Школа анализа данных
Машинное обучение, часть 1
Теоретическое домашнее задание №1

Решите предложенные задачи. Решения необходимо оформить в виде PDF документа. Каждая задача должна быть подробно обоснована, задачи без обоснования не засчитываются. Решения пишутся в свободной форме, однако так, чтобы проверяющие смогли разобраться. Если проверяющие не смогут разобраться в решении какой-нибудь задачи, то она автоматически не засчитывается.

Задача 1 (0.5 балла). Объясните, стоит ли использовать оценку leave-one-out-CV или k-fold-CV с небольшим k в случае, когда:

- обучающая выборка содержит очень малое количество объектов;
- обучающая выборка содержит очень большое количество объектов.

Задача 2 (0.5 балла). Метрика Минковского определяется как

$$\rho_p(x, y) = \Big(\sum_{i=1}^{n} |x_i - y_i|^p\Big)^{\frac{1}{p}}$$

для $p \geq 1$. Частными случаями данной метрики являются:

- Евклидова метрика ($p = 2$)
- Манхэттенское расстояние ($p = 1$)
- Метрика Чебышева ($p = \infty$)

Изобразите линии уровня функции $f(x) = \rho_p(x, 0)$ для трех приведенных случаев в двумерном пространстве ($n = 2$).

Задача 3 (2 балла). Рассмотрим l точек, равномерно распределенных в n-мерном единичном шаре с центром в начале координат. Предположим, что мы хотим применить метод ближайшего соседа для точки начала координат. Зададимся вопросом, на каком расстоянии будет расположен ближайший объект. Выведите выражение для медианы расстояния от начала координат до ближайшего объекта. Чтобы проинтерпретировать полученный результат, подставьте в формулу конкретные значения, например, $l = 500$, $n = 10$. Что будет происходить при дальнейшем увеличении размерности пространства?

Задача 4 (1 балл). Известно, что метод ближайших соседей неустойчив к шуму. Рассмотрим модельную задачу бинарной классификации с одним признаком и двумя объектами обучающей выборки: $x_1 = 0.1$, $x_2 = 0.5$. Первый объект относится к первому классу, второй — ко второму. Добавим к объектам новый шумовой признак, распределенный равномерно на отрезке $[0, 1]$. Теперь каждый объект описывается уже двумя признаками. Пусть требуется классифицировать новый

объект $u = (0, 0)$ в этом пространстве методом одного ближайшего соседа (метрика евклидова). Какова вероятность того, что после добавления шума второй объект окажется к нему ближе, чем первый?

Задача 5 (1 балл) Чем больше метод машинного обучения склонен к переобучению, тем больше настраиваемых параметров у него должно быть. Действительно, склонность к переобучению свидетельствует о гибкости модели, а гибкость говорит о большом количестве «степеней свободы» модели или, другими словами, параметров.

Рассмотрим несколько моделей. Линейные алгоритмы классификации имеют порядка n настраиваемых параметров (вектор весов), где n — размерность пространства. На рис. 1 показан результат работы линейного алгоритма в случае бинарной классификации. Метод K ближайших соседей (K-NN) имеет один настраиваемый параметр K. На рис. 2 показан результат работы обученного K-NN на тех же данных.

На этих и других примерах можно легко видеть, что K-NN куда более гибок, нежели линейная модель. Однако K-NN обладает всего одним параметром, а линейная модель целым набором из n параметров. Почему так происходит, ведь это противоречит рассуждениям выше?

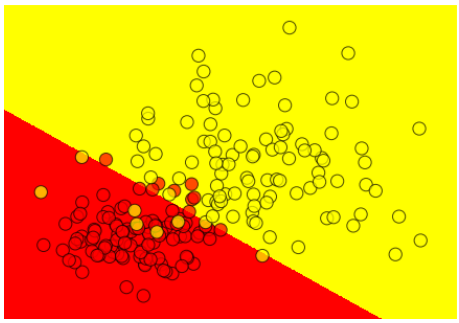

Рис. 1: Линейная модель

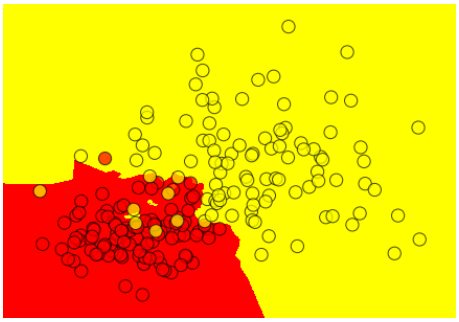

Рис. 2: Метод ближайших соседей

Задача 6 (2 балла). Покажите, что асимптотическая сложность алгоритма построения бинарного решающего дерева на l объектах в n-мерном пространстве равна $O(n\,l \log l)$. Для простоты можно считать, что дерево сбалансированное, в каждом листе остается по одному объекту и в качестве функционала качества используется Mean Squared Error (MSE):

$$Q = \frac{1}{l}\sum_{i=1}^{l}(y_i - \tilde{y}_i)^2.$$

Задача 7 (3 балла). Пусть имеется построенное решающее дерево для задачи многоклассовой классификации. Рассмотрим лист дерева с номером m и объекты R_m, попавшие в него. Обозначим за p_{mk} долю объектов k-го класса в листе m. *Индексом Джини* этого листа называется величина

$$\sum_{k=1}^{K} p_{mk}(1 - p_{mk}),$$

где K — общее количество классов. Индекс Джини обычно служит мерой того, насколько хорошо в данном листе выделен какой-то один класс. Решите следующие задачи:

1. Сопоставим в соответствие листу m алгоритм классификации $a(x)$, который предсказывает класс случайно, причем класс k выбирается с вероятностью p_{mk}. Покажите, что матожидание частоты ошибок этого алгоритма на объектах из R_m равно индексу Джини.

2. Какая стратегия поведения в листьях решающего дерева приводит к меньшей вероятности ошибки: рассмотренный алгоритм $a(x)$ или предсказание преобладающего в листе класса?

3. *Дисперсией класса k* назовем дисперсию выборки $\{[y(x_i) = k] : x_i \in R_m\}$, где $y(x_i)$ — класс объекта x_i, а R_m — множество объектов в листе. Покажите, что сумма дисперсий всех классов в заданном листе равна его индексу Джини.