

Фамилия, имя, номер группы:

.....

Внесите сюда ответы на тест:

Вопрос	1	2	3	4	5	6	7	8	9	10
Ответ										

Табличка для проверяющих работу:

Тест	1	2	3	4	5	Итого



Совы не то, чем они кажутся

Великан из Твин Пикс (1990)

Это нулевой вариант мидтёрма. Он нужен для того, чтобы его формат не стал для вас сюрпризом. Работа состоит из трёх частей: тестовая, задачи и ответы на открытые вопросы. Списывание карается обнулением работы. Удачи!

Часть первая: тестовая


Дайте ответ на 10 тестовых вопросов. Каждый вопрос стоит 3 балла. Никакие дополнительные пояснений в этой части работы от вас не требуются.

Вопрос 1. У Ратибора есть онлайн-кинотеатр с огромным количеством фильмов. Ратибор знает жанры для каждого из них. В понедельник база фильмов пополнится новыми релизами студии Disney. К сожалению, новые фильмы придут без указанных жанров. Ратибор хочет предсказать их с помощью машинного обучения. Какую задачу ему предстоит решить?

☐ A Кластеризация

☐ C Регрессия

☐ E Рекомендации

☐ B Классификация 


☐ D Ранжирование

☐ F Нет верного ответа.

Вопрос 2. Качество чего оценивается с помощью кросс-валидации или отложенной выборки?

☐ A Метода обучения параметров для конкретной модели

☐ C Конкретного набора признаков


☐ E Модели с конкретным набором параметров 

☐ B Конкретной обучающей выборки



☐ D Регуляризации

☐ F Нет верного ответа.




Вопрос 3. Допустим, мы обучаем линейную модель на MSE с L_2 -регуляризатором. Как будет разумно измерять её ошибку на тестовой выборке?

- ☐ A Из MSE надо вычесть значение L_2 -регуляризатора
- ☐ B По MSE с L_2 -регуляризатором
- ☐ C По значению L_2 регуляризатора
- ☐ D По MSE 
- ☐ E Подойдёт любой из вышеперечисленных способов
- ☐ F Нет верного ответа.

Вопрос 4. Выберите верные утверждения про стохастический градиентный спуск

- ☐ A В SGD, скорее всего, требуется больше итераций для сходимости, чем в обычном градиентном спуске 
- ☐ B SGD НЕ используется в современном машинном обучении
- ☐ C В SGD одна итерация требует больше вычислений, чем в обычном градиентном спуске
- ☐ D В SGD, скорее всего, требуется меньше итераций для сходимости, чем в обычном градиентном спуске
- ☐ E В SGD одна итерация требует меньше вычислений, чем в обычном градиентном спуске 
- ☐ F Нет верного ответа.


Вопрос 5. Какие из способов приведённых ниже можно использовать для работы с пропусками в действительных переменных при обучении линейных моделей?

- ☐ A Если пропусков очень много, выкинуть переменную 
- ☐ B Заполнить пропуски аномальным значением
- ☐ C Выделить пропуски в отдельную категорию и сделать ONE-преобразование
- ☐ D Заполнить нулями 
- ☐ E Заполнить пропуски медианами, посчитанными по каждой колонке 
- ☐ F Нет верного ответа.




Вопрос 6. Драгомир пытается предсказать продажи видео-игр. Он предсказывает продажи по возрасту игры, x . Целевая переменная y — количество продаж. Драгомир оценил линейную регрессию:

$$\ln y = 5 - 6 \cdot \ln x.$$




Предположим, что мы отгружаем на рынок новую партию игры, выпущенной в прошлом году. Спрогнозируйте, сколько экземпляров этой игры будет продано?

- ☐ A 4
- ☐ B 20
- ☐ C 148 
- ☐ D 5
- ☐ E 0
- ☐ F Нет верного ответа.


Вопрос 7. Выберите все верные утверждения про регуляризацию линейных моделей.

- ☐ A Регуляризация штрафует модель за слишком **большие** по модулю значения коэффициентов 
- ☐ B L_2 -регуляризация зануляет коэффициенты
- ☐ C Регуляризация штрафует модель за слишком **маленькие** по модулю значения коэффициентов
- ☐ D L_1 -регуляризация зануляет коэффициенты 
- ☐ E Регуляризатор обычно приплюсовывают к функции потерь 
- ☐ F Нет верного ответа.



Вопрос 8. Велимудр обучает метод ближайших соседей для классификации молекул на токсичные и обычные. Выберите все верные утверждения про KNN:

- ☐ A Велимудр может подобрать оптимальное число соседей с помощью кросс-валидации 
- ☐ B Для обучения KNN важно нормировать данные 
- ☐ C KNN на этапе обучения подбирает оптимальное число соседей с помощью градиентного спуска
- ☐ D KNN на этапе обучения заготавливает выборку 
- ☐ E Если взять число соседей очень большим, модель точно не переобучится
- ☐ F Нет верного ответа.

