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

КУРСОВАЯ РАБОТА

по дисциплине
«Математическая статистика»

Выполнил студент
группы 3630102/80401 Веденичев Дмитрий Александрович

Проверил
Доцент, к.ф.-м.н. Баженов Александр Николаевич

Санкт-Петербург
2021

Содержание

Список иллюстраций

1 Постановка задачи

Есть два набора данных, связанных с красными и белыми вариантами португальского вина "Vinho Verde". Из-за проблем с конфиденциальностью и логистикой доступны только физико-химические (входные данные) и сенсорные (выходные данные) переменные. Данные не сбалансированы, вин среднего качества больше, чем с низкой и высокой оценкой соответственно.

Входные данные представляют собой:

1. фиксированную кислотность
2. летучую кислотность
3. лимонную кислоту
4. остаточный сахар
5. хлориды
6. свободный диоксид серы
7. общий диоксид серы
8. плотность
9. pH
10. сульфаты
11. алкоголь
12. качество вино

В нашей работе необходимо:

1. Считать и обработать полученные наборы данных.
2. Обучить наивный байесовский классификатор с помощью тренировочных данных.
3. Провести тестирование обученного классификатора и исследовать результаты.

2 Теория

Часто бывает необходимо определить принадлежность объекта какому-то классу. Для данной задачи используются инструменты, называемые классификаторами. Одним из них является наивный Байесовский классификатор, который основан на применении теоремы Баейса со строгими (наивными) предположениями о независимости определяющих объект характеристик. В зависимости от природы вероятностной модели, Байесовский наивный классификатор способен показывать результаты лучше чем обучаемые нейросети. Это, в совокупности с возможностью обучаться по малым объемам данных, делает из него серьезный инструмент для решения жизненных задач.

2.1 Модель наивного Байесовского классификатора

Вероятностная модель для классификатора — это условная модель $p(Class|X_1, X_2, .., X_n)$ над зависимой переменной $Class$ с малым количеством значений (классов) от нескольких переменных $X_1, X_2, .., X_n$. Используя теорему Байеса запишем:

$$p(Class|X_1, X_2, .., X_n) = \frac{p(Class)p(X_1, X_2, .., X_n|Class)}{p(X_1, X_2, .., X_n)}$$

На практике интересен лишь числитель, так как знаменатель не зависит от $Class$, и значения свойств X_i даны, так что знаменатель - константа. Числитель эквивалентен совместной вероятности модели $p(Class, X_1, ..., X_n)$, для которой можно выполнить следующие преобразования

$$p(Class, X_1, ..., X_n) = p(Class)p(X_1, .., X_n|Class) = p(Class)p(X_1|Class)p(X_2, ..., X_n|Class, X_1) =$$

$$= p(Class)p(X_1|Class)...p(X_n|Class, X_1, ..., X_{n-1})$$

Используем 'наивное' предположение о независимости переменных, которое позволяет нам сделать преобразование $p(X_i|Class, X_j) = p(X_i|Class), i \neq j$. Тогда полученное выше можно преобразовать в виде

$$p(Class, X_1, ..., X_n) = p(Class)p(X_1|Class)...p(X_n|Class, X_1, ..., X_{n-1}) =$$

$$= p(Class)p(X_1|Class)...p(X_n|Class) = p(Class)\prod_{i=1}^{n} p(X_i|C)$$

Таким образом, из предположения о независимости, условное распределение $Class$ можно выразить в виде

$$p(Class|X_1, X_2, .., X_n) = \frac{1}{Z}p(Class)\prod_{i=1}^{n} p(X_i|C)$$

Где $Z = p(X_1, .., X_n)$ - это коэффициент масштабирования, зависящий только от $X_1, .., X_n$, таким образом это константа, если значения известны.

2.2 Оценка параметров и построенние классификатора по модели

Параметры модели могут быть аппроксимированы относительными частотами из набора данных обучения. Это оценки максимального правдоподобия вероятностей. Непрерывные свойства, обычно оцениваются через нормальное распределение. Для математического ожидания и дисперсии вычисляются среднее арифметическое и среднеквадратическое отклонение.

Наивный Байесовский классификатор объединяет Баейсовскую модель(2.1) с правилом решения. Одно общее правило должно выбирать наиболее вероятную гипотезу, оно также известно как апостериорное правило принятия решения (MAP). Соответствующий наивный Баейсовский классификатор — это функция определённая следующим образом:

$$classify(x_1, ..., x_n) = \underset{Class}{\operatorname{argmax}}\, p(Class = class) \prod_{i=1}^{n} p(X_i = x_i | Class = class)$$

3 Программная реализация

Код программы был написан на языке Python в среде разработки PyCharm. В ходе работы использовались дополнительные библиотеки:

1. seaborn
2. matplotlib
3. numpy
4. pandas
5. sklearn

В приложении находится ссылка на GitHub репозиторий с исходныи кодом.

