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:

Задачи	1	2	3	4	5	6	7	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.

Oxxxymiron 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. Архимед хочет отбить римскую атаку на Сиракузы. С помощью беспилотника Архимед делает съёмки римских кораблей, а затем специальным алгоритмом компьютерного зрения пытается предсказать, сколько римских солдат находится на каком корабле. Какую задачу решает Архимед?

- [A] Регрессия
- [B] Классификация
- [C] Кластеризация
- [D] Ранжирование
- [E] Рекомендаии
- [F] Нет верного ответа.

Вопрос 2. Что такое гиперпараметр?

- [A] Параметр, который оказывает ключевое влияние на производительность модели
- [B] Ровно то же самое, что и параметр
- [C] Параметр, чьё оптимальное значение нельзя подобрать по обучающей выборке
- [D] Параметр, который является многомерным
- [E] Параметр, от которого выход модели зависит нелинейно
- [F] Нет верного ответа.

Вопрос 3. В чём потенциальный недостаток кросс-валидации по двум блокам?

- [A] Оценка качества модели на новых данных может оказаться очень заниженной из-за того, что тестирование проводится на слишком маленькой выборке
- [B] Оценка качества модели на новых данных может оказаться очень заниженной из-за того, что тестирование проводится на слишком большой выборке
- [C] Оценка качества модели на новых данных может оказаться очень заниженной из-за того, что обучение проводится на выборке сильно больше исходной
- [D] Оценка качества модели на новых данных может оказаться очень заниженной из-за того, что обучение проводится на выборке сильно меньше исходной
- [E] Нужно обучать слишком много моделей, вычислительных мощностей для этого может не хватить, новые сервера в Россию из-за санкций не поставляют, поэтому кросс-валидация по двум блокам не имеет никакого смысла
- [F] Нет верного ответа.

Вопрос 4. Что из этого формула для шага в градиентном спуске?

- [A] $w_t = w_{t-1} - \eta \cdot \nabla L(w_t)$
- [B] $w_t = w_{t-1} + \eta \cdot \nabla L(w_t)$
- [C] $w_t = w_{t-1} - \eta \cdot \nabla L(w_{t-1})$
- [D] $w_t = w_{t-1} - \eta \cdot \nabla L(w_0)$
- [E] $w_t = w_{t-1} + \eta \cdot \nabla L(w_{t-1})$
- [F] Нет верного ответа.

Вопрос 5. Какие из сопособов приведённых ниже можно использовать для работы с пропусками в категориальных 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 OHE-преобразование
- [D] Заполнить нулями
- [E] Заполнить пропуски медианами, посчитанными по каждой колонке
- [F] Нет верного ответа.

Вопрос 6. Бог плодородия Дионис спустился с Олимпа вкусить вина свежего урожая. Зевс пытается понять, сколько дней будет идти кутёж Диониса. Для этого он использует линейную регрессию, обученную на предыдущих кутежах:

$$y_i = 7 + 0.5 \cdot x_1 + 0.2 \cdot x_2,$$

где x_1 — качество вина по десятибальной шкале, x_2 — количество событульников. Выберите все верные утверждения об этой модели.

- [A] Каждые дополнительные 5 собутыльников будут затягивать кутёж на день
- [B] In vino veritas, in aqua sanitas
- [C] Кутёж затянется минимум на 14 дней
- [D] Кутёж затянется минимум на 7 дней
- [E] Каждый дополнительный собутыльник будет затягивать кутёж на один день
- [F] Нет верного ответа.

Вопрос 7. Что из этого можно использовать для регуляризации? Под регуляризацией мы понимаем штрафование моделей за сильно отличающиеся от нуля веса.

[A] $|w_1| + |w_2| + ... + |w_d|$

[B] $\frac{1}{|w_1|} + \frac{1}{|w_2|} + ... + \frac{1}{|w_d|}$

[C] $w_1^2 + w_2^2 + ... + w_d^2$

[D] $w_1^3 + w_2^3 + ... + w_d^3$

[E] $|w_1|^3 + |w_2|^3 + ... + |w_d|^3$

[F] Нет верного ответа.

Вопрос 8. У нас есть 2 класса. Классификатор предсказывает, что объект равновероятно относится к каждому из них. Какое значение принимает logloss на этом объекте?

[A] $-\log 2$

[B] $-0.5 \log 2$

[C] $-2 \log 2$

[D] $-2 \log 0.5$

[E] $\log 2$

[F] Нет верного ответа.

Вопрос 9. Леонид предсказывает цены на квартиры в Спарте с помощью метода ближайших соседей. При построении предсказания он хочет учитывать расстояние до соседей. Какая из формул ниже поможет Леониду корректно построить прогноз? Все суммы ищутся по ближайшим соседям, y_j — цена квартиры, ρ_j — расстояние от объекта для которого строится предсказание до соотвествующего соседа.

