Проверочная работа 3

Факультатив «Введение в анализ данных и машинное обучение на Python»

15 февраля 2020

Время на выполнение: 30-40 минут

Решите как можно больше заданий. Их много, но все они не очень сложные. **При выполнении можно пользоваться любыми материалами** (и Интернетом, и Google, и записями с прошлых занятий) – за исключением помощи соседей, социальных сетей и мессенджеров. Шпаргалка в начале содержит достаточно информации для выполнения заданий на деревья и kNN, даже если Вы пропустили эти темы (а Интернет – ещё больше). Задания выполняйте на отдельном листе.

Шпаргалка

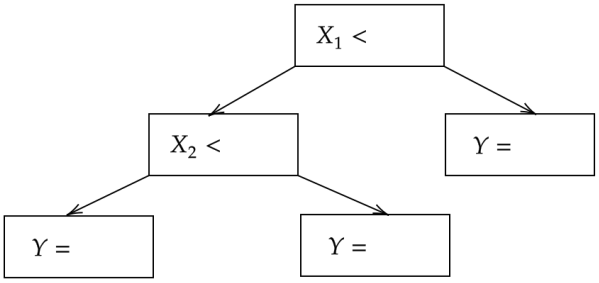
- Решающее дерево строится так, чтобы при каждом разбиении уменьшить неопределённость среди классов в каждом листе. Например, разбиение [1, 0, 0] более предпочтительно, чем разбиение [1, 2, 3], потому что в первом случае в лист попали объекты двух классов (0 и 1), а во второй – трёх (1, 2, 3), и если мы случайным образом выберем объект из листа, то в первом случае мы более уверены, какой это будет класс, чем во втором.
- Разбиение проходит по признакам. В наших задачах порог можно выбирать любой разумный: если мы хотим отделить значения признака 1 и 2, то и $X_j < 3$, и $X_j < 3.1$, и $X_j < 2.5$ подойдут.
- Случайный лес – композиция (особым образом построенное объединение) решающих деревьев.
- Принцип работы *kNN*: отнести новый объект к тому классу, который наиболее распространён среди k ближайших соседей. Ближайших – в смысле обычного расстояния.

Задание 1

Решающее дерево для задачи бинарной классификации строится по следующей выборке:

X_1	X_2	Y
1	0	0
1	2	1
3	3	1
4	3	1

Заполните схематичное представление обученного дерева. Задача имеет бесконечное множество вариантов верных решений, в качестве правильного принимается любой из них.



Задание 2

a) Опишите одним словом, к какой проблеме склонны решающие деревья.

b) Назовите не менее одного способа решения данной проблемы.

c) Склонен ли случайный лес к этой проблеме?

Задание 3

Пусть решается задача бинарной классификации, в которой Y может принимать значения «круг» или «звезда». Обучающая выборка состоит из 5 объектов, 2 из которых относятся к классу «звезда», а 3 – к классу «круг». Рисунок ниже представляет собой выборку, изображённую в пространстве признаков.

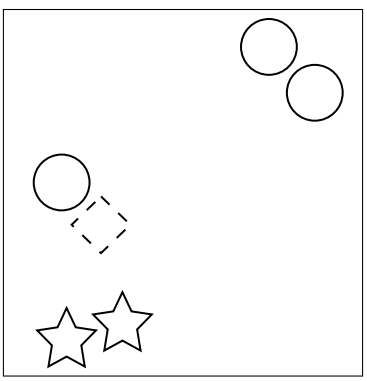
В выборку поступает новое наблюдение, обозначенное на рисунке пунктирным ромбом. Используя метод kNN, определите, к какому классу будет отнесено это новое наблюдение для:

a) $k = 1$

b) $k = 3$

c) $k = 5$

Расстояние задаётся обычным образом.



Задание 4

Выберите **один** наиболее подходящий ответ.

Модель:

$$\hat{Height}_i = 30 + 120 \times Weight_i,$$

где $Height_i$ – рост индивида в сантиметрах, $Weight_i$ – вес индивида в килограммах, скорее:

a) Не является линейной.

b) Корректно обучена.

c) Является моделью множественной регрессии.

d) Не имеет свободного члена.

e) Некорректно обучена.

f) Покажет высокое качество на тестовой выборке.

Задание 5

Одним предложением опишите, что делают следующие фрагменты кода:

1. `lr_model = LinearRegression()`
2. `tree.fit(X_train, Y_train)`

Задание 6

Как исследователь может определить, какое число ближайших соседей в kNN или решающих деревьев в случайном лесе следует выбрать?

Подсказка: вспомните или посмотрите, как мы решали эту проблему в коде к соответствующим занятиям.

Задание 7 (необязательное)

Как Вы считаете, какая из следующих моделей лучше подходит для описания зависимости уровня дохода индивида от его возраста:

1. $income_i = \beta_0 + \beta_1 \times age_i + u_i$
2. $income_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times age_i^2 + u_i$?

Почему?

Бонусное задание

Так как ввиду не очень высокой посещаемости велика вероятность, что наш факультатив будет досрочно закрыт и эта проверочная работа станет последней, напишите, пожалуйста, фидбэк по тому, что Вам понравилось, а что нет, а также предложения по тому, что следует изменить в следующем году. Также можете нарисовать картинку.