Липецкий государственный технический университет

Факультет автоматизации и информатики

Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №2

по дисциплине «Прикладные интеллектуальные системы и экспертные системы»

Бинарная классификация

Студент Крутских А.Ю.

Группа М-ИАП-22

Руководитель Кургасов В.В.

Липецк 2022 г.

Цель работы

Получить практические навыки решения задачи бинарной классификации данных в среде Jupiter Notebook. Научиться загружать данные, обучать классификаторы и проводить классификацию. Научиться оценивать точность полученных моделей.

Задание кафедры

1) в среде Jupiter Notebook создать новый ноутбук (Notebook);

2) импортировать необходимые для работы библиотеки и модули;

3) загрузить данные в соответствие с вариантом;

4) вывести первые 15 элементов выборки (координаты точек и метки класса);

5) отобразить на графике сгенерированную выборку. Объекты разных классов должны иметь разные цвета;

6) разбить данные на обучающую (train) и тестовую (test) выборки в пропорции 75% - 25% соответственно;

7) отобразить на графике обучающую и тестовую выборки. Объекты разных классов должны иметь разные цвета;

8) реализовать модели классификаторов, обучить их на обучающем множестве. Применить модели на тестовой выборке, вывести результаты классификации:

• Истинные и предсказанные метки классов

• Матрицу ошибок (confusion matrix)

• Значения полноты, точности, f1-меры и аккуратности

• Значение площади под кривой ошибок (AUC ROC)

• Отобразить на графике область принятия решений по каждому классу

В качестве методов классификации использовать:

a) Метод к-ближайших соседей (n_neighbors = {1, 3, 5, 9})

b) Наивный байесовский метод

c) Случайный лес (n_estimators = {5, 10, 15, 20, 50})

9) по каждому пункту работы занести в отчет программный код и результат вывода;

10) по результатам п.8 занести в отчет таблицу с результатами классификации всеми методами и выводы о наиболее подходящем методе классификации ваших данных.

Ход работы

Вариант 8.

Вариант	**8**
Вид классов	moons
Random_state	15
cluster_std	-
noise	0.2
Centers	-

Для всех вариантов, использующих для генерации make_classification, дополнительные параметры: n_features=2, n_redundant=0, n_informative=1, n_clusters_per_class=1.

1) в среде Jupiter Notebook создать новый ноутбук (Notebook);

2) импортировать необходимые для работы библиотеки и модули;

3) загрузить данные в соответствие с вариантом;

4) вывести первые 15 элементов выборки (координаты точек и метки класса);

5) отобразить на графике сгенерированную выборку. Объекты разных классов должны иметь разные цвета;

6) разбить данные на обучающую (train) и тестовую (test) выборки в пропорции 75% - 25% соответственно;

7) отобразить на графике обучающую и тестовую выборки. Объекты разных классов должны иметь разные цвета;

8) реализовать модели классификаторов, обучить их на обучающем множестве. Применить модели на тестовой выборке, вывести результаты классификации:

• Истинные и предсказанные метки классов

• Матрицу ошибок (confusion matrix)

• Значения полноты, точности, f1-меры и аккуратности

• Значение площади под кривой ошибок (AUC ROC)

• Отобразить на графике область принятия решений по каждому классу

В качестве методов классификации использовать:

a) Метод к-ближайших соседей (n_neighbors = {1, 3, 5, 9})

b) Наивный байесовский метод

c) Случайный лес (n_estimators = {5, 10, 15, 20, 50})

9) по каждому пункту работы занести в отчет программный код и результат вывода;

10) по результатам п.8 занести в отчет таблицу с результатами классификации всеми методами и выводы о наиболее подходящем методе классификации ваших данных.

Код программы

Вывод

В ходе выполнения данной лабораторной работы мы получили базовые навыки работы с языком python и набором функций для анализа и обработки данных. Получили практические навыки решения задачи бинарной классификации данных в среде Jupiter Notebook. Научились загружать данные, обучать классификаторы и проводить классификацию. Научились оценивать точность полученных моделей.

Контрольные вопросы

1) Постановка задачи классификации данных. Что такое бинарная классификация?

Задача классифика́ции — 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.

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

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

Бинарная классификация – это один из типов задач классификации в машинном обучении, когда мы должны классифицировать два взаимоисключающих класса. Например, классифицировать сообщения как спам или не спам, классифицировать новости как фальшивые или настоящие.

2) Общий алгоритм решения задачи классификации данных.

1. Конструирование модели: описание множества предопределенных классов.

- Каждый пример набора данных относится к одному предопределенному классу.

- На этом этапе используется обучающее множество, на нем происходит конструирование модели.

- Полученная модель представлена классификационными правилами, деревом решений или математической формулой.

2. Использование модели: классификация новых или неизвестных значений.

