Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Многомерный статистический анализ

Отчет по лабораторной работе

Тема: Классификация объектов на основе дискриминантного анализа

Выполнил:
студент гр. 3630102/60401
Камалетдинова Ю. А.

Проверил:
к. ф-м. н., доцент
Павлова Л. В.

Санкт-Петербург
2020

Содержание

Постановка задачи

В данной работе рассматривается задача классификации объектов на основе дискриминантного анализа. . Пусть есть некоторые популяции W_1, W_2. Тогда задача состоит в отнесении объекта w на основе вектора признаков $X = x_1, \ldots, x_p$. Необходимо построить и протестировать классификатор на разных данных:

1. Модельные данные
2. Данные из репозитория
   `https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/`

1. Ход работы

1.1. Построение дискриминантной функции

Сгенерируем обучающие выборки, представляющие популяции W_1, W_2, в предположении, что выборки имеют многомерное нормальное распределение:

$$x \sim N\left(\mu_1, \sum\nolimits_1\right), \;\; x \sim N\left(\mu_2, \sum\nolimits_2\right)$$

, где $\mu_1, \mu_2 \in R^p$ — векторы средних, $\sum_1 = \sum_2 \in R^p \times R^p$ — матрицы ковариаций.

Дискриминантная функция имеет вид

$$z(x) = \alpha_1 x_1 + \cdots + \alpha_p x_p, \tag{1}$$

где вектор коэффициентов $\alpha \in R^p$ находится по формуле:

$$\alpha = \sum\nolimits^{-1} (\mu_1 - \mu_2) \tag{2}$$

Наблюдение из x_i из выборки будем относить к W_1, если $z \geq c$, и к W_2, если $z < c$, где c — постоянная.

Константу c будем вычислять по формулам:

$$c = \frac{(\xi_1 - \xi_2)^2}{\sigma_z^2}, \tag{3}$$

где $\xi_1 = \sum_{j=1}^{p} \alpha_j \mu_{1_j}, \quad \xi_2 = \sum_{j=1}^{p} \alpha_j \mu_{2_j}$ — средние z для $x \in W_1, W_2$ соответственно, $\sigma_z^2 = \sum_{m=1}^{p} \sum_{j=1}^{p} \alpha_m \sigma_{mj} \alpha_j$ — дисперсия

Для измерения "расстояния" между двумя популяциями воспользуемся расстоянием Махалонобиса:

$$\Delta^2 = \frac{(\xi_1 - \xi_2)^2}{\sigma_z^2} \tag{4}$$

Итак, для решения задачи требуется найти коэффициенты $\alpha_1, \ldots, \alpha_p$, минимизарующие Δ^2. Для классификации вектора X будем использовать следующее правило:

$$\begin{aligned} x_i \in W_1 : \sum_{i=1}^{p} \alpha_i x_i \geq \frac{\xi_1 + \xi_2}{2} + \ln \frac{q_2}{q_1} \\ x_i \in W_2 : \sum_{i=1}^{p} \alpha_i x_i < \frac{\xi_1 + \xi_2}{2} + \ln \frac{q_2}{q_1}, \end{aligned} \tag{5}$$

где $P(x_i \in W_1) = q_1, \;\; P(x_i \in W_2) = q_2$ — априорные вероятности

1.2. Численный эксперимент

Проведем два эксперимента на смоделированных данных: в случае, когда данные "хорошо" разделимы и наоборот. Будем считать, что данные "хорошо" разделимы, когда их диапазоны значений векторов средних μ не пересекаются с учетом дисперсий.

Смоделируем две обучающие выборки объемами n_1, n_2 из нормального трехмерного распределения с заданными параметрами: $X_1 \sim N(\mu_1, \sum_1), \; X_2 \sim N(\mu_2, \sum_2), \; X_1, X_2 \in R^3$. Сгенерируем тестовую выборку, распределенную по трехмерному нормальному распределению, включающую наблюдения из обеих популяций. Объемы выборок $n_1 = n_2 = 150$. Приведем значения $\mu_1, \; \mu_2, \; \sum$ для "хорошо" разделимых данных

1.862	0.943	0.007
0.943	0.521	-0.061
0.007	-0.0614	0.108

Таблица 1: Матрица ковариаций для "хорошо" разделимых данных

μ_1	3.727	3.324	3.204
μ_2	1.365	1.235	1.030

Таблица 2: Векторы средних для "хорошо" разделимых данных

Сформируем тестовую выборку из смеси данных, распределенных как исходные популяции W_1, W_2. Зададим априорные вероятности $q_1 = 0.4,\ q_2 = 0.6$. Объем тестовой выборки $n_{test} = 100$. Представим результат классификации в виде матрицы ошибок:

Real \ Pred	W_1	W_2
W_1	40	0
W_2	0	60

Таблица 3: Матрица ошибок для "хорошо" разделимых данных

Исходя из заданных $q_1,\ ,q_2,\ n_{test}$ получаем, что в тестовой выборке 40 наблюдений популяции W_1 и 60 наблюдений популяции W_2. По результатам таблицы 3 делаем вывод, что все объекты были безошибочно классифицированы.

Сгенерируем "плохо" разделимые данные: значения векторов средних μ пересекаются. Тестовую выборку будем формировать как раньше: $q_1 = 0.4,\ q_2 = 0.6,\ n_{test} = 100,\ n_1 = n_2 = 150$. Приведем параметры для генерации данных и матрицу ошибок в виде таблиц:

1.658	-0.333	1.179
-0.333	1.683	-0.756
1.179	-0.756	1.101

Таблица 4: Матрица ковариаций для "плохо" разделимых данных

μ_1	1.366	1.514	1.272
μ_2	1.268	1.126	1.083

Таблица 5: Векторы средних для "плохо" разделимых данных

Real \ Pred	W_1	W_2
W_1	20	20
W_2	10	50

Таблица 6: Матрица ошибок для "плохо" разделимых данных

По результатам из таблицы 3 виидим, что из 100 наблюдений тестовой выборки только 70 удалось классифицировать верно.

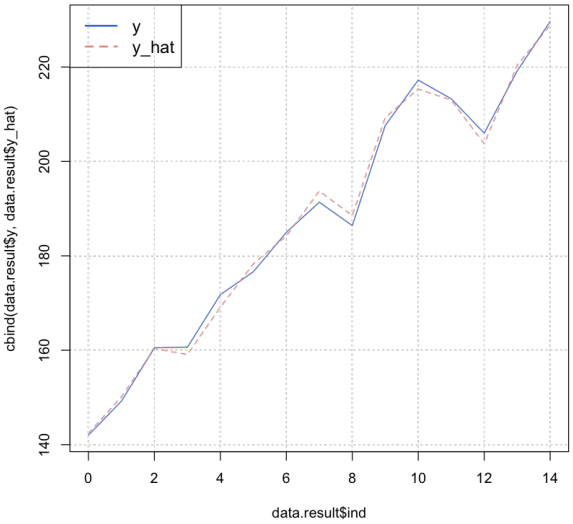

Рис. 1: Оцененные и реальные значения редуцированной (r) модели

Заключение

Список литературы