Обучение модели логистической регрессии на датасете о винах

1. Введение

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 произвольный объект из исходного множества.

В машинном обучении задача классификации относится к разделу обучения с учителем. Существует также обучение без учителя, когда разделение объектов обучающей выборки на классы не задаётся, и требуется классифицировать объекты только на основе их сходства друг с другом.

2. Описание проблемы и решение

Набор данных относится к белому и красному вариантам португальских вин. Данные содержат

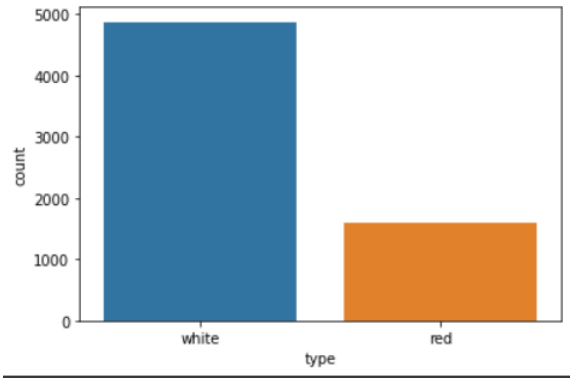
	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

физико-химические и органолептические признаки, по которым вино можно отнести к первому или второму типу. Этот набор данных можно рассматривать для задачи классификации или регрессии. Реализуем задачу по предсказанию качества вина, на основе физико-химического состава.

Цель данного исследование в реализации разных алгоритмов классификации и оценки их выходного качества на предложенном датасете о винах.

3. Признаки и предобработка данных

В задаче классификации данные называются несбалансированными (Imbalanced Data), если в обучающей выборке доли объектов разных классов существенно различаются, также говорят, что «классы не сбалансированы».



Нормализация – это процедура предобработки входной информации (обучающих, тестовых и валидационных выборок, а также реальных данных), при которой значения признаков во входном векторе приводятся к некоторому заданному диапазону, например, [0…1] или [-1…1] [1].

Поскольку наш набор данных несбалансирован (т. Е. Есть функции, которые имеют высокие диапазоны по отношению к другим), мы нормализуем наши значения от 0 до 1.

Разделим данные на два класса, в зависимости от качества вина.

```
data_full['Excellent quality'] = [ 1 if x>=0.765 else 0 for x in data_full.quality]
y = data_full["Excellent quality"]
X = data_full.drop(['Excellent quality' ,'quality'], axis = 1)

X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.8, random_state = 42)
```

4. Обучающие модели

Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

4.1 Модели машинного обучения для задачи классификации на основе признаков:

KNN - Алгоритм K-ближайших соседей (KNN) использует «сходство признаков» для прогнозирования значений новых точек данных, что также означает, что новой точке данных будет присвоено значение на основе того, насколько близко он соответствует точкам в обучающем наборе.

DecisionTreeClassifier - использует дерево решений (как предиктивную модель[en]), чтобы перейти от наблюдений над объектами (представленными в ветвях) к заключениям о целевых значениях объектов (представленных в листьях). Это обучение является одним из подходов моделирования предсказаний, используемых в статистике, интеллектуальном анализе данных и обучении машин. Модели деревьев, в которых целевая переменная может принимать дискретный набор значений, называются деревьями классификации. В этих структурах деревьев листья представляют метки классов, а ветки представляют конъюнкции признаков, которые ведут в эти метки классов. Деревья решений, в которых целевая переменная может принимать непрерывные значения (обычно, вещественные числа) называются деревьями регрессии.

Random forest — алгоритм машинного обучения, предложенный Лео Брейманом[1][2] и Адель Катлер , заключающийся в использовании комитета (ансамбля) решающих деревьев. Алгоритм сочетает в себе две основные идеи: метод бэггинга Бреймана, и метод случайных подпространств (англ.)рус., предложенный Тин Кам Хо . Алгоритм применяется для задач классификации, регрессии и кластеризации. Основная идея заключается в использовании большого ансамбля решающих деревьев, каждое из которых само по себе даёт очень невысокое качество классификации, но за счёт их большого количества результат получается хорошим.

SVM - набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа. Принадлежит семейству линейных классификаторов и может также рассматриваться как частный случай регуляризации по Тихонову. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как *метод классификатора с максимальным зазором*. Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с наибольшим зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. *Разделяющей гиперплоскостью* будет гиперплоскость, создающая наибольшее расстояние до двух параллельных гиперплоскостей. Алгоритм основан на допущении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

5. Результаты

В этом разделе мы сравниваем относительную эффективность трех реализованных моделей обучения.

в нашем наборе данных. В таблице и рисунке показаны ошибка обучения и ошибка проверки различных моделей на
набор данных.

	models	Acurracy score
2	RandomForest	0.975251
3	SVM	0.966744
0	Knn	0.962104
1	Decision Tree	0.957463