Автор:

Сахибназарова Виктория Бахтиёровна

магистрант

ФГАОУ ВО «Самарский государственный

аэрокосмический университет

им. академика С.П. Королёва (НИУ)»

г. Самара, Самарская область

РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ ДАННЫХ ПРИ ПОМОЩИ ИСПОЛЬЗОВАНИЯ РЕГРЕССИОННОЙ МОДЕЛИ

Аннотация: *в данной работе составлена и реализована в виде программного продукта математическая модель линейной регрессии, решающая задачу классификации. Автором также производится анализ влияния параметров модели на результат классификации.*

Ключевые слова: *задача классификации, регрессионная модель, линейная регрессия.*

В качестве исходных данных мы имеем выборку из значительного количества наблюдений зависимости интересующего нас результирующего значения от ряда других значений, но не имеем функции, однозначно выражающую данную зависимость. Для того чтобы предположить, какое результирующее значение мы получим из очередного наблюдения, можно использовать регрессионную модель.

В качестве математической модели для данной работы была выбрана линейная регрессия:

$$y = f(x, b) + \varepsilon,$$

где b – параметры модели, ε – случайная ошибка модели, $f(x, b)$ – функция регрессии, имеющая вид:

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon,$$

где x_i – факторы модели, k – количество факторов модели.

Также линейная регрессия может быть представлена в матричном виде $y = Xb + \varepsilon$, $X \in \mathbb{R}^{n \times k}$, $b \in \mathbb{R}^{k \times 1}$, где n – количество наблюдений [1]:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$$ – вектор наблюдений зависимой переменой y

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$ – матрица факторов

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$ – вектор случайных ошибок

$$b = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_k \end{pmatrix}$$ – вектор коэффициентов.

При представлении регрессии в матричной форме вектор коэффициентов b будут высчитываться по следующей формуле, исходя из метода наименьших квадратов:

$$b = (X^T X)^{-1} X^T y.$$

Значения погрешности рассчитываются по формуле:

$$\Delta = y_k - y'_k$$

где y_k – значение качества, взятое из исходных данных, y'_k – прогнозируемое значение качества.

В качестве исходного выбран набор данных, содержащих информацию о характеристиках и соответствующих им уровнях качества красного вина. Количество исходных характеристик равно 10. Уровень качества варьируется от 0 до 10.

Описанная выше математическая модель реализована в программном продукте. После считывания данных по обучающей выборке происходит расчет коэффициентов b линейной регрессии, затем прогнозируются значения качества для тестовой выборки, и вычисляется разница между прогнозируемыми и исходными значениями качества. На экран выводятся графики разностей исходных и

прогнозируемых значений качества, а также график оценки прогнозирования [2] для разного количества учитываемых характеристик.

Как видно из рисунка 5 и таблицы 1, наибольшее расхождение между прогнозируемым значением качества и значением из исходных данных можно наблюдать при учете 6 признаков из 10. Наилучшая оценка достигается при учете 8 признаков, что позволяет сделать вывод, что точность прогноза зависит от количества и последовательности наблюдаемых признаков. Также из сравнения таблиц 1 и 2 следует, что оценка зависит и от объема тестовой выборки.

Рис. 1. График зависимости оценки от количества признаков

Таблица 1

Оценки для тестовой выборки объемом 400 наблюдений

Количество признаков	Оценка
4	0,00097140398027
5	54,66706371937510
6	249,79184290830900
7	0,01001580458585
8	0,00764995517674
9	0,45022525995150
10	0,20398732467449

Таблица 2

Оценки для тестовой выборки объемом 600 наблюдений

Количество признаков	Оценка
4	0,00071851540525
5	60,76292402607450
6	310,85297856234900
7	0,01009903902687
8	0,00676029152085
9	0,50350479846911
10	0,20680512671192

Список литературы

1. Линейная регрессия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Линейная_регрессия

2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным [Текст]: Учебник. – М.: Наука, 1979. – 448 с.