Машинное обучение
Теоретическое домашнее задание №5

Задача 1 (3 балла). Известно, что решающие деревья плохо подходят для онлайн-обучения — если приходят новые размеченные объекты, то нельзя быстро дообучить на них дерево, поскольку это может потребовать полного изменения структуры. Можно пересчитать прогнозы в листьях, но это вряд ли можно назвать полноценным дообучением.

Как будет выглядеть дообучение модели k ближайших соседей, если поступили новые размеченные объекты? Сколько и каких операций для этого потребуется? Быстрее ли это, чем обучение модели с нуля на расширенной выборке? Для определённости считайте, что речь идёт о задаче многоклассовой классификации.

Задача 2 (7 баллов). Обучение метода одного ближайшего соседа заключается в запоминании обучающей выборки. На этапе применения модели к новому объекту x мы ищем ближайший к нему объект из обучения и возвращаем класс этого объекта. Воспользуемся несколько иным взглядом на применение такой модели.

Допустим, мы решаем задачу бинарной классификации, а в качестве функции расстояния используем евклидову метрику.

Диаграммой Вороного, соответствующей выборке X, назовём такое разбиение пространства на области, что каждая область состоит из точек, для которых одна и та же точка из выборки X является ближайшей. Более формально, диаграмма Вороного для выборки X состоит из ℓ областей $R_1, \ldots, R_\ell$, определяемых как

$$R_i = \{x \in \mathbb{R}^d \,|\, \rho(x, x_i) < \rho(x, x_j), j \neq i\}.$$

Покажем, что при использовании классификатора одного ближайшего соседа граница между классами является подмножеством границ между такими областями.

1. Покажите, что для указанной постановки разделяющую поверхность между классами можно описать через n гиперплоскостей для некоторого n:

$$a(x) = \prod_{j=1}^{n} [\langle w_j, x\rangle < t_j], \tag{1.1}$$

   причём каждая из этих гиперплоскостей образует одну из сторон одной из ячеек диаграммы Вороного.

2. Допустим, мы решили вместо обучающей выборки хранить n гиперплоскостей, образующих разделяющую поверхность, и на этапе применения использовать формулу (1.1). Даст ли такой подход гарантированный выигрыш по памяти или по времени применения для одного объекта в худшем случае? Ответ обоснуйте.

Задача 3 (6 балла). На лекциях говорилось, что критерий информативности для набора объектов R вычисляется на основе того, насколько хорошо их целевые переменные предсказываются константой (при оптимальном выборе этой константы):

$$H(R) = \min_{c \in \mathbb{Y}} \frac{1}{|R|} \sum_{(x_i, y_i) \in R} L(y_i, c),$$

где $L(y, c)$ — некоторая функция потерь. Соответственно, чтобы получить вид критерия при конкретной функции потерь, необходимо аналитически найти оптимальное значение константы и подставить его в формулу для $H(R)$.

Выведите критерии информативности для следующих функций потерь:

1. $L(y, c) = (y - c)^2$;
2. $L(y, c) = \sum_{k=1}^{K} (c_k - [y = k])^2$;
3. $L(y, c) = -\sum_{k=1}^{K} [y = k] \log c_k$.

У вас должны получиться дисперсия, критерий Джини и энтропийный критерий соответственно.

Задача 4 (4 балла). Запишите оценку сложности построения одного решающего дерева в зависимости от размера обучающей выборки ℓ, числа признаков d, максимальной глубины дерева D. В качестве предикатов используются пороговые функции $[x_j > t]$. При выборе предиката в каждой вершине перебираются все признаки, а в качестве порогов рассматриваются величины t, равные значениям данного признака на объектах, попавших в текущую вершину.

Задача 5 (задача не оценивается). Ответьте на вопросы:

1. Что такое решающее дерево? Как по построенному дереву найти прогноз для объекта?
2. Зачем в вершинах нужны предикаты? Какие типы предикатов вы знаете? Приведите примеры.
3. Почему для любой выборки можно построить решающее дерево, имеющее нулевую ошибку на ней?
4. Почему не рекомендуется строить небинарные деревья (т.е. имеющие больше двух потомков у каждой вершины)?
5. Как устроен жадный алгоритм построения дерева? Какие у него параметры?
6. Зачем нужны критерии информативности?
7. Как задается критерий ошибки классификации? Критерий Джини? Энтропийный критерий? Какой у них смысл?
8. Как задается критерий информативности, основанный на среднеквадратичной ошибке, в задачах регрессии?

9. Какие критерии останова вы знаете?
10. Что такое стрижка дерева?
11. Какие методы обработки пропущенных значений вы знаете?
12. Как можно учитывать категориальные признаки в решающем дереве?
13. Как можно свести задачу перебора всех разбиений категориального признака к задаче поиска оптимального разбиения для вещественного признака?