1

Какая стратегия поведения в листьях решающего дерева приводит к меньшей вероятности ошибки: выдавать в ответе тот класс, который преобладает в листе, или отвечать случайно с тем же распределением классов, что и в листе? Считайте, что рассматривается задача многоклассовой классификации.

2

При обучении SVM с линейным ядром на наборе данных с очень большим числом разреженных признаков точность многоклассовой классификации на тестовой выборке получилась равной 95%, а при обучении SVM с радиальным ядром — 34%. Почему SVM с более сложным ядром показал менее высокое качество? Когда можно ожидать от радиального ядра улучшения качества по сравнению с линейным?

3

В одном проекте заказчик очень хотел, чтобы исследователь решал не поставленную задачу классификации на классы 0 и 1, а задачу регрессии на тех же метках с модулем отклонения в качестве функции потерь. Замысел заказчика был в том, что оцененные числа получатся дробными, и это будет приближением для вероятности класса 1. Считая, что алгоритм старается минимизировать математическое ожидание потерь при условии известного объекта x, выясните, какими будут получаться прогнозы при такой функции потерь и насколько они будут соответствовать замыслу заказчика.

4

Какой должна быть функция потерь в предыдущей задаче, чтобы действительно оценивались вероятности? Покажите, что это так.

5

Когда в задаче бинарной классификации предпочтительней использовать ROC-AUC, а не accuracy? Как изменится ROC-AUC решения задачи бинарной классификации, если умножить все прогнозы на 2?

6

Имеется файл log.txt размером 1Tb, содержащий в себе лог в следующем формате:

номер записи, тип запроса, время отклика

Пример начала лога:

1,/index,0.06

2,/test,0.03

3,/home,0.561

4,/home,0.87

5,/index,1.02

Напишите программу на Python 2, которая для каждого типа запроса подсчитывает среднее время отклика и 95% доверительный интервал для этой величины. Реализуйте также проверку гипотезы о равенстве средних времён отклика для типов запроса /index и /test на уровне значимости 5%.

(оставьте пожалуйста в виде ссылки на код на любой из удобных для вас ресурсов)