4 Ход работы

На вход подаются два .csv файла с данными о красном и белом винах. Все последующие этапы описывают взаимодействие с одним из двух файлов. Таким образом рассматривается один тип вина за итерацию работы.

1. Обработка данных. Данные из файла собираются в pandas.dataframe. Выполняется проверка на наличие отсутствующих значений. Составление таблицы корреляции переменных

в датасете. Для упрощения задачи классификации, количество классов с 11 (от 0 до 10) уменьшается до 2. Теперь от 0 до 6.5 будет 'плохое' вино(ему соотвествует 0), от 6.5 до 10 'хорошее' вино(ему соответствует 1). Данное упрощение связано как с недостатком данных для крайних значений, так называемой несбалансированностью, так и с упрощением работы для классификатора.

2. Создание тренировочных и тестовых датасетов. Разбиваем исходные данные на тренировочные и тестовые, используя встроенные функции sklearn. Перед разделением на тестовые и тренировочные данные - датасет перемешивается. Тестовые данные будут составлять 20% от входных данных. Каждый из тренировочных и тестовых датасетов представляет из себя два элемента: матрица, где на строках представлены значения x_i, и вектор, где элементами являются зависимые переменные c_i.

3. Обучить классификатор на тренировочных данных. Тренировочные данные, полученные на предыдущем этапе, используются для обучения классификатора.

4. Проверить классификатор на тестовых данных. Получить предсказания от классифиактора на тестовых данных. Сравнить предсказания с имеющимися ответами. Собрать результаты сравнения.

5. Отобразить метрики. Из собранных данных отобразить результат обучения классификатора.

