Основы машинного обучения
Контрольная работа
Вариант 1

Задача 1 (1 балл). Рассмотрим метод k ближайших соседей с парзеновским окном:

$$a(x) = \underset{y \in \mathbb{Y}}{\arg\max} \sum_{i=1}^{k} w_i [y_{(i)} = y]; \quad w_i = [i \leqslant 10] \frac{1}{1 + \rho(x, x_{(i)})}.$$

(здесь мы взяли конкретное ядро и специально убрали ширину окна h для простоты)

Пусть у нас задача классификации на 5 классов. Считайте, что выборка достаточно большая (скажем, больше 100.000 объектов). Ниже приведены утверждения, выполненные для обычного метода kNN. Какие из них останутся верными для нашего метода, а какие нет? Ответы обоснуйте.

- При $k = 1$ доля верных ответов на обучающей выборке будет равна единице.
- При $k = \ell$ мы получим константную модель, которая выдаёт один и тот же прогноз для любого объекта (здесь ℓ — число объектов в обучающей выборке).

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

1. Иногда в задачах регрессии применяют функционал MAPE, измеряющий относительную ошибку:

   $$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

   Говорят, что этот функционал связан с MAE, средней абсолютной ошибкой. А как именно? И на каких данных MAPE в точности совпадёт с MAE?

2. Мы знаем, что в нашей задаче ожидаются выбросы, и очень этого боимся. Чтобы выбросы точно не повлияли на модель, мы разработали специальную функцию потерь:

   $$L(y, z) = [|y - z| < 100](y - z)^2.$$

   Что не так с этой функцией потерь и почему обучение на неё, скорее всего, приведёт к абсолютно бесполезной модели? А как её исправить, чтобы правда избежать проблемы с выбросами?

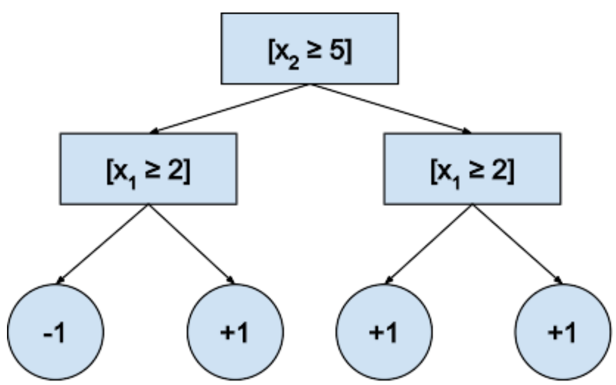
3. Известно, что линейные модели часто обучаются градиентным спуском. В нём важно определить, при каких условиях нужно остановить обучение — это условие называется критерием останова. Ниже перечислено несколько критериев останова. Для каждого из них прокомментируйте, является ли он осмысленным или же приведёт к некорректной работе метода оптимизации. Через Q_t обозначается величина функционала ошибки после итерации t.

   - $\|w^t - w^{t-1}\| < \varepsilon$;
   - $\|w^t\| < \varepsilon$;
   - $\|w^t + w^{t-1}\| < \varepsilon$;
   - $|Q_t| < \varepsilon$;
   - $|Q_t - Q_{t-1}| < \varepsilon$;
   - $Q_t - Q_{t-1} > \varepsilon$.

Задача 3 (2 балла). Рассмотрим следующую выборку с двумя признаками для задачи бинарной классификации:

x_1	3	4	1	0	0	1	4	6
x_2	6	5	7	6	2	1	3	1
y	+1	+1	-1	+1	-1	-1	-1	+1

Для этой выборки мы построили решающее дерево глубины два. Считайте, что в случае выполнения условия мы идём вправо.



Ответьте на вопросы:

1. Чему равна энтропия той части обучающей выборки, которая попала в самую левую листовую вершину? Используйте логарифм по основанию 2.

2. Чему равен прогноз дерева для объекта с признаками $(3, 2)$?

3. Как нужно изменить предикат в корне, чтобы для объекта из предыдущего пункта прогноз изменился на противоположный?

Задача 4 (2 балла). Ответьте на вопросы по решающим деревьям и композициям:

1. Построим бэггинг над случайными деревьями — то есть деревьями, у которых каждый предикат выбирается случайно (и признак, и порог берутся случайным образом независимо от данных). Деревья строятся, пока в каждом листе не окажется меньше 100 объектов. Что вы можете сказать про такие композиции? Они будут лучше, хуже или такими же по качеству, как и в классическом бэггинге над деревьями? Попробуйте провести рассуждения с использованием понятий смещения и разброса.

2. Перейдём к градиентному бустингу. Решаем задачу бинарной классификации. Пусть мы уже обучили $N-1$ дерево: $a_{N-1}(x) = \sum_{n=1}^{N-1} b_n(x)$. Следующее дерево мы собираемся обучать так:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - \left.\frac{\partial L(b_{N-1}(x_i), z)}{\partial z}\right|_{z=y_i}\right) \to \min_{b_N(x)}$$

Найдите и исправьте все ошибки в этой формуле. Кратко поясните для каждой ошибки, почему она является таковой.

