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) Кроссвалидация, LOO, k-fold.

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) Метрики, визуализация.

Метрика Минковского с параметром p определяется как

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

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 (1.5 балла) Метрические методы, kNN, проклятие размерности.

Рассмотрим l точек, распределенных равномерно по объему n-мерного единичного шара с центром в нуле. Предположим, что мы хотим применить метод ближайшего соседа для точки начала координат. Зададимся вопросом, на каком расстоянии будет расположен ближайший объект. Для ответа на этот вопрос выведите выражение для **медианы** расстояния от начала координат до ближайшего объекта. Чтобы проинтерпретировать полученный результат, подставьте в формулу конкретные значения: $l = 500$ и $n = 10$. Покажите, как будет меняться значение медианы при дальнейшем увеличении размерности пространства при фиксированном количестве точек и постройте график этой зависимости. Поясните, почему полученная для медианы формула наглядно демонстрирует проклятие размерности. Для размерности n посчитайте, сколько точек $l = f(n)$ необходимо взять, чтобы побороть проклятье размерности.

Задача 4 (1 балл) Метрические методы, kNN, устойчивость к шуму.

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 с евклидовой метрикой. Какова вероятность того, что после добавления шума второй объект окажется ближе к объекту u, чем первый?

Задача 5 (1 балл) Число степеней свободы алгоритма обучения.

Известно, что чем больше у метода машинного обучения настраиваемых параметров, тем больше он склонен к переобучению. Действительно, склонность к переобучению свидетельствует о «гибкости» модели, а «гибкость» говорит о большом количестве «степеней свободы» модели или, другими словами, параметров. Здесь под «гибкостью» будем неформально подразумевать способность алгоритма подстроиться под любые данные.

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

На этих примерах можно легко видеть, что K-NN куда более «гибок», чем линейная модель. Однако K-NN обладает всего одним параметром (число соседей), а линейная модель целым набором из n параметров. Почему так происходит? Что не так с приведенными выше рассуждениями?

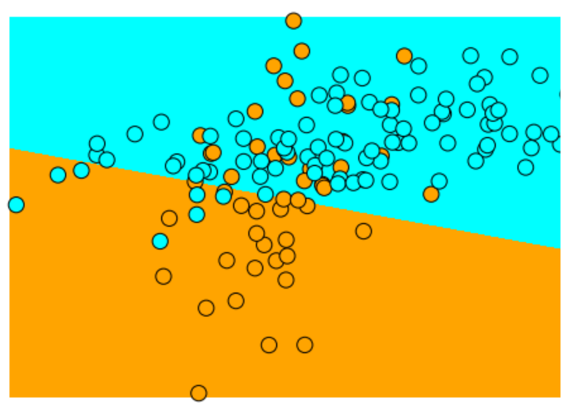

Рис. 1: Линейная модель

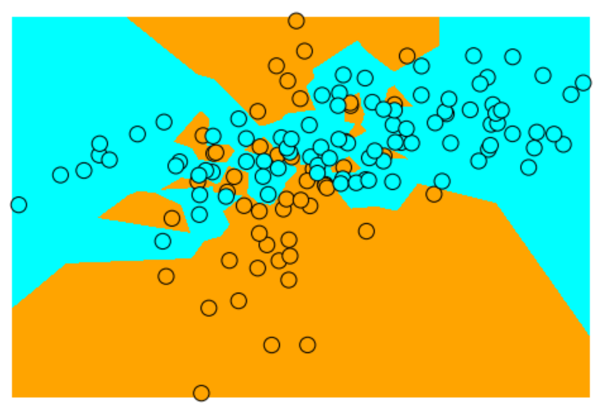

Рис. 2: Метод ближайших соседей

Задача 6 (0.5 балла) Функции потерь.

Рассмотрим выборку из объектов $x_1, \dots, x_l$ и ответов на них $y_1, \dots, y_l$, где $y_i \in \{0, 1\}$. Пусть решается задача бинарной классификации, и некоторый классификатор выдает $p(x_i)$ — вероятность принадлежности объекта x_i классу 1. Будем считать, что объекты получаются из распределения $p(x)$ независимо. Правдоподобие выборки, описываемое распределением $p(x)$, показывает, насколько вероятно пронаблюдать данные $x_1, y_1, \ldots, x_l, y_l$. Чем точнее классификатор предсказывает вероятность принадлежности классу 1, тем выше правдоподобие выборки. Часто на практике оказывается вычислительно удобнее не максимизировать правдоподобие выборки, а минизировать отрицательный логарифм правдоподобия выборки. Покажите, что в случае бинарной классификации отрицательный логарифм правдоподобия соответствует следующему выражению

$$LogLoss = -LogLikelihood = -\sum_{i=1}^{l}(y_i \log p(x_i) + (1-y_i)\log(1-p(x_i)))$$

Задача 7 (1.5 балла) Функции потерь, константное предсказание, решающие деревья.