Вопрос 9. Рагнеда классифицирует фотографии. Она хочет обучить логистическую регрессию, которая будет отличать гусей от уток. Пусть \hat{p}_i — это предсказание вероятности того, что на фото гусь, $y_i = 1$, если на фото гусь. Какую функцию потерь надо минимизировать Рагнеде для обучения модели?

- ☐ A $p_i \cdot \ln(1 - y_i) + (1 - p_i) \cdot \ln y_i$
- ☐ B $y_i \cdot \ln(1 - p_i) + (1 - y_i) \cdot \ln p_i$
- ☐ C $y_i \cdot \ln p_i + (1 - y_i) \cdot \ln(1 - p_i)$ 
- ☐ D $p_i \cdot \ln y_i + (1 - p_i) \cdot \ln(1 - y_i)$
- ☐ E $(y_i - \hat{p}_i)^2$
- ☐ F Нет верного ответа.

Вопрос 10. Выберите все метрики классификации, которые не зависят от выбора порога

- ☐ A Доля правильных ответов (accuracy)
- ☐ B Точность (precision)
- ☐ C Площадь под ROC-кривой (ROC-AUC) 
- ☐ D f-мера
- ☐ E Площадь под PR-кривой (PR-AUC) 
- ☐ F Нет верного ответа.

Часть вторая: открытые вопросы

Эта часть состоит из открытых вопросов. На них необходимо дать краткие, но ёмкие ответы. За каждый ответ вы можете получить 10 баллов.

Вопрос 11. Предложите для каждой из перечисленных ниже задач, как сформулировать их в терминах машинного обучения: укажите, что будет являться объектом и целевой переменной, а также напишите тип задачи.

1. Мы делаем сервис доставки еды. Мы хотим показывать пользователю, который собирает заказ, через сколько этот заказ будет доставлен.
2. Служба поддержки в одном сервисе такси не справляется с нагрузкой. Однако некоторые запросы связаны с тем, что клиент забыл в машине документы или вещи, поэтому надо быстрее помочь ему связаться с водителем. Мы хотим с помощью машинного обучения определять по запросу, стоит ли рассматривать его в первую очередь.

Вопрос 12. Машинлёрнерша Рагнеда не доверяет пакетным реализациям алгоритмов машинного обучения. Поэтому она написала свой собственный градиентный спуск. Для того, чтобы делать шаг градиентного спуска, она использовала следующие формулы.

