

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

Позитивная мотивация — явно не мой конёк, и мы все умрём.
 Всё уже было до нас, можно выдохнуть страх, и уставить глаза в небосклон.
 Если этой контрольной и сопротивляться, то не с печальным лицом.
 Всё повторится не раз, но мы живы сейчас, нами рано удобрять чернозём.
Охххутирон и Тося Чайкина про мидтёрм по ML (2022)

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 императора Куско есть база данных обо всех бизонах, пасущихся на просторах его необъятной империи. Для удобства император хочет разбить всех бизонов на 30 стад таким образом, чтобы в каждом были похожие по своим характеристикам особи. Какую задачу решает император солнца?

- | | | |
|---|---|---|
| <input type="checkbox"/> A Регрессия | <input type="checkbox"/> C Кластеризация | <input type="checkbox"/> E Рекомендации |
| <input type="checkbox"/> B Классификация | <input type="checkbox"/> D Ранжирование | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 2. Для чего можно использовать кросс-валидацию?

- | | | |
|--|--|--|
| <input type="checkbox"/> A Для подбора лучшего коэффициента регуляризации | <input type="checkbox"/> C Для подбора оптимального способа измерения ошибки модели на новых данных | <input type="checkbox"/> E Для улучшения качества обучающей выборки |
| <input type="checkbox"/> B Для оценки того, как модель будет работать на новых данных | <input type="checkbox"/> D Для выбора лучшего типа модели | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 3. Выберите все верные утверждения про переобучение (overfitting).

- | | | |
|--|---|---|
| <input type="checkbox"/> A Метод ближайших соседей на этапе обучения просто запоминает всю выборку, поэтому он не переобучается | <input type="checkbox"/> C Если качество модели на тестовой выборке ниже, чем на обучающей, скорее всего, модель переобучилась | <input type="checkbox"/> E Чтобы избежать переобучения, качество модели достаточно измерить на той же самой выборке, на которой обучается модель |
| <input type="checkbox"/> B L_2 -регуляризатор добавляют в линейную модель, чтобы не дать ей переобучиться | <input type="checkbox"/> D Валидационную выборку выделяют, чтобы подобрать на ней гиперпараметры | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 4. Выберите все верные утверждения про градиентный спуск

- | | | |
|---|--|--|
| <input type="checkbox"/> A Градиентный спуск гарантированно находит глобальный оптимум | <input type="checkbox"/> C Градиентный спуск не гарантирует нахождение глобального оптимума | <input type="checkbox"/> E При обучении логистической регрессии с помощью градиентного спуска напрямую оптимизируют долю верных ответов, ассигасу |
| <input type="checkbox"/> B Метод ближайших соседей обучается градиентным спуском | <input type="checkbox"/> D Масштабирование признаков ускоряет работу градиентного спуска | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 5. Какие из сопособов приведённых ниже можно использовать для борьбы с выбросами при обучении линейных моделей?

- | | | |
|--|---|---|
| <input type="checkbox"/> A Заменить выбросы на какой-нибудь квантиль, например, медиану | <input type="checkbox"/> C Использовать функцию потерь, которая нечувствительна к выбросам (Huber Loss и тп) | <input type="checkbox"/> E Выбросы — это категориальные переменные, для них можно сделать ONE-преобразование |
| <input type="checkbox"/> B Выбросить все наблюдения-выбросы | <input type="checkbox"/> D Стандартизировать данные | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 6. Рассмотрим выборку, состоящую из одного признака, и линейную модель над этой выборкой. Говорят, что эту модель можно изобразить как прямую. А в каких координатах? Признак обозначается x , целевая переменная y , веса модели — w_0 и w_1

- | | | |
|--|--|---|
| <input type="checkbox"/> A x | <input type="checkbox"/> C (w_1, y) | <input type="checkbox"/> E (x, y, w_0, w_1) |
| <input type="checkbox"/> B (x, y) | <input type="checkbox"/> D (w_0, w_1) | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 7. Какие из функций ниже логично использовать для оценки качества линейной модели на тестовой выборке?

- | | | |
|--|---|--|
| <input type="checkbox"/> A $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$ | <input type="checkbox"/> C $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{j=1}^k w_j^2$ | <input type="checkbox"/> E $\sum_{j=1}^k w_j^2$ |
| <input type="checkbox"/> B $\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $ | <input type="checkbox"/> D $\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i + \sum_{j=1}^k w_j^2$ | <input type="checkbox"/> F Нет верного ответа. |

Вопрос 8. Льюис и Кларк идут с экспедицией на Дикий Запад и предсказывают население каждого нового племени с помощью метода двух ближайших соседей. Все прогнозы строятся с учётом того, каким получилось расстояние между племенами.






Если нарисовать регион на карте, окажется что племя Сиу численностью 10 тыс. живёт по координатам $(0, 0)$. Племя Ото численностью 20 тыс. живет по координатам $(3, 5)$. Племя Миссури живёт по координатам $(0, 1)$. Каким будет примерный прогноз для его численности?