Допустим, при построении решающего дерева в некоторый лист попало N объектов $x_1, ..., x_N$ с метками $y_1, ..., y_N$. Предсказание в каждом листе дерева — константа. Найдите, какое значение $\tilde{y}$ должен предсказывать этот лист для минимизации следующих функций потерь:

1. Mean Squared Error (средний квадрат ошибки) для задачи регрессии:

$$Q = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{y})^2;$$

2. Mean Absolute Error (средний модуль отклонения) для задачи регрессии:

$$Q = \frac{1}{N}\sum_{i=1}^{N}|y_i - \tilde{y}|.$$

3. LogLoss (логарифмические потери) для задачи классификации:

$$Q = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log \tilde{y} + (1-y_i)\log(1-\tilde{y})), \quad \tilde{y} \in [0,1], \quad y_i \in \{0,1\}.$$

Задача 8 (1 балл) Решающие деревья, функции потерь, impurity functions.

$$\Phi(U) - \frac{|U_1|}{|U|}\Phi(U_1) - \frac{|U_2|}{|U|}\Phi(U_2) \to \max$$

таким выражением в лекции задается критерий, по которому происходит ветвление вершины решающего дерева. Давайте разберемся подробнее.

Impurity function $\Phi(U)$ («функция нечистоты» или «функция неопределенности») используется для того, чтобы измерить степень неоднородности целевых меток $y_1, \ldots, y_l$ для множества объектов U размера l. Например, при обучении решающего дерева в текущем листе выбирается такое разбиение множества объектов U на два непересекающихся множества U_1 и U_2, чтобы impurity function $\Phi(U)$ исходного множества U как можно сильнее превосходила нормированную impurity function в новых листьях $\frac{|U_1|}{|U|}\Phi(U_1) + \frac{|U_2|}{|U|}\Phi(U_2)$. Отсюда и получается, что нужно выбрать разбиение, решающее задачу

$$\Phi(U) - \frac{|U_1|}{|U|}\Phi(U_1) - \frac{|U_2|}{|U|}\Phi(U_2) \to \max.$$

Полученную разность называют Gain (выигрыш), и она показывает, на сколько удалось уменьшить «неопределенность» от разбиения листа два новых.

В соответствии с одним из возможных определений, impurity function — это значение функционала ошибки $Q = \frac{1}{l} \sum\limits_{i=1}^{l} \mathcal{L}(y_i, \tilde{y})$ в листе с множеством объектов U при константном предсказании $\tilde{y}$, оптимальном для Q (см. задачу 7):

$$\Phi(U) = \frac{1}{l} \sum_{i=1}^{l} \mathcal{L}(y_i, \tilde{y}).$$

Понятно, что каждому критерию разбиения соответствует своя impurity function $\Phi(U)$, а в основе каждой $\Phi(U)$ лежит некоторая функция потерь. Давайте разберемся, откуда берутся различные критерии разбиения.

1. Покажите, что для квадратичных потерь $\mathcal{L}(y_i, \tilde{y}) = (y_i - \tilde{y})^2$ в задаче регрессии $y_i \in \mathbb{R}$ impurity function $\Phi(U)$ равна выборочной дисперсии целевых меток объектов, попавших в лист дерева.

2. Покажите, что для функции потерь Logloss $\mathcal{L}(y_i, \tilde{y}) = -y_i \log(\tilde{y}) - (1 - y_i) \log(1 - \tilde{y})$ в задаче классификации $y_i \in \{0, 1\}$ impurity function $\Phi(U)$ соответствует энтропийному критерию разбиения.

Задача 9 (1 балл)Решающие деревья, индекс Джини.

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 (см. impurity function в предыдущей задаче).

1. Поставим 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. *Дисперсией класса* k назовем дисперсию выборки $\{[y_i = k] : \ x_i \in R_m\}$, где y_i — класс объекта x_i, $[f]$ — индикатор истинности выражения f, равный 1 если f верно, и нулю в противном случае, а R_m — множество объектов в листе. Покажите, что сумма дисперсий всех классов в заданном листе равна его индексу Джини.

Задача 10 (1.5 балла) Бинарные решающие деревья, MSE.

Предложите алгоритм построения **оптимального** бинарного решающего дерева для задачи регрессии на l объектах в n-мерном пространстве с асимптотической сложностью $O(nl \log l)$. В качестве предикатов нужно рассматривать пороговые правила (наиболее распространенный случай на практике). Для простоты можно считать, что получающееся дерево близко к сбалансированному (т.е. его глубина имеет порядок $O(\log l)$) и в качестве функции ошибки используется Mean Squared Error (MSE):

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

Под оптимальностью в данной задаче подразумевается, что в каждом узле дерева делается оптимальное с точки зрения MSE разбиение на два поддерева.