$$w_t = w_{t-1} + (\nabla Q(w_t))^2$$

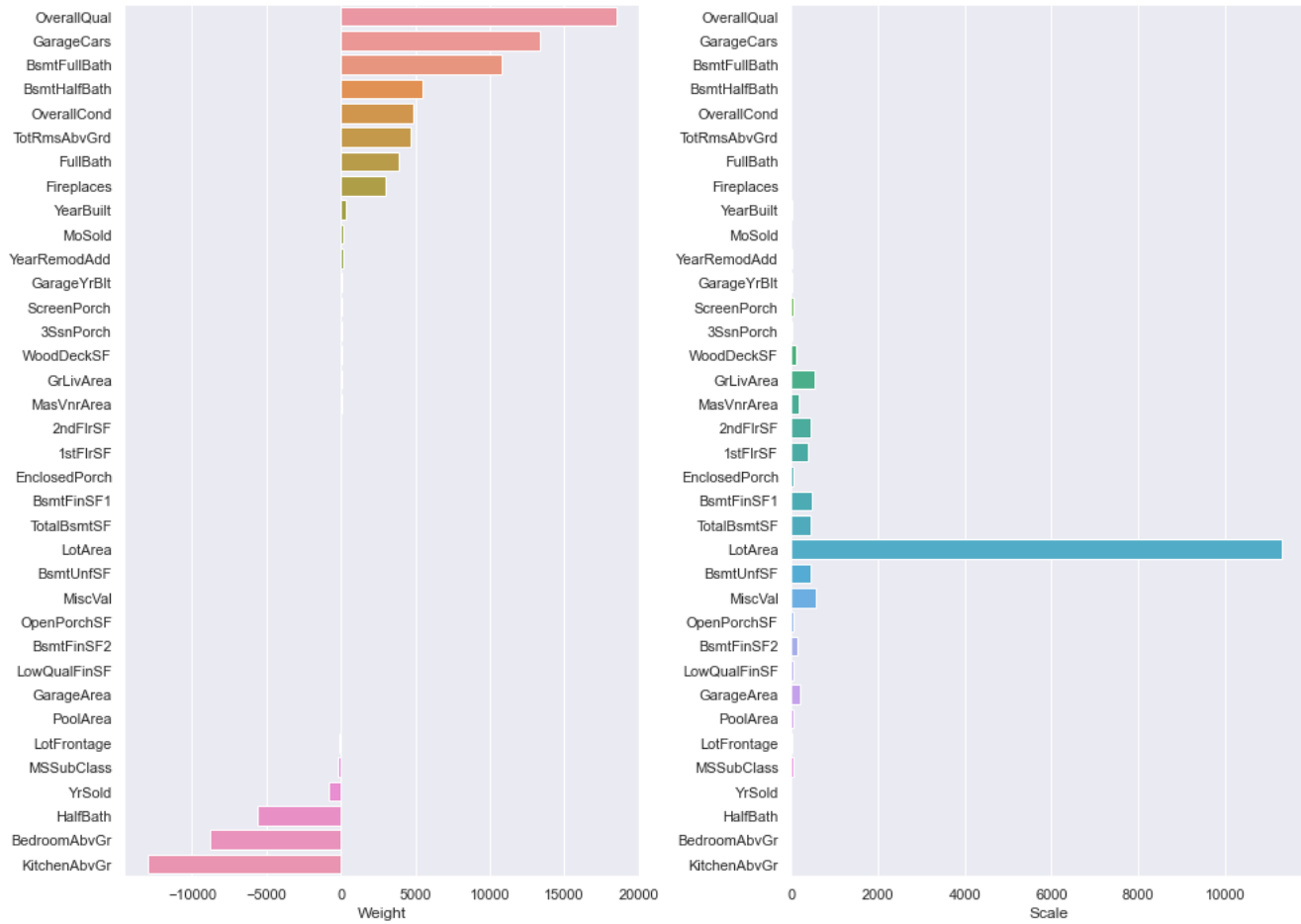
Какие ошибки вы тут видите? Для каждой объясните, к каким последствиям и почему она приведёт, а также как это исправить.

Вопрос 13. Объясните мем



Вопрос 14. Результатом обучения логистической регрессии является вектор весов w . Если нам дан объект x , как посчитать вероятность того, что он относится к положительному классу? Объясните все компоненты в формуле, которую запишите.

Вопрос 15. Доброгнева обучает линейную регрессию для предсказания цен на недвижимость. Она хочет, чтобы коэффициенты перед соответствующими факторами отражали их важность для итогового прогноза. На левой картинке Доброгнева построила значения коэффициентов. На правой она визуализировала стандартное отклонение каждого фактора. Добилась ли Доброгнева нужной ей интерпретации коэффициентов? Что они в действительности отражают? Как это исправить?



Часть третья: задачи

Решите все задания. Все ответы должны быть обоснованы. Решения должны быть прописаны для каждого пункта. Рисунки должны быть чёткими и понятными. Все линии должны быть подписаны. За решение каждой задачи вы можете получить 10 баллов.

Вопрос 16. Винни-Пух и Пятачок классифицируют пчёл на правильных и неправильных. У них есть выборка X , состоящая из 7 объектов, и классификатор $a(x)$, предсказывающий оценку принадлежности пчелы к правильным. Предсказания $a(x)$ и реальные метки пчёл представлены ниже:

$b(x)$	0.2	0.6	0.3	0.7	0.5	0.9	0.6
y	-1	+1	-1	-1	+1	+1	-1

1. Пятачок считает, что порог $t = 0.55$ самый лучший. Вычислите precision и recall для такого классификатора.
2. Винни-Пух согласен с Пятачком, но у него в голове опилки. При подсчёте тех же метрик он перепутал местами предсказания классификатора и реальные метки. Какие значения метрик он получит?
3. Что будет происходить с precision и recall, если Винни-Пух и Пятачок увеличат порог?

Вопрос 17. На плоскости расположены колонии рыжих и чёрных муравьёв. Рыжих колоний три и они имеют координаты $(-1, -1)$, $(1, 1)$ и $(3, 3)$. Чёрных колоний тоже три и они имеют координаты $(2, 2)$, $(4, 4)$ и $(6, 6)$.

- а) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод одного ближайшего соседа (надо явно указать как будет классифицирована каждая точка плоскости и провести разделяющую поверхность).
- б) Поделите плоскость на «зоны влияния» рыжих и чёрных муравьёв, используя метод трёх ближайших соседей.
- в) С помощью кросс-валидации с выкидыванием отдельных наблюдений выберите оптимальное число соседей k перебрав $k \in \{1, 3, 5\}$. Целевой функцией является количество верных предсказаний (accuracy).