ЛАБОРАТОРНАЯ РАБОТА 7. НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР. ВИДЫ КЛАССИФИКАТОРОВ

ЦЕЛЬ И ЗАДАЧИ РАБОТЫ, ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ЕЕ ВЫПОЛНЕНИЯ

Целью выполнения лабораторной работы является формирование практических навыков проведения классификации с использованием наивного Байесовского классификатора.

Основными задачами выполнения лабораторной работы являются:

1. Составить модель на основе наивного Байесовского классификатора.
2. Изучить виды наивного Байесовского классификатора.
3. Результатами работы являются:
4. Визуализированные и проанализированные данные классификации.
5. Составленная модель.
6. Подготовленный отчет.

КРАТКАЯ ХАРАКТЕРИСТИКА ОБЪЕКТА ИЗУЧЕНИЯ, ИССЛЕДОВАНИЯ

Задача классификации — задача, в которой имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать (см. ниже) произвольный объект из исходного множества.

Классифицировать объект — значит, указать номер (или наименование) класса, к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

В математической статистике задачи классификации называются также задачами дискриминантного анализа. В некоторых прикладных областях, и даже в самой математической статистике, из-за близости задач часто не различают задачи кластеризации от задач классификации.

Существуют следующие типы классов:

Двухклассовая классификация. Наиболее простой в техническом отношении случай, который служит основой для решения более сложных задач.

Многоклассовая классификация. Когда число классов достигает многих тысяч (например, при распознавании иероглифов или слитной речи), задача классификации становится существенно более трудной.

Непересекающиеся классы.

Пересекающиеся классы. Объект может относиться одновременно к нескольким классам.

Нечёткие классы. Требуется определять степень принадлежности объекта каждому из классов, обычно это действительное число от 0 до 1.

НАИВНЫЙ БАЙЕСОВСКИЙ КЛАССИФИКАТОР

Наивный байесовский классификатор — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости.

В зависимости от точной природы вероятностной модели, наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практических приложениях для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью, не веря в байесовскую вероятность и не используя байесовские методы.

Несмотря на наивный вид и, несомненно, очень упрощенные условия, наивные байесовские классификаторы часто работают намного лучше во многих сложных жизненных ситуациях.

Достоинством наивного байесовского классификатора является малое количество данных необходимых для обучения, оценки параметров и классификации.

Байесовский подход к классификации основан на теореме, утверждающей, что если плотности распределения каждого из классов известны, то искомый алгоритм можно выписать в явном аналитическом виде. Более того, этот алгоритм оптимален, то есть обладает минимальной вероятностью ошибок.

На практике плотности распределения классов, как правило, не известны. Их приходится оценивать (восстанавливать) по обучающей выборке. В результате байесовский алгоритм перестаёт быть оптимальным, так как восстановить плотность по выборке можно только с некоторой погрешностью. Чем короче выборка, тем выше шансы подогнать распределение под конкретные данные и столкнуться с эффектом переобучения.

Байесовский подход к классификации является одним из старейших, но до сих пор сохраняет прочные позиции в теории распознавания. Он лежит в основе многих достаточно удачных алгоритмов классификации.

Модель наивного байесовского классификатора

Вероятностная модель для классификатора — это условная модель:

$$p(C \mid F_1, \dots, F_n)$$

над зависимой переменной класса C с малым количеством результатов или классов, зависимая от нескольких переменных $F_1, \dots, F_n$. Проблема заключается в том, что когда количество свойств n очень велико или когда свойство может принимать большое количество значений, тогда строить такую модель на вероятностных таблицах становится невозможно. Поэтому мы переформулируем модель, чтобы сделать её легко поддающейся обработке. Используя теорему Байеса, запишем

$$p(C \mid F_1, \dots, F_n) = \frac{p(C)\, p(F_1, \dots, F_n \mid C)}{p(F_1, \dots, F_n)}$$

На практике интересен лишь числитель этой дроби, так как знаменатель не зависит от C и значения свойств F_i даны, так что знаменатель — константа. Числитель эквивалентен совместной вероятности модели

$$p(C, F_1, \dots, F_n)$$

которая может быть переписана следующим образом, используя повторные приложения определений условной вероятности:

$$\begin{aligned} p(C, F_1, \dots, F_n) &= p(C)\, p(F_1, \dots, F_n \mid C) \\ &= p(C)\, p(F_1 \mid C)\, p(F_2, \dots, F_n \mid C, F_1) \\ &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1) p(F_3, \dots, F_n \mid C, F_1, F_2) \\ &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C, F_1) \dots p(F_n \mid C, F_1, F_2, F_3, \dots, F_{n-1}) \end{aligned}$$

