Syllabus of the course
"Multivariate analysis and nonparametric statistics"

Author: Vladimir Panov
email: vpanov@hse.ru
tel. 8(495)7729590*26215

Introduction

This course handles various methods of solving popular statistical tasks like probability density estimation, describing the dependence structures via regression models, and providing statistical tests. All methods considered in this course require only few assumptions about the probabilistic properties of the model from which the data were obtained. For instance, they forgo the assumption that the original distribution is normal.

In this course, we show the implementation of considered appoaches in statistical software (preferably in the R-language), and demonstrate how these methods can be used for the solution of some real-world problems.

Content

Part I: Probability density estimation.

1. Statement of the problem. Estimation of the distribution function.
2. Histogram as a density estimate. Bias-variance tradeoff. General concept and particular results for the histogram. Bias-variance decomposition for histograms. Minimization of AMISE for histogram: Scott and Friedman-Diaconis rules for bandwidth selection. Other ideas for the choice of the amount of bins: Sturges rule. "Pretty" procedure in the R language.
3. Kernel density (Parzen-Rosenblatt) estimates. Bias-variance decomposition for kernel estimates. Minimization of AMISE for kernel estimates with respect to the kernel: Epanechnikov kernel. The notion of the kernel efficiency. Minimization of AMISE for kernel estimates with respect to the bandwidth: nrd and nrd0 options. (Unbiased) cross-validation for the probability density estimates.
4. Rates of convergence for histogram and kernel density estimates. Lower bounds for density estimates: van der Vaart's theorem.

Part II. Nonparametric regression.

1. Statement of the problem.
2. Nearest neighboors algorithm, local averaging. Method "Super smoother". Cross-validation approach and the bass parameter.
3. Local regression. Method "Loess" ("Lowess").
4. Generalized cross-validation. Motivation of the algorithm: theorem about the closed form of the cross-validation error for linear regression.
5. Akaike criterion.
6. The Nadaraya-Watson kernel estimator. Modifications of this estimator (local polynomial estimator, Gasser-Muller estimator).

7. The notion of linear smoother. Regressogram.

Part III. Nonparametric tests.

1. Tests for independence I: Kendall's tau. Unbiased estimate for Kendall's tau. Exact distribution of this estimate for n=3. Large-sample approximation for the constructed estimate. Calculation of the mean and the variance. Construction of asymptotic confidence intervals. The notion of bootstrap. Relation between Kendall's tau and the Pearson correlation coefficient.
2. Tests for independence II: Spearman's rho. Equivalent form of the Spearman's rho. Exact distribution of Spearman's rho for n=3. Large-sample approximation for the constructed estimator. Calculation of the mean and the variance.
3. Paired replicates data. Wilcoxon test. Exact distribution for n=3. Large-sample approximation. Calculation of the mean and the variance.
4. 2 independent samples. Wilcoxon statistics and Mann-Whitney statistics. Mann-Whitney test. Exact distribution for n=3 and m=2. Large-sample approximation. Calculation of the mean and the variance.
5. Many independent samples. Kruskal-Wallis test. Relation to the ANOVA test. Exact distribution for k=3, n1=n2=n3=2 (only general idea). Large-sample approximation (without proof).
6. Two-way layout. Friedman's test (only general idea).

Part IV. Bonus lecture

Wavelets. Haar basis. The notion of resolution. Application of this idea to the regression problem. Hard and soft thresholding. Motivation of the soft thresholding: theorem about the situation, when all coefficients are equal to zero.

Recommended literature

a) Book about possible applications of nonparametric methods with examples in the R language:
   1. Hollander, M., Wolfe, D. and Chicken, E. Nonparametric statistical methods. John Wiley and Sons, 2014.

b) Theoretical books:
   2. Wasserman, L. All of nonparametric statistics. Springer, 2006.
   3. Tsybakov, A. Introduction to nonparametric estimation. Springer, 2009.

c) Some other useful books
   4. Sprent, P., Smeeton, N. Applied nonparametric statistical methods. 4th ed., 2007.
   5. Hastie, T., Tibshirani, R. and Friedman, J. The elements of statistical learning: data mining, inference, and prediction. Springer, 2001.
   6. Silverman, B. Density estimation for statistics and data analysis. Springer, 1986.

7. Лагутин, М.Б. Наглядная математическая статистика. Бином, 2007
8. Haerdle, W., Spokoiny, V., Panov, V. and Wang, W. Basics of modern mathematical statistics. Springer, 2013

Student evaluation

The final evaluation grade is calculated according to the formula:

[Final mark] = 0.3 * [cumulative mark for the work during the modulus]
+ 0.7 * [mark for the final test].

The cumulative mark for the work during the modulus is based on the mark for the home tasks and on the activity during the seminars. If the student missed 5 pairs or more, the final mark can be reduced by 25 %.