5 Результаты

5.1 Корреляция переменных в исходных датасетах

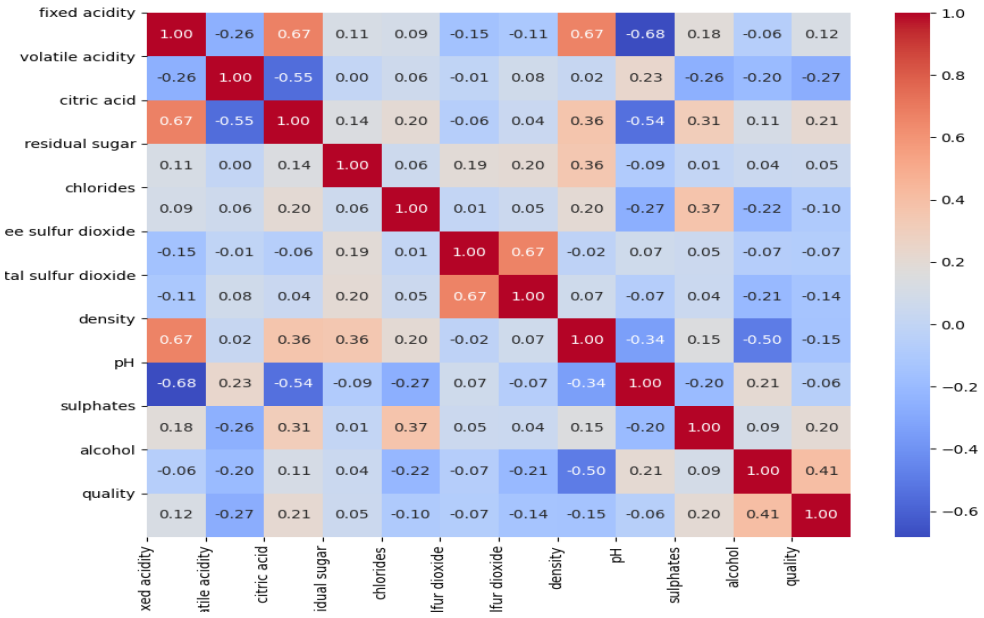

Рис. 1: Корреляция переменных в красном вине

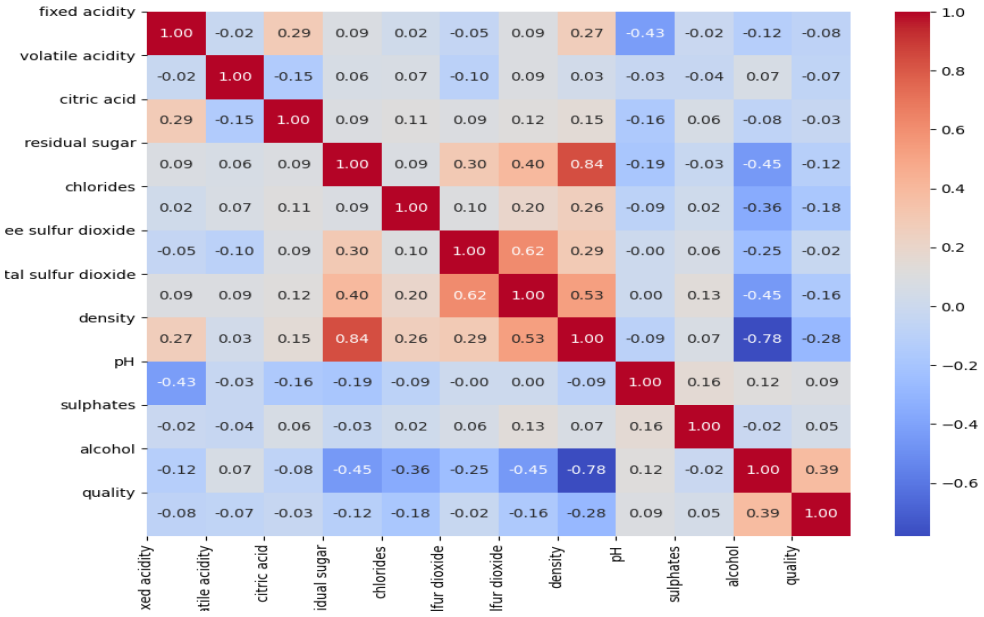

Рис. 2: Корреляция переменных в белом вине

5.2 Результаты классификации наивным Баейсовским классификатором

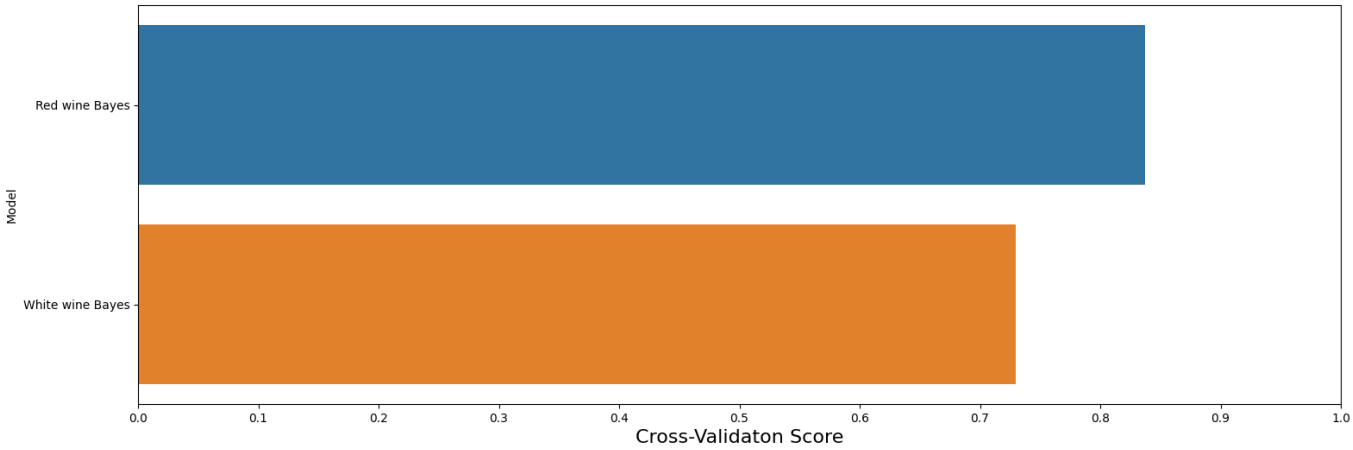

Рис. 3: Результаты перекрестной проверки классификаторов белого и красного вина

Model	True Positive	False Positive	True Negative	False Negative	Accuracy (training)	Accuracy (test)	Cross-Validation
White wine Bayes	539	214	159	68	0.732	0.712	0.730
Red wine Bayes	234	39	37	10	0.839	0.847	0.837

Таблица 1: Матрица ошибок

6 Анализ

При изучении графика (3) и таблицы (1) получаем следующее:
Наивный Баейсовский классификатор показывает хороший результат в классификации красных вин, несмотря на малый размер датасета. Белые вина имеют больший датасет, однако классификатор показывает себя хуже, чем с красных вином.

7 Приложение

Код программы GitHub URL:

https://github.com/PopeyeTheSailorsCat/Bayes_estimator-mat_stat_course-/tree/main/src

Список литературы

[1] Hand, DJ, Yu, K. (2001). «Idiot's Bayes — not so stupid after all?» International Statistical Review. Vol 69 part 3, pages 385—399.

[2] J.P. Sacha. "New synthesis of Bayesian network classifiers and interpretation of cardiac SPECT images Ph.D. Dissertation, University of Toledo, 1999., page 48

[3] Электронный ресурс: https://scikit-learn.org/stable/modules/naive_bayes.html

[4] Электронный ресурс: https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html