Задача 5 (2 балла). Вам выдали классификатор $b(x)$, и вам предстоит разобраться, насколько он хорош. Для этого у вас есть тестовая выборка из 8 объектов. Ниже указаны правильные ответы и уверенности модели в положительном классе:

$b(x)$	-0.1	0.3	1	3	3.5	4	5	7
y	+1	-1	-1	-1	+1	+1	-1	+1

Ответьте на вопросы:

1. Нарисуйте ROC-кривую и посчитайте AUC-ROC.

2. Укажите самый близкий к нулю порог t, при котором модель $a(x) = [b(x) \geqslant t]$ будет иметь максимальную полноту.

3. Чему будет равен AUC-ROC на этой же выборке для модели $b_1(x) = (b(x))^2 - 0.3$?

Задача 1

Задача 1 (1 балл). Рассмотрим метод k ближайших соседей с парзеновским окном:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i [y_{(i)} = y]; \quad w_i = [i \leqslant 10] \frac{1}{1 + \rho(x, x_{(i)})}.$$

(здесь мы взяли конкретное ядро и специально убрали ширину окна h для простоты)

Пусть у нас задача классификации на 5 классов. Считайте, что выборка достаточно большая (скажем, больше 100.000 объектов). Ниже приведены утверждения, выполненные для обычного метода kNN. Какие из них останутся верными для нашего метода, а какие нет? Ответы обоснуйте.

1) • При $k = 1$ доля верных ответов на обучающей выборке будет равна единице.

2) • При $k = \ell$ мы получим константную модель, которая выдаёт один и тот же прогноз для любого объекта (здесь ℓ — число объектов в обучающей выборке).

вес w_i обратно пропорционален расстоянию $\rho(x, x_i)$

1) Да

2) Нет, т.к. $w_i \neq 0 \Leftarrow$ $i > 10$, т.е. вес далёких (>10 по близости) вершин будет занулён

Задача 2

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

1. Иногда в задачах регрессии применяют функционал MAPE, измеряющий относительную ошибку:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left| \frac{a(x_i) - y_i}{y_i} \right|$$

Говорят, что этот функционал связан с MAE, средней абсолютной ошибкой. А как именно? И на каких данных MAPE в точности совпадёт с MAE?

2. Мы знаем, что в нашей задаче ожидаются выбросы, и очень этого боимся. Чтобы выбросы точно не повлияли на модель, мы разработали специальную функцию потерь:

$$L(y, z) = [|y - z| < 100](y - z)^2.$$

Что не так с этой функцией потерь и почему обучение на неё, скорее всего, приведёт к абсолютно бесполезной модели? А как её исправить, чтобы правда избежать проблемы с выбросами?

1) MAPE по сути вычисляет среднюю относительную неточность, а MAE : $Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$ вычисляет среднюю абсолютную неточность

Они в точности совпадут на данных, где $y_i \in \{-1, 1\}$

2) $L(y, z) = [|y - z| < 100](y - z)^2$

Тут при $|y - z| \geqslant 100$ будет 0 ошибка

Поэтому модели будет очень легко минимизировать ошибку : просто взять z так, чтобы $|y - z| \geqslant 100$

$L(y, z) = [|y| < 100](y - z)^2$

такая модель будет игнорировать $y : |y| \geqslant 100$

можно подобрать другое число — гиперпараметр

3) Известно, что линейные модели часто обучаются градиентным спуском. В нём важно определить, при каких условиях нужно остановить обучение — это условие называется критерием останова. Ниже перечислено несколько критериев останова. Для каждого из них прокомментируйте, является ли он осмысленным или же приведёт к некорректной работе метода оптимизации. Через Q_t обозначается величина функционала ошибки после итерации t.

- $\|w^t - w^{t-1}\| < \varepsilon$; ✓ веса не сильно меняются при шаге
- $\|w^t\| < \varepsilon$; ✗ маленькие веса? не обязательно, что минимум в 0
- $\|w^t + w^{t-1}\| < \varepsilon$; ✗ не осмысленный
- $|Q_t| < \varepsilon$; ✓ приблизились к нулевой ошибке
- $|Q_t - Q_{t-1}| < \varepsilon$; ✓ ошибка не сильно меняется при шаге
- $Q_t - Q_{t-1} > \varepsilon$. ✗ не осмысленный

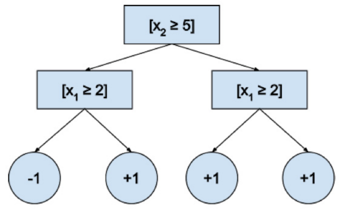
$\boxed{\text{✓ – можно использовать; ✗ – нельзя}}$

Задача 3

Задача 3 (2 балла). Рассмотрим следующую выборку с двумя признаками для задачи бинарной классификации:

x_1	3	4	1	0	0	1	4	6
x_2	6	5	7	6	2	1	3	1
y	+1	+1	-1	+1	-1	-1	-1	+1

Для этой выборки мы построили решающее дерево глубины два. Считайте, что в случае выполнения условия мы идём вправо.



