Работа №2: Классификация многомерных объектов при наличии обучающей выборки

(Классификация объектов на основе дискриминантного анализа)

a) построение и тестирование классификатора с использованием модельных данных;
b) построение и тестирование классификатора с использованием данных из репозитория;

План проведения численных экспериментов

I. Работа с модельными данными

1. Моделируете ОВ1 – выборку заданного объема n_1 из нормального **трехмерного** распределения, т.е. моделируете последовательность векторов

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}^{(1)}, \sum_1);$$

моделируете ОВ2, объема n_2 из нормального **трехмерного** распределения, т.е. моделируете последовательность векторов

$$\boldsymbol{x} \sim N\left(\boldsymbol{\mu}^{(2)}, \sum_2\right);$$

моделируете тестовую выборку (ТВ) (или тестовую последовательность) заданного объема n, представленную смесью (других) объектов из **двух** предыдущих нормальных **трехмерных** распределений; $\boldsymbol{q_1}$ и $\boldsymbol{q_2}$ в этом случае определяются как доля объектов соответствующих популяций.

$\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)} \in R^p, \quad \sum_1 = \sum_2 = \sum \in R^p \times R^p$ **– заданные вами значения параметров распределений, *p*=3.**

2. Вычисляете выборочные оценки параметров распределения и (см. лек.,стр.8):
3. Далее строите классификатор, используя вместо значений параметров распределения – их оценки (подробно - см. лек.):

$$\text{ОВ1}: n_1 \qquad \boldsymbol{\mu}^{(1)} \to \hat{\boldsymbol{\mu}}^{(1)}$$
$$\text{ОВ2}: n_2 \qquad \boldsymbol{\mu}^{(2)} \to \hat{\boldsymbol{\mu}}^{(2)}$$

$$\hat{\mu}_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^{(k)}, k = 1,2$$

$$\boldsymbol{\Sigma} \to S: \quad S = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S^{(1)} + (n_2 - 1)S^{(2)}\right]$$

$$S^{(k)} = \left(s_{lj}^{(k)}\right), l, j = \overline{1,p}, k = 1,2$$

$$s_{lj}^{(k)} = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (x_{il}^{(k)} - \hat{\mu}_l^{(k)})(x_{ij}^{(k)} - \hat{\mu}_j^{(k)}), k = 1,2$$

$$\boldsymbol{\alpha} = \sum^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) \to \hat{\boldsymbol{\alpha}} = \boldsymbol{a} = S^{-1}(\hat{\boldsymbol{\mu}}^{(1)} - \hat{\boldsymbol{\mu}}^{(2)})$$

$$\text{ОВ1}: \quad z_i^{(1)} = a_1 x_{i1}^{(1)} + \cdots + a_p x_{ip}^{(1)}, \qquad i \in \overline{1, n_1} \quad ; \quad \xi_1 \to \bar{z}_1 = \frac{1}{n_1} \sum_{i=1}^{n} z_i^{(1)}$$

$$\text{ОВ2: } z_i^{(2)} = a_1 x_{i1}^{(2)} + \cdots + a_p x_{ip}^{(2)}, \qquad i \in \overline{1, n_2} \quad \xi_2 \to \bar{z}_2 = \frac{1}{n_2}\sum_{i=1}^{n} z_i^{(2)},$$

$$\sigma_z^2 \to s_z^2 = \sum_{l=1}^{p}\sum_{j=1}^{p} a_l s_{lj} a_j$$

И далее:

$$\boldsymbol{x} \to D_1 : \sum_{j=1}^{p} a_j x_j \geq \frac{(\bar{z}_1 + \bar{z}_2)}{2} + ln\frac{q_2}{q_1}$$

$$\boldsymbol{x} \to D_2 : \sum_{j=1}^{p} a_j x_j < \frac{(\bar{z}_1 + \bar{z}_2)}{2} + ln\frac{q_2}{q_1}$$

Оценки расстояния Махаланобиса (смещенная и несмещенная):

$$D^2 = \frac{(\bar{z}_1 - \bar{z}_2)^2}{s_z^2}$$

$$D_H^2 = \frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2} D^2 - p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Оцениваете вероятности ошибочной классификации: см. лек., стр. 11.

4. Работа с ТВ: классифицируете векторы ТВ, результаты приводите в виде Таблицы сопряженности.
5. Анализируете полученные результаты .

Замечание. Численные эксперименты можно проводить для различных модельных данных –«хорошо» разделенных, «плохо» разделенных и т.д.
Диапазон значений n_2 , n_1 , n : 100÷1000.
$\boldsymbol{q_1}$ и $\boldsymbol{q_2}$ можно определять как долю объектов ОВ1 и ОВ2, соответственно.
Любые доп.исследования приветствуются!!!

II. РАБОТА С ДАННЫМИ ИЗ РЕПОЗИТОРИЯ

В репозитории находите данные о кредитовании физических лиц (https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/);

В файле german.doc найдете описание данных, сами данные – в файле german.data-numeric.

Надеюсь, с извлечением данных проблем не возникнет, отлично разберетесь без меня.

Данные – многомерные, каждый объект характеризуют 24 признака, объектов – 1000.
Итак, каждый объект представлен *вектор - строкой* признаков.

Последний столбец – это метка!!, каждый из объектов отнесен к одному из двух классов.

Вы **сами формируете** ОВ1, ОВ2 и ТВ.

Можно взять другие многомерные данные из этого репозитория.

Очевидно, что в ОВ1 должны попасть объекты с одной и той же меткой,

В ОВ2 - с одной и той же, но другой! меткой.

Далее работаете с этими данными по описанной выше схеме.

III. Классификация тех же данных (модельных и из репозитория) после перехода к главным компонентам (PCA).

IV. Общие ВЫВОДЫ

Удачи!