Основы машинного обучения
Контрольная работа
Вариант 1

Задача 1 (3 балла). В этой задаче мы будем обсуждать модель линейной регрессии $a(x) = \langle w, x \rangle$. Ответьте на вопросы:

1. Часто линейные модели обучают с помощью градиентного спуска. Формула шага в нём выглядит так:

   $$w^{(t)} = w^{(t-1)} - \eta_t \nabla_w Q(w^{(t-1)}).$$

   Ниже предложено три варианта формул для длины шага. Объясните для каждого из них, какие проблемы могут возникнуть в градиентном спуске при его использовании.

   - $\eta_t = t$;
   - $\eta_t = \sin(t)$;
   - $\eta_t = \frac{1}{100^t}$.

2. Мы обучаем линейную регрессию на среднеквадратичную ошибку:

   $$\frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - y_i\right)^2 \to \min_w$$

   Такой подход имеет какие-то смещение и разброс. Напишите, как по сравнению с этим подходом изменятся смещение и разброс у следующих модификаций, и поясните, почему вы так считаете:

   - $\frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - y_i\right)^2 + \|w\|_2^2 \to \min_w$;
   - $\frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - 1\right)^2 \to \min_w$;

3. Ниже записан функционал для некоторой модели:

   $$\frac{1}{\ell} \sum_{i=1}^{\ell} \left(w_0 + w_1 x_{i1}^{\alpha} + w_2 x_{i2}^{\beta} - y_i^{\gamma}\right)^2$$

   Здесь требуется подобрать значения для $w_0, w_1, w_2, \alpha, \beta, \gamma$. Что из этого будет параметром, а что гиперпараметром? Иными словами, что из этого можно подбирать по обучающей выборке, а что нельзя? Поясните свои ответы.

Задача 2 (3 балла). В этой задаче мы будем обсуждать решающие деревья и их композиции.

1. Как известно, решающие деревья обучаются жадным алгоритмом. При построении очередной вершины ключевым шагом является подбор предиката:

   $$Q(R_m, j, t) \to \min_{j,t}$$

   Здесь мы выбираем такой предикат $[x_j > t]$, при котором, скажем, энтропия в дочерних вершинах будет минимальна. А как именно решается записанная выше задача выбора оптимальных j и t? Например, задача выбора оптимальных весов для линейной регрессии решается градиентным спуском — а как здесь?

2. Говорят, что решающее дерево может идеально подогнаться под любую обучающую выборку (при условии, что там все объекты различны). Почему это так? Поясните на каком-нибудь примере выборки.

3. Ниже записаны формулы для построения N-й базовой модели $b_N(x)$ в градиентном бустинге:

   $$s_i = -\frac{\partial}{\partial z} L(a_{N-1}(x_i), z)\Big|_{a_{N-1}(x_i)}$$

   $$\frac{1}{\ell}\sum_{i=1}^{\ell}(b_N(x_i) - y_i)^2 \to \min_{b_N}$$

   Найдите все ошибки. Объясните их. Укажите, как их исправить.

4. Объясните, что из перечисленного ниже является параметрами, а что гиперпараметрами при обучении решающего дерева:

   - Максимальная глубина дерева;
   - Прогнозы в листьях дерева;
   - Размер случайной подвыборки, из которой выбирается оптимальный признак для предиката.

5. Что из перечисленного ниже приведёт к понижению смещения решающего дерева, а что к повышению? Поясните ваши ответы.

   - уменьшение максимально допустимой глубины дерева;
   - обучение дерева на случайном подмножестве признаков (вместо обучения на всех признаках);
   - замена поиска оптимального предиката в каждой вершине на выбор случайного предиката.

Задача 3 (2 балла). Ответьте на вопросы по функциям потерь:

1. Ниже записаны некоторые функции потерь для задачи регрессии. В формулах через y обозначается правильный ответ, через z — прогноз модели. Объясните, в чём особенности этих функций потерь и где они могут оказаться полезны.

- $\max\left(1, (y-z)^2\right)$;
- $\min\left(100, (y-z)^2\right)$;
- $\max(0, y-z)$.

2. Рассмотрим задачу бинарной классификации. Будем обучать линейный классификатор на следующий функционал:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(y_i\langle w, x_i\rangle - 1\right)^2.$$

Почему это будет плохой идеей?

Задача 4 (2 балла). Вам выдали классификатор $b(x)$, и вам предстоит разобраться, насколько она хороша. Для этого у вас есть тестовая выборка из 8 объектов. Ниже указаны правильные ответы и уверенности модели в положительном классе:

$b(x)$	3	1	7	0	2	9	4	-2
y	-1	-1	+1	+1	+1	-1	+1	-1

Выполните следующие шаги:

1. Нарисуйте ROC-кривую и посчитайте AUC-ROC.
2. Посчитайте точность и полноту этой модели при пороге $t = 5$ (считайте, что мы относим к положительному классу объекты, на которых $b(x) > t$).
3. Можно ли достичь полноты в хотя бы 25% при точности в 100%? Если да, укажите, при каком пороге. Если нет, поясните, почему.
4. Можно ли достичь полноты в 50% при точности в хотя бы 60%? Если да, укажите, при каком пороге. Если нет, поясните, почему.