и т. д. Теперь можно использовать «наивные» предположения условной независимости: предположим, что каждое свойство F_i условно независимо от любого другого свойства F_j при $j \neq i$. Это означает:

$$p(F_i \mid C, F_j) = p(F_i \mid C)$$

таким образом, совместная модель может быть выражена как:

$$\begin{aligned} p(C, F_1, \dots, F_n) &= p(C)\, p(F_1 \mid C)\, p(F_2 \mid C) p(F_3 \mid C) \dots p(F_n \mid C) \\ &= p(C) \prod_{i=1}^{n} p(F_i \mid C) \end{aligned}$$

Это означает, что из предположения о независимости, условное распределение по классовой переменной C может быть выражено так:

$$p(C \mid F_1, \ldots, F_n) = \frac{1}{Z}\, p(C) \prod_{i=1}^{n} p(F_i|C)$$

где $Z=p(F_1, \ldots, F_n)$— это масштабный множитель, зависящий только от $F_1, \ldots, F_n$, то есть константа, если значения переменных известны.

Оценка параметров

Все параметры модели могут быть аппроксимированы относительными частотами из набора данных обучения. Это оценки максимального правдоподобия вероятностей. Непрерывные свойства, как правило, оцениваются через нормальное распределение. В качестве математического ожидания и дисперсии вычисляются статистики — среднее арифметическое и среднеквадратическое отклонение соответственно.

Если данный класс и значение свойства никогда не встречаются вместе в наборе обучения, тогда оценка, основанная на вероятностях, будет равна нулю. Это проблема, так как при перемножении нулевая оценка приведет к потере информации о других вероятностях. Поэтому предпочтительно проводить небольшие поправки во все оценки вероятностей так, чтобы никакая вероятность не была строго равна нулю.

Построение классификатора по вероятностной модели

Наивный байесовский классификатор объединяет модель с правилом решения. Одно общее правило должно выбрать наиболее вероятную гипотезу; оно известно, как апостериорное правило принятия решения (MAP). Соответствующий классификатор — это функция classify определённая следующим образом:

$$classify(f_1, \ldots, f_n) = arg \underbrace{max}_{c}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i \,|C = c)$$

ВИДЫ НАИВНЫХ БАЙЕСОВСКИХ КЛАССИФИКАТОРОВ

В библиотеке sklearn реализовано несколько видов наивного Байесовского классификатора:

- Гауссовский наивный Байес
- Полиномиальный наивный Байес
- Дополненный (complement) наивный Байес
- Наивный Байес Бернулли

Рассмотрим каждый из них.

Gaussian Naive Bayes

GaussianNB реализует алгоритм Гауссовского наивного Байеса для классификации. При этом предполагается, что данные имеют Гауссовское распределение:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}}$$

Параметры σ_y и μ_y оцениваются методом максимального правдоподобия.

Multinomial Naive Bayes (MNB)

MultinomialNB реализует алгоритм наивного Байеса для полиномиально распределенных данных и является одним из двух классических наивных байесовских вариантов, используемых в классификации текста (где данные, как правило, представлены в виде векторов счетчиков слов). Распределение параметризовано векторами $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ для каждого класса y, где n – число признаков (в классификации текста – размер набора слов), а θ_{yi} – вероятность $P(x_i \mid y)$ появления признака i в объекте, относящемуся к классу y.

Параметры θ_{yi} оцениваются при помощи относительного подсчета частот:

$$\widehat{\theta_{yi}} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Здесь $N_{yi} = \sum_{x \in T} x_i$ – количество появлений признака i в объекте класса y в обучающей выборке T, а $N_y = \sum_{i=1}^{n} N_{yi}$ – общее число всех признаков класса y.

Приоритеты сглаживания $\alpha \geq 0$ учитывают признаки, отсутствующие в обучающих примерах, и предотвращают нулевые вероятности в дальнейших вычислениях. Установка параметра $\alpha = 1$ называется сглаживанием Лапласа, в то время как установка $\alpha < 1$ называется сглаживанием Лидстона.

Bernoulli Naive Bayes

BernoulliNB реализует алгоритм наивного Байеса для классификации данных, которые имеют многомерное распределение Бернулли; то есть, могут обладать несколькими признаками, но при этом предполагается, что каждый из них является бинарной переменной. Следовательно, этот класс требует, чтобы выборки были представлены в виде векторов объектов с двоичными числами. Если передаются данные другого типа, BernoulliNB может преобразовать свой вход в двоичную форму (в зависимости от значения параметра binarize).