- ☐ A 12 тыс. ☐ C 20 тыс. ☐ E 14 тыс.
☐ B 15 тыс. ☐ D 10 тыс. ☐ F Нет верного ответа.

Вопрос 9. Мир пришёл на Великие Равнины. Несколько индейских племён зарыли топор войны и решили раскошегарить трубку мира. Шаман племени Апачи умеет предсказывать, будет ли заключён мирный договор. Он делает с помощью логистической регрессии, обученной на предыдущих конфликтах:

$$P(y_i = 1 | x_i) = \frac{1}{1 + \exp(100 - 20 \cdot x_i)},$$

где x — время в минутах, которое вожди провели за трубкой мира. Выберите все верные утверждения об этой модели.

- ☐ A Если трубку мира не раскурят вообще, мир наступит с очень низкой вероятностью
☐ B Если вожди курили трубку 5 минут, мир будет заключен с вероятностью 0.25
☐ C С каждой дополнительной минутой за трубкой мира, вероятность заключить мир уменьшается
☐ D     
☐ E С каждой дополнительной минутой за трубкой мира, вероятность заключить мир растёт
☐ F Нет верного ответа.

Вопрос 10. Мы хотим обучать модель, напрямую максимизируя полноту. Выберите в списке ниже все причины, почему это плохая идея?

- ☐ A Максимальную полноту можно получить тривиальной моделью, которая все объекты относит к положительному классу
☐ B Полнота зависит от порога
☐ C Полноту нельзя продифференцировать
☐ D Полнота очень плохо работает, если выборка несбалансированная
☐ E Максимальную полноту можно получить тривиальной моделью, которая все объекты относит к отрицательному классу
☐ 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. Фармкомпания Wachowski Inc. производит два вида таблеток: красные и синие. К сожалению, конвейерная лента на заводе иногда выходит из строя и в одну блистерную упаковку попадают таблетки разного цвета. Мы хотели бы отбраковывать такие упаковки.
2. Студент решил воспользоваться высокими технологиями при подготовке к экзамену и строит модель, которая предскажет его оценку.

Вопрос 12. Несколько архитекторов строят мемориал в честь вождя племени Оглала, Неистового Коня. Архитекторы хотят оптимизировать поставки стройматериалов. Для этого они предсказывают время прибытия поставки по разным факторам: трафик на дорогах, погода, загруженность складов и т.п.

Архитекторы хотят понять, какие факторы влияют на доставку сильнее всего. Они обучают линейную регрессию и хотят занулить переменные перед всеми факторами, которые несущественны для их прогнозов. Используется следующая функция потерь:

$$Q(w) = \frac{1}{n} \sum_{i=1}^n (a(x_i) - y_i) + \lambda \cdot \sum_{i=1}^d w_i^2 \rightarrow \max_w$$

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

Вопрос 13. Объясните мем

model: overfits on training data

world: new data

model:



Вопрос 14. Линейные модели в машинном обучении обучаются с помощью градиентного спуска. На каждой итерации вектор весов w изменяется по какой-то формуле. Выпишите эту формулу. Объясните в ней каждую компоненту и обозначение.

Вопрос 15. Вождь Маквагабо¹ хочет классифицировать бусы из бисера. У него есть 1000 синих бус и 10^6 красных.

Почему для оценки качества классификации не получится использовать долю правильных ответов, ассигасу? Приведите примеры моделей, где эта метрика показывает неадекватный результат. Что можно сделать для корректной оценки работы модели в таком случае?

¹Есть теория, что Голландцам землю за бусы продали не местные индейцы, а шайка мошенников, которая даже не владела этой землёй.

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. Индейские Шаманы предсказывают стоимость недвижимости в Сиэтле. Шаман Одэхингум (лёгкое колебание воды) использует метод ближайших соседей. Шаман Пэпина (виноградная лоза, растущая вокруг дуба) использует линейную регрессию. Шаман Апониви (где ветер вырывает промежуток с корнем) использует случайный лес.

В тестовой выборке у них есть три дома y_i . Для каждого из них шаманы построили прогнозы.

настоящие y_i	4	7	1
KNN	4	5	2
линейная регрессия	5	6	0
случайный лес	1	1	1

1. Найдите для прогнозов MAE, MSE, RMSE и MAPE.
2. Объясните, зачем от MSE обычно переходят к RMSE.
3. Объясните, почему MAE считается более устойчивой к выбросам.

Вопрос 17. Покахонтас спасла от смерти капитана Джона Смита, а затем решила задачу классификации. У неё получились следующие прогнозы.

y_i	\hat{p}_i
1	0.9
0	0.1
0	0.75
0	0.56
0	0.2
1	0.37
0	0.25

1. Бинаризируйте ответ по порогу t и посчитайте точность и полноту для $t = 0.3$ и для $t = 0.8$.
2. Постройте ROC-кривую и найдите площадь под ней.