Program of the course «Nonparametric statistics»

Lecturer: Vladimir Panov (vpanov@hse.ru, https://www.hse.ru/en/org/persons/93419930)

Department of Statistics and Data Analysis

Meeting Minute # 1 dated 22 May 2019

Course Description

a) Pre-requisites. Basic course on probability theory and mathematical statistics.

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

In this course, we show the implementation of considered approaches in statistical software (preferably in the R-language), and demonstrate how the subsequent application of different methods leads to complete statistical analysis of the real-world problems.

Learning objectives

1. Study of the basic concepts of nonparametric statistics.

2. Study mathematical background of the nonparametric statistical methods.

Learning outcomes

1. Understanding the difference between parametric, nonparametric and semiparametric methods.

2. Understanding the methodology of application of the nonparametric statistical methods.

Course Plan

Part I: Probability density estimation.

1. Estimation of the distribution function.
2. Histogram as a density estimator.
3. Bias-variance tradeoff. General concept and particular results for the histogram.
4. Bias-variance decomposition for histograms.
5. Minimization of AMISE for histogram: Scott and Friedman-Diaconis rules for bandwidth selection.
6. Other ideas for choice of the amount of bin: Sturges rule.

7. “Pretty” procedure in the R language.
8. Kernel density estimates.
9. Bias-variance decomposition for kernel estimates.
10. Minimization of AMISE for kernel estimates with respect to the kernel: Epanechnikov kernel. The notion of efficiency of a kernel.
11. Minimization of AMISE for kernel estimates with respect to bandwidth: nrd and nrd0 options.
12. Cross-validation (biased and unbiased) for the probability density estimators.
13. Plug-in estimates for bandwidth selection: Park and Marron estimate.

Part II. Nonparametric regression.

1. Linear smoothers. First examples: regressogram and local averages.
2. Cross-validation for regression models.
3. Local regression. The Nadaraya-Watson kernel estimator.
4. Spline approach: wavelets, multivariate adaptive regression splines (MARS).

Part III. Nonparametric tests.

1. Tests on the location parameter: Wilcoxon, sign test.
2. Two-sample location problem: Mann-Whitney test.
3. Tests for many samples: Kruskal-Wallis test.
4. Testing the independence: Kendall’s test.

Part IV. Dimension reduction techniques.

1. Principal component analysis.
2. Independent component analysis. Negentropy and curtosis.
3. Projection pursuit regression. Alternating procedure.
4. Neural networks. Single-layer perceptron.

Reading list

Required

Wasserman, L. All of nonparametric statistics. Springer, 2006. Electronic resource,

https://link.springer.com/book/10.1007/0-387-30623-4

Optional

Tsybakov, A. Introduction to nonparametric estimation. Springer, 2009. Electronic resource,

https://link.springer.com/book/10.1007/b13794

Grading System

The final evaluation grade of the students 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 %.

Examination type

March 26, 2019

Course "Multivariate analysis and nonparametric statistics". Exam

Duration: 2 hours 40 minutes

Each practical task (1,2,3) - 6 points, each theoretical task (T1 - T3) - 4 points, each question (Q1 - Q10) - 2 points. The maximal score is equal to

$$3 * 6 + 3 * 4 + 10 * 2 = 50.$$

The total scores will be converted into marks (1-10) in accordance to the rule, which will be announced after the exam (e.g., 45-50 can be converted into 10, etc.).

Consider the dataset "mtcars", which includes the data on fuel consumption of 32 automobiles produced in 1973-74. For information about this database, type "?mtcars" in the R language.

The aim of this study is to construct the regression model which can explain how the fuel consumption (first variable, "mpg") depends on other numeric characteristics of the car (variables 2-7).

1. (Tests.) Display the boxplots which show the fuel consumption of the cars depending on the amount of cylinders. Test the hypothesis that medians of the fuel consumption of cars with various amounts of cylinders coincide. For any car with 8 cylinders (Group 1) find a car with 4 or 6 cylinders with similar standardised characteristics "disp","hp","drat", "wt", "qsec" (Group 2). Test the hypothesis that the fuel consumption of cars in Group 1 and Group 2 are the same.

   T1 Find the mean and the variance of the Mann-Whitney statistics.

   Q1 Explain what is the large-sample approximation in statistical tests, and how it can be used for testing the hypothesis for paired replicates data.

Q2 Explain the relation between the Kruskal-Wallis statistics and the ANOVA test.

Q3 Explain why the Kendall tau is equal to 1 if and only if the Pearson correlation coefficient is equal to 1, provided that the data follow the multivariate normal distribution.

2. (Regression.)

(i) Fit the supsmu (Super Smoother) model describing the dependence between "mpg"(as y-variable) and variable "disp" (as x-variable). Construct the estimators

- under various choices of span parameter (0.05, 0.2, 0.5),
- with span chosed by cross-validation.

Find the best model in the sense that the mean-squared error is minimal.

(ii) For the same variables, fit the kernel regression under various choices of kernels (Gaussian, Epanechnikov), and various methods for bandwidth selection (Akaike criterion, least-squares cross-validation). Find the best model in the sense that the mean-squared error is minimal.

(iii) Repeat the steps (i) and (ii) with other 4 variables ("hp","drat", "wt", "qsec") instead of "disp". As the result, provide a table, which summarizes the information about the best models for each variable.

(iv) Based on the results obtained on the step (iii), find 2 variables which (in your opinion) are at most useful for describing the fuel consumption. Argue your choice. Construct the multivariate LOESS estimator based on these 2 variables. Plot the corresponding 3-dimensional graph.

T2 Formulate the Akaike criterion for the choice of parameters in regression problems. Show the mathematical origins of this criterion.

Q4 Formulate an optimization problem, whose solution is the Nadaraya-Watson estimator.

Q5 What are the possible choices for the span parameters in the method "super smoother"?

Q6 What is the effective degrees of freedom? What does the value of this characteristic mean for linear regression?

3. (Density estimation.) Find the residuals of the linear regression estimator between "mpg" and "disp". Estimate the probability density function of these residuals using the histogram and the kernel density estimator, and display the corresponding graphs on the same plot. Use also other methods for showing whether the distribution is normal or not (various tests and graphs). Interpret the results.

T3 Assuming that the density is supported on [0,1], prove the explicit formula for the main terms of the MISE for the histogram estimator. Explain what is the bias-variance tradeoff in this case.

Q7 For which classes of densities the histogram and the kernel density estimators are optimal density estimators?

Q8 Which essential idea is used in "nrd" and "nrd0" rules for the choice of the bandwidth parameter in the kernel density estimation?

Q9 What is the cross-validation score for the density estimation?

Bonus question:

Q10 What is meaning of the term "resolution" in the context of the multiresolution analysis (MRA)?