Принятие решения для наивного Байеса Бернулли основывается на формуле:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

которая отличается от правила полиномиального наивного Байеса тем, что она явно “штрафует” за отсутствие элемента, который является признаком класса, в то время как полиномиальный наивный Байес просто проигнорировал бы отсутствующий элемент.

Complement Naive Bayes

ComplementNB реализует дополненный наивный байесовский (CNB) алгоритм. CNB - это адаптация стандартного полиномиального наивного байесовского алгоритма (MNB), который в частности подходит для несбалансированных наборов данных. В частности, CNB использует статистику из дополнения каждого класса для вычисления весов модели. Изобретатели CNB эмпирически показывают, что оценки параметров для CNB более стабильны, чем оценки для MNB. Кроме того, CNB регулярно опережает MNB (часто со значительным

отрывом) по задачам классификации текста. Процедура расчета весов выглядит следующим образом:

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{ij}}$$

$$w_{ci} = log\hat{\theta}_{ci}$$

$$w_{ci} = \frac{w_{ci}}{\sum_j |w_{ci}|}$$

где суммирование осуществляется по всем документам j, не относящимся к классу C, d_{ij} является значением счетчика или значением tf-idf термина i в документе j, α_i - это сглаживающий гиперпараметр, подобный найденному в MNB, и $\alpha_i = \sum_i \alpha_i$. Вторая нормализация учитывает тенденцию(склонность) более длинных документов оказывать доминирующее воздействие на оценки параметров в MNB. Правило классификации:

$$\hat{c} = arg \underbrace{min}_{c} \sum_i t_i w_{ci}$$

то есть документ присваивается классу, который является самым плохим соответствием дополнения.

ОБРАЗЕЦ ВЫПОЛНЕНИЯ ЗАДАНИЯ

```
from sklearn.naive_bayes import GaussianNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.datasets import load_digits
from matplotlib import pyplot as plt
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns;


digits = load_digits()

# разделение данных на обучающие и тестовые
X_train, X_test, y_train, y_test = train_test_split(digits.data,
digits.target)

# обучение модели
clf = GaussianNB()
clf.fit(X_train, y_train)

# прогнозы модели на тестовых данных
predicted = clf.predict(X_test)
expected = y_test

# вывод цифр (зеленый - совпадение, красный - нет)
fig = plt.figure(figsize=(6, 6))
fig.subplots_adjust(left=0, right=1, bottom=0, top=1, hspace=0.05,
wspace=0.05)

for i in range(64):
    ax = fig.add_subplot(8, 8, i + 1, xticks=[], yticks=[])
    ax.imshow(X_test.reshape(-1, 8, 8)[i], cmap=plt.cm.binary,
              interpolation='nearest')

    if predicted[i] == expected[i]:
        ax.text(0, 7, str(predicted[i]), color='green')
    else:
        ax.text(0, 7, str(predicted[i]), color='red')

# построение матрицы несоответствий
fig = plt.figure(figsize=(10, 10))
mat = confusion_matrix(predicted, expected)
names = np.unique(predicted)
sns.heatmap(mat, square=True, annot=True, fmt='d', cbar=True,
            xticklabels=names, yticklabels=names)
```

```
    plt.xlabel('Реальные значения')
    plt.ylabel('Данные модели')

    # оцениваем точность модели
    print("Точность модели = {}%".format(round(accuracy_score(expected,
predicted) * 100)))
```

В результате выполнения программы будут получены следующие результаты (рис. 7.1, 7.2):

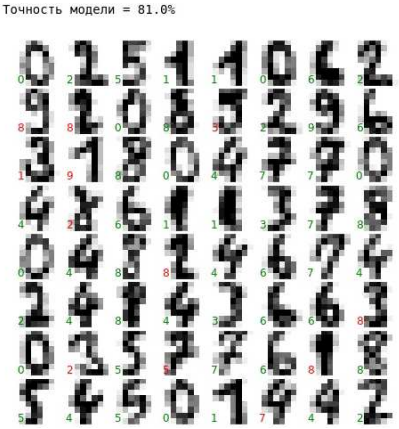

Рис. 7.1 Визуализация результатов прогноза модели (зеленый – верно, красный - неверно)

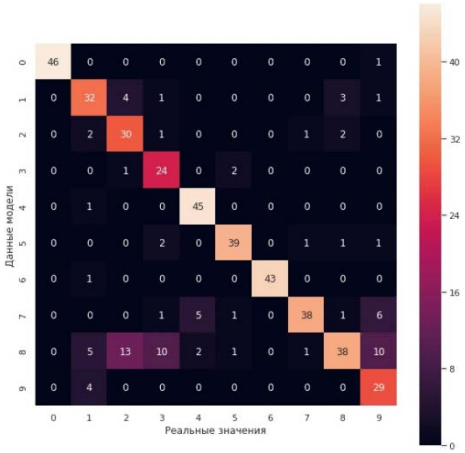

Рис. 7.2 Матрица несоответствия

ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ

Для всех вариантов необходимо провести классификацию с помощью наивного Байесовского классификатора. Составить модель для классификации. Сравнить данные, полученные моделью, с реальными данными. Данные для вариантов необходимо брать из набора sklearn.datasets.