Ответьте на вопросы:

1. Чему равна энтропия той части обучающей выборки, которая попала в самую левую листовую вершину? Используйте логарифм по основанию 2.
2. Чему равен прогноз дерева для объекта с признаками $(3, 2)$?
3. Как нужно изменить предикат в корне, чтобы для объекта из предыдущего пункта прогноз изменился на противоположный?

1) Туда попали:

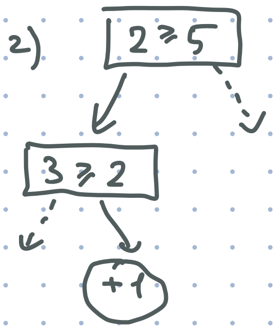
x_1	0	1
x_2	2	1
y	-1	-1

то есть $p_{-1} = 1$

$$H = p_{-1} \log_2 p_{-1} = 1 \cdot \log_2 2 = 0$$

Ответ: 0

2)



Ответ: +1

3) Никак, потому что в правой части дерева только положительные прогнозы (+1), а в левой части (3,2) точно попадает в +1

Задача 4

Задача 4 (2 балла). Ответьте на вопросы по решающим деревьям и композициям:

1. Построим бэггинг над случайными деревьями — то есть деревьями, у которых каждый предикат выбирается случайно (и признак, и порог берутся случайным образом независимо от данных). Деревья строятся, пока в каждом листе не окажется меньше 100 объектов. Что вы можете сказать про такие композиции? Они будут лучше, хуже или такими же по качеству, как и в классическом бэггинге над деревьями? Попробуйте провести рассуждения с использованием понятий смещения и разброса.
2. Перейдём к градиентному бустингу. Решаем задачу бинарной классификации. Пусть мы уже обучили $N-1$ дерево: $a_{N-1}(x) = \sum_{n=1}^{N-1} b_n(x)$. Следующее дерево мы собираемся обучать так:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - \frac{\partial L(b_{N-1}(x_i), z)}{\partial z}\bigg|_{z=y_i}\right) \to \min_{b_N(x)}$$

Найдите и исправьте все ошибки в этой формуле. Кратко поясните для каждой ошибки, почему она является таковой.

1) Бэггинг над случайными деревьями будет лучше, т.к. ошибка = смещение + разброс + шум

смещение итоговой модели такое же

$$\text{разброс}(a_n) = \frac{1}{N}(\text{разброс}(b_n)) + \underline{\text{ранжирование}(b_1, b_k)}$$

то есть модели $b(x)$ не должны быть похожи для более хорошей итоговой модели, и случайные признаки в каждом признаке нам в этом помогут

2) $$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(b_N(x_i) - \frac{\partial L(b_{N-1}(x_i), z)}{\partial z}\bigg|_{z=y_i}\right) \to \min_{b_N(x)}$$

y_i

$z = a_{N-1}$

т.к. корректирует ошибку итоговой модели, а не предыдущей, и частную производную берём по предсказанию модели, а не по итоговому ответу

Задача 5

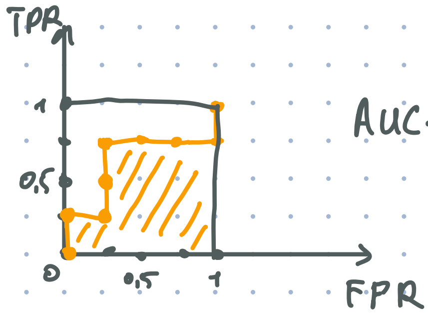
Задача 5 (2 балла). Вам выдали классификатор $b(x)$, и вам предстоит разобраться, насколько он хорош. Для этого у вас есть тестовая выборка из 8 объектов. Ниже указаны правильные ответы и уверенности модели в положительном классе:

$b(x)$	-0.1	0.3	1	3	3.5	4	5	7
y	+1	-1	-1	-1	+1	+1	-1	+1

Ответьте на вопросы:

1. Нарисуйте ROC-кривую и посчитайте AUC-ROC.
2. Укажите самый близкий к нулю порог t, при котором модель $a(x) = [b(x) \geqslant t]$ будет иметь максимальную полноту.
3. Чему будет равен AUC-ROC на этой же выборке для модели $b_1(x) = (b(x))^2 - 0.3$?

1) $TPR = \frac{TP}{TP + FN}$ $FPR = \frac{FP}{FP + TN}$



$AUC\text{-}ROC = \frac{10}{16} = \frac{5}{8} = 0{,}625$

2) $t = -0{,}1$, всё занесёт в класс +1,

полнота recall = 1

3)

$b(x)$	$-0{,}1$	$0{,}3$	1	3	$3{,}5$	4	5	7
$b^2(x) - 0{,}3$	$0{,}01 - 0{,}3$	$0{,}09 - 0{,}3$	$0{,}7$	$8{,}7$	?	$15{,}7$	$24{,}7$	$48{,}7$

что-то между 8,7 и 16,7, порядок не поменяется

Порядок не изменился $\Rightarrow$ AUC-ROC не изменится:

AUC-ROC = 0,625