Илясов Семён Алексеевич 4 вариант, 20930гр

Расчетное задание по математической статистике. Часть 4

4.5. Задание

По заданному набору данных двух классов и одной переменных (приложение 2):

4.1. Построить байесовский классификатор в предположении, что переменная распределена нормально для обоих классов (5 баллов).
4.2. Оценить точность классификации при случайном разбиении выборки на обучающую (80%) и контрольную выборки (20%) (5 баллов).

Замечание: программе необходимы файлы: iris.data, приложенные в письме. В этих файлах содержатся таблицы с данными, необходимыми для решения. Их необходимо поместить в ту же директорию, из которой запускается сама программа.

Данные из таблицы вариантов для 4 варианта:

```
X = "Iris-versicolor”
Y = "Iris-setosa"
TYPE = "sepal_length"
```

4.1

В этом пункте нужно построить Байесовский классификатор для классов с нормально распределённой переменной.

Алгоритм:

1) Т.к. распределение нормальное:

$p_1(x) \sim N(a_1, \sigma_1)$ и $p_2(x) \sim N(a_2, \sigma_2)$, где:

x – переменная, которую мы хотим классифицировать

a_1, a_2 – матожидание для 1 и 2 классов,

σ_1, σ_2 – стандартное отклонение для 1 и 2 классов соответственно.

2) Найдём оценки параметров моделей нормального распределения по методу максимального правдоподобия:

$$a_1 = \mu_1 = \frac{1}{n_1} \sum_{i:Y(i)=1} x_i, \qquad a_2 = \mu_2 = \frac{1}{n_2} \sum_{i:Y(i)=2} x_i,$$

$$\sigma_1^2 = \widehat{\sigma_1}^2 = \frac{1}{n_1} \sum_{i:Y(i)=1} (x_i - \mu_1)^2, \qquad \sigma_2^2 = \widehat{\sigma_2}^2 = \frac{1}{n_2} \sum_{i:Y(i)=2} (x_i - \mu_2)^2$$

где n_1, n_2 – число объектов в 1 и 2 выборке соответственно.

3)Байесовский классификатор:

$C = argmax(\boldsymbol{f}n(x, \mu 1, \widehat{\ \ 1}), \boldsymbol{f}n(x, \mu 2, \widehat{\ \ 1}))$

$\boldsymbol{f}n(x, \mu, \widehat{\ \ })$ - ф-я плотности нормального распределения.

C − класс, к которому классификатор отнес переменную x.

$\mu 1$ == 5.9799

$\mu 2$ == 5.0325

$\widehat{\ 1}$ == 0.2766

$\widehat{\ 1}$ == 0.1211

$C = argmax(\boldsymbol{f}n(x, 5,9799, 0.2766), \boldsymbol{f}n(x, 5.0325, 0.1211))$

4.2

В этом пункте нужно оценить точность классификации(accuracy) на случайном разбиении данных на тренировочную и контрольную выборки в соотношении тренировочная/контрольная = 4/1 (80%/20%)

Алгоритм:

1) Посчитаем количество верноположительных результатов:

Positive = $\sum i = 1 \ \{1| C\, i = 1\}$ = 8

2) Посчитаем количество верноотрицательных результатов:

Negative = $\sum i = 1 \ \{1 | C\, i \ = \ 2\}$ = 10

3) Посчитаем accuracy:

accuracy = (Negative + Positive) / N =

= (10 + 8)/20 = 0.9 = 90%

N – кол-во объектов в контрольной выборке

Результат:

```
m1 ==  5.9799999999999995 , m2 ==  5.032500000000001
var1 ==  0.2766 , var2 ==  0.12119374999999999
Result of prediction for Positive:  8
Result of prediction for Negative:  10
classifier accuracy = 90.0%
```