ТРЕБОВАНИЯ К РЕАЛИЗАЦИИ

В качестве результата работы необходимо визуализировать результаты классификации и отобразить их в Jupyter Notebook. Также необходимо проанализировать полученные результаты и сделать выводы.

ВАРИАНТЫ ЗАДАНИЙ

Вариант 1

Используя Гауссовский наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) цветков ириса. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 2

Используя Гауссовский наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) сортов вина. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 3

Используя полиномиальный наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) цветков ириса. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 4

Используя полиномиальный наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) сортов вина. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 5

Используя дополненный (complement) наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) цветков ириса. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 6

Используя дополненный (complement) наивный Байес, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) сортов вина. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 7

Используя наивный Байес Бернулли, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) цветков ириса. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 8

Используя наивный Байес Бернулли, построить модель (на обучающей выборке) и произвести классификацию (на тестовой выборке) сортов вина. Визуализировать полученную классификацию. Построить матрицу несоответствий и интерпретировать результаты. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний).

Вариант 9

Провести классификацию цифр из примера при помощи полиномиального и дополненного (complement) наивного Байеса. Оценить точность каждой модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний для каждого наивного Байеса). Сравнить результаты. В случае значительных отклонений от результатов примера объяснить эти отклонения. Визуализировать полученную классификацию (для одной из моделей). Построить матрицу несоответствий и интерпретировать результаты (для той же модели).

Вариант 10

Провести классификацию цифр из примера при помощи наивного Байеса Бернулли и полиномиального наивного Байеса. Оценить точность каждой модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний для каждого наивного Байеса). Сравнить результаты. В случае значительных отклонений от результатов примера объяснить эти отклонения. Визуализировать полученную классификацию (для одной из моделей). Построить

матрицу несоответствий и интерпретировать результаты (для той же модели).

Вариант 11

Провести классификацию цифр из примера при помощи наивного Байеса Бернулли и дополненного (complement) наивного Байеса. Оценить точность каждой модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний для каждого наивного Байеса). Сравнить результаты. В случае значительных отклонений от результатов примера объяснить эти отклонения. Визуализировать полученную классификацию (для одной из моделей). Построить матрицу несоответствий и интерпретировать результаты (для той же модели).

Вариант 12

Провести классификацию цифр из примера при помощи наивного Байеса Бернулли. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний). Построить график зависимости точности модели от размера обучающей выборки. Сделать выводы.

Вариант 13

Провести классификацию цифр из примера при помощи дополненного (complement) наивного Байеса. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний). Построить график зависимости точности модели от размера обучающей выборки. Сделать выводы.

Вариант 14

Провести классификацию цифр из примера при помощи полиномиального наивного Байеса. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний). Построить график зависимости точности модели от размера обучающей выборки. Сделать выводы.

Вариант 15

Провести классификацию цифр из примера при помощи Гауссовского наивного Байеса. Оценить точность модели (для оценки точности необходимо взять среднее арифметическое оценки 10 испытаний). Построить график зависимости точности модели от размера обучающей выборки. Сделать выводы.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Назовите основную задачу классификации.
2. Дайте определение понятию классификация.
3. Перечислите и охарактеризуйте основные типы классов.
4. Дайте определение понятию наивный Байесовский классификатор.
5. Приведите теорему Байеса.
6. Опишите принцип построения модели наивного Байесовского классификатора.
7. Опишите принцип оценки параметров модели.
8. Опишите принцип построения классификатора по вероятностной модели.
9. Перечислите виды наивного Байесовского классификатора.
10. Опишите Гауссовский наивный Байес.
11. Опишите полиномиальный наивный Байес.
12. Опишите дополненный (complement) наивный Байес.
13. Опишите наивный Байес Бернулли.

ФОРМА ОТЧЕТА ПО ЛАБОРАТОРНОЙ РАБОТЕ

На выполнение лабораторной работы отводится 2 занятия (4 академических часа: 3 часа на выполнение и сдачу лабораторной работы и 1 час на подготовку отчета).

Отчет на защиту предоставляется в печатном виде.

Структура отчета (на отдельном листе(-ах)): титульный лист, формулировка задания, описание процесса выполнения лабораторной работы, результаты выполнения работы, выводы.