariate) distributions are equal or not. The idea behind this method is mixing machine learning algorithms with Hypothesis testing. First labeling the two samples and then perform machine learning algorithm to compute the score of the sample and only then perform a two-sample hypothesis test (Testing whether the two score distributions are equal or not) using a two-sample statistic test as Kolmogorov–Smirnov(Measure the difference between the two cdf $D_{n_1,n_2} = \sup_x |F_{1,n_1} - F_{2,n_2}|$) test or other as chi-squared, Mann Whitney. To perform so, it's needed to build reject/accept regions, and therefore it's fundamental to know the distribution under the null hypothesis of the statistical test. When the same data is used for both training and subsequent scoring, these univariate null distributions are not valid. To solve this, it can be possible to perform Fisher's permutation. In Friedman's paper, it was proposed the following procedure to compute the distribution of the statistical test under H_0 instead :
Suppose that the sample z_i came from a reference distribution (say p_0).

- Drawn a sample of size n_1 from p_0.
- Use them with the actual data to train the classification model and compute the scores.
- Compute the statistics test between the two scores sample.
- Repeat P times.

At the end, it produces a set of statistic values $t_1, \dots t_P$ that can be used as the distribution of statistical test under H_0 building the reject region R_α as the values of test statistic greater or equal then $1 - \alpha$ quantile of $t_1, \dots, t_P$. Using the additional information from the reference distribution p_0 this procedure has the potential for increased power of the test.

Comments

The performance of this procedure strictly depends on ML model and statistic test used and consequently on sample size, the dimensionality of the distributions, and distribution distance(To compute the power). Model less complex would perform better when the dimension of the distributions grows, in fact (under perfect information) it is much less likely for any two distributions to be found close together (after all, they must be close in every single dimension to be close overall). This means that distribution in high dimensions tends to be more separated, on average, than distribution in low dimensions, assuming the new dimensions actually add information. For this reason, simplistic classifiers tend to work as well or better than more complicated classifiers as the dimensionality grows.