

Методы прикладной статистики

Домашнее задание 2

Максимум- 10 баллов

Домашнее задание необходимо направить на почту a.a.egorov07@gmail.com. Допустимые форматы - R html или PDF файл с описание результатов + скрипт файл с кодом, на основе которого были получены результаты.

Дедлайн – 21.12.2020 23:59

1. Применить метод главных компонент для понижения размерности данных о результативности деятельности российских вузов. Данные в файле «задание1.xlsx».

а) Определить минимальное количество компонент, которые необходимо использовать для сохранения 75% первоначальной информации.

б) Выписать формулы зависимости главных компонент из пункта а) от первоначальных данных.

2. Задана случайная величина X , распределение которой неизвестно. Из данной случайной величины была получена выборка $x = \{3, 5, 7, 9, 9, 10, 11, 13, 17, 19\}$. Найдите оценки $f(x)$ в точках 5 и 10, используя:

а) Ядро Епанечникова

б) Гауссово ядро

Решите данную задачу без использования статистических пакетов

3. Используйте набор данных «задание3.xlsx», содержащий основные характеристики социально-экономического развития регионов России. Выберите 3-4 переменные, которые могут использоваться для классификации регионов. Обоснуйте свой выбор. Осуществите классификацию методом k -средних. Попробуйте несколько вариантов классификации с разным количеством кластеров. Определите наилучшее разбиение. Опишите состав кластеров. Реализуйте алгоритм иерархической кластеризации. Сравните результаты с разбиением, полученным на основе метода k -средних.

4. Используйте набор данных «задание4.xlsx». Постройте ядерные оценки плотности, используя ядро Гаусса, ядро Епанечникова, а также треугольное ядро. Постройте графики. Проинтерпретируйте результаты.

5. Используйте набор данных «Weekly» из пакета «ISLR». Постройте модель линейного дискриминантного анализа, используя в качестве единственного предиктора переменную «Lag2». Проинтерпретируйте полученные результаты.