[A] $\frac{1}{k} \sum_{j=1}^{k} y_j$

[B] $\frac{\sum_{j=1}^{k} \rho_j \cdot y_j}{\sum_{j=1}^{k} \rho_j}$

[C] $\frac{1}{k} \sum_{j=1}^{k} \frac{y_j}{\rho_j}$

[D] $\frac{\sum_{j=1}^{k} \rho_j \cdot y_j}{\sum_{j=1}^{k} \frac{1}{\rho_j}}$

[E] $\frac{\sum_{j=1}^{k} \frac{1}{\rho_j} \cdot y_j}{\sum_{j=1}^{k} \rho_j}$

[F] Нет верного ответа.

Вопрос 10. Какие из метрик перечисленных ниже используются для решения задачи классификации?

[A] MSE

[B] f-мера

[C] MAPE

[D] logloss

[E] ROC-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. В заповеднике голод — белки часто остаются некормленными, так как зайцы оказываются слишком прожорливыми и съедают чужую еду. Мы приняли решение поставить кормушки с фотоловушками, которые бы открывались для того животного, которое подошло к кормушке, чтобы ограничить ненасытных зайцев.

2. По статистике некоторая доля сотрудников одной компании ежедневно опаздывает из-за пробок, сокращая свой рабочий день на некоторое количество времени. Мы решили разработать приложение, которое будет подсказывать сотрудникам, на какое время им стоит запланировать выход из дома, чтобы не опоздать на работу.

Вопрос 12. Аполоний обожает логистическую регрессию. Поэтому он решил закодить её самостоятельно, без всяких пакетов. Он записал модель следующим образом:

$$b(x_i) = \sigma(b + \langle x_i, w \rangle)$$
$$\sigma(t) = \frac{1}{1 - exp(t)}$$

Для обучения Аполоний использует функцию потерь

$$L(b(x_i), y_i) = b(x_i \cdot \ln y_i + (1 - b(x_i) \cdot \ln(1 - y_i)$$

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

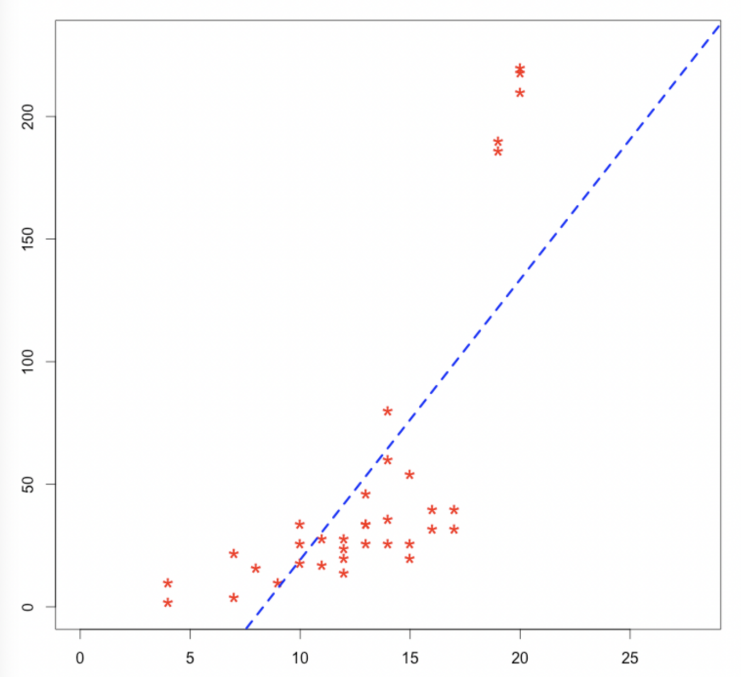
Вопрос 13. Объясните мем

Вопрос 14. Пигмалион оживил Галатею, вырезанную из слоновой кости! Теперь они хотят обучить метод ближайших соседей для классификации горных пород. В качестве признаков используются: вес пород, размер, цвет и тп.

Галатея хочет строить свои прогнозы с учётом того, каким получилось расстояние между объектами. Пигмалион хочет строить свои прогнозы с помощью метода большинства. Выпишите формулы, которые будут использованы для расчёта прогнозов. Объясните в них каждую компоненту и обозначения.

Вопрос 15. Аргонавты плавают вокруг нимф и пытаются понять, сколько их нужно слушать, чтобы весь экипаж зачаровало. По собранным данным оценивается линейная регрессия. В качестве объясняющей переменной, x используется длина песни, в качестве объясняемой число зачарованных аргонавтов, y. В качестве функции потерь используется MSE.

Ясон отложил по оси x длину песни, а по оси y число зачарованных аргонавтов. Поверх облака точек он нарисовал линейную регрессию. Получилась такая картинка:



С какими проблемами при обучении модели столкнулся Ясон? Предложите как минимум два способа исправить возникшую проблему.

Часть третья: задачки

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. Мойры предсказывают судьбу. Клото использует для этого метод ближайших соседей. Ла́хеси использует линейную регрессию, а А́тропо случайный лес. В тестовой выборке у них есть три наблюдения y_i. Для каждого из них мойры построили прогнозы.

настоящие y_i	1	2	3
KNN	2	3	1
линейная регрессия	2	3	4
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. Константин основал Константинополь, а затем решил задачу классификации. У него получились следующие прогнозы.

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

1. Бинаризуйте 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.