- Оценка правильности (точности) модели.

1. Известные значения из тестового примера сравниваются с результатами использования полученной модели.

2. Уровень точности - процент правильно классифицированных примеров в тестовом множестве.

3. Тестовое множество, т.е. множество, на котором тестируется построенная модель, не должно зависеть от обучающего множества.

- Если точность модели допустима, возможно использование модели для классификации новых примеров, класс которых неизвестен

3) Чем отличаются обучающая и тестовая выборки? Какие существуют способы формирования обучающей и тестовой выборок?

Обучающая выборка - это набор, который подается на вход модели в процессе обучения вместе с ответами, с целью научить модель видеть связь между этими признаками и правильным ответом

Тестовая выборка используется для проверки модели. Модель не получает целевой признак на вход и, более того, должна предсказать его величину используя значения остальных признаков. Эти предсказания потом сравниваются с реальными ответами.

Способы формирования выборок:

- метод удерживания
- метод k-кратной перекрёстной проверки
- скользящий экзамен
- стратификация
- самонастройка

4) Как рассчитываются значения полноты и точности классификации?

Точность (precision) и полнота (recall) являются метриками которые используются при оценке большей части алгоритмов извлечения информации. Иногда они используются сами по себе, иногда в качестве базиса для производных метрик, таких как F-мера или R-Precision.

TP — истино-положительное решение;

TN — истино-отрицательное решение;

FP — ложно-положительное решение;

FN — ложно-отрицательное решение.

Тогда, точность и полнота определяются следующим образом:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

5) Как рассчитывается значение площади под кривой ошибок?

AUC-ROC (или ROC AUC) — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve). Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

6) Что показывает и как рассчитывается матрица ошибок?

На практике значения точности и полноты гораздо более удобней рассчитывать с использованием матрицы неточностей (confusion matrix). В случае если количество классов относительно невелико (не более 100-150 классов), этот подход позволяет довольно наглядно представить результаты работы классификатора.

Матрица неточностей – это матрица размера N на N, где N — это количество классов. Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора.

Матрица ошибок позволяет оценить эффективность прогноза не только в качественном, но и в количественном выражении

Это таблица с 4 различными комбинациями прогнозируемых и фактических значений.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

7) Алгоритм и особенности метода к-ближайших соседей.

Шаг 1 − Загружаем обучающий и тестовый dataset.

Шаг 2 − Выбираем значение K, то есть ближайшие точки данных. Оно может быть любым целым числом.

Шаг 3 - Вычисляем расстояние между тестовыми данными и каждой строкой обучающих данных с помощью любого из методов. Наиболее часто используемый метод вычисления расстояния - евклидов.

Шаг 4− Отсортировываем в порядке возрастания, основываясь на значении расстояния.

Шаг 5 − Алгоритм выбирает верхние K строк из отсортированного массива.

Шаг 6 − Назначаем класс контрольной точке на основе наиболее частого класса этих строк.

Особенности:

- Алгоритм прост и легко реализуем.

- Не чувствителен к выбросам.

- Нет необходимости строить модель, настраивать несколько параметров или делать дополнительные допущения.

- Алгоритм универсален. Его можно использовать для обоих типов задач: классификации и регрессии.

8) Алгоритм и особенности метода случайного леса.

Порядок действий в алгоритме

- Загрузите ваши данные.

- В заданном наборе данных определите случайную выборку.

- Далее алгоритм построит по выборке дерево решений.

- Дерево строится, пока в каждом листе не более n объектов, или пока не будет достигнута определенная высота.

- Затем будет получен результат прогнозирования из каждого дерева решений.

На этом этапе голосование будет проводиться для каждого прогнозируемого результата: мы выбираем лучший признак, делаем разбиение в дереве по нему и повторяем этот пункт до исчерпания выборки.

В конце выбирается результат прогноза с наибольшим количеством голосов. Это и есть окончательный результат прогнозирования.

Особенности:

— имеет высокую точность предсказания, на большинстве задач будет лучше линейных алгоритмов; точность сравнима с точностью бустинга

— практически не чувствителен к выбросам в данных из-за случайного сэмлирования

— не чувствителен к масштабированию значений признаков, связано с выбором случайных подпространств

— способен эффективно обрабатывать данные с большим числом признаков и классов

— одинаково хорошо обрабатывет как непрерывные, так и дискретные признаки

— редко переобучается, на практике добавление деревьев почти всегда только улучшает композицию, но на валидации, после достижения определенного количества деревьев, кривая обучения выходит на асимптоту

— для случайного леса существуют методы оценивания значимости отдельных признаков в модели

— хорошо работает с пропущенными данными; сохраняет хорошую точность, если большая часть данных пропущенна

— предполагает возможность сбалансировать вес каждого класса на всей выборке, либо на подвыборке каждого дерева