Практическая работа №2: Подготовка статистических данных для обработки в среде R

Цель работы

Научится использовать инструменты R для подготовки к обработке статистических данных.

Основные теоритические положения

Рассмотрим традиционный способ представления результатов эксперимента – матрицу данных. Пусть исследователь располагает совокупностью из N наблюдений над состоянием исследуемого явления. При этом явление описано набором из n характеристик, значения которых тем или иным способом измерены в ходе эксперимента. Данные характеристики носят название признаков, показателей или параметров. Такая информация представляется в виде двухмерной таблицы чисел $\mathbf X$ размерности $N \times n$ или в виде матрицы X размерности $N \times n$: $$ X = \begin{pmatrix} x_{11} & \ldots & x_{1j} & \ldots & a_{1n}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ x_{i1} & \ldots & x_{ij} & \ldots & a_{in}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ x_{N1} & \ldots & x_{Nj} & \ldots & a_{Nn}\\ \end{pmatrix} $$ Строки матрицы X соответствуют наблюдениям или, другими словами, объектам наблюдения. В качестве объектов наблюдения выступают, например: в социологии – респонденты (анкетируемые люди), в экономике – предприятия, виды продукции и т. д. Столбцы матрицы X соответствуют признакам, характеризующим изучаемое явление. Как правило, это наиболее легко измеряемые характеристики объектов. Например, предприятие характеризуется численностью, стоимостью основных фондов, видом выпускаемой продукции и т. д. Очевидно, что элемент X_{ij} представляет собой значение признака j, измеренное на объекте i. Часто матрица данных X приводится к стандартной форме следующим преобразованием (для элементов матрицы в стандартной форме используется обозначение x'): $$ x'_j = \frac{x_{ij} - \bar x_j}{\sigma_j}, $$ $$ \bar x_j = \frac1N\sum_{i = 1}^Nx_{ij}, $$ $$ \sigma_j^2 = \frac1N\sum_{i = 1}^N(x_{ij} - \bar x_j)^2, \, i = 1..N, \, j = 1..n, $$ где $\bar x_j$, σ_j^2 – среднее и дисперсия по столбцу с номером j, после которого стандартная матрица X' обладает следующими свойствами: $$ \overline{x'_j} = \frac1N\sum_{i = 1}^Nx'_{ij} = 0, $$ $$ (\sigma'_j)^2 = \frac1N\sum_{i = 1}^N(x'_{ij})^2 = 1, \, i = 1..N, \, j = 1..n. $$ Зачастую признаки, описывающие некоторый объект, имеют существенно различный физический смысл. Это приводит к тому, что величины в различных столбцах исходной матрицы трудно сопоставлять между собой, например, *килограмм* и *метр*. Поэтому получение стандартизованной матрицы можно понимать как приведение всех признаков к некоторой единой условной физической величине, выраженной в одних и тех же условных единицах.

Общая формулировка задачи

Выбрав набор данных, одобренный преподавателем, подготовить данные для последующей

работы с помощью инструментов R.

Порядок выполнения работы

1. Импортировать данные из файла в рабочее пространство R.
2. Разбить общий файл данных на группы файлов с одинаковым номером класса.
3. Подготовить файлы для обучения классификации и проверочные файлы двумя способами: первый способ подразумевает формирование файла обучения из первой половины выборок по всем классам, а контрольный файл сформировать из второй половины данных. Второй способ подразумевает сформировать файлы для обучения и контроля на основе данных с четными и нечетными номерами.
4. Подготовить аналогичные файлы с использованием центрирования и нормировки данных.
5. Построить графики зависимостей значений признаков (полигон) для всех классов и для каждого по отдельности.
6. Построить гистограммы для каждого параметра для всех классов и для каждого по отдельности
7. С помощью функции *summary()* вывести на экран описательную статистику для всех классов и для каждого по отдельности. Объяснить результаты.
8. Построить диаграммы размахов (“ящик с усами”) для всех классов и для каждого по отдельности.
9. Построить матрицы корреляций для всех классов и для каждого класса по отдельности.