

Решается задача бинарной классификации с классами $\{0, 1\}$. Алгоритм выдаёт некоторую оценку, принадлежащую отрезку $[0, 1]$, что объект относится к классу 1. Качество алгоритма $\text{ROC-AUC}=0.5$. Как меняется значение метрики качества, если возвести каждое предсказание в квадрат?

Выберите один ответ:

- ☐ Зависит от данных: может как улучшиться, так и ухудшиться
- ☐ Улучшится
- ☐ Ухудшится
- ☒ Не меняется ✓

Ваш ответ верный.

Правильный ответ: Не меняется

Дан текст: "Но не каждый хочет что-то исправлять :("

После некоторой обработки получилось:

```
['но', '', 'не', '', 'каждый', '', 'хотеть', '', 'что-то', '', 'исправлять', ':(\n']
```

Выберите все шаги, которые были сделаны с исходным текстом:

Выберите один или несколько ответов:

- ☐ Стемминг (stemming)
- ☒ Лемматизация (lemmatization) ✓
- ☐ Векторизация (vectorization)
- ☒ Токенизация (tokenization) ✓

Ваш ответ верный.

Правильные ответы: Токенизация (tokenization), Лемматизация (lemmatization)

Василий пытается отправить СМС в условиях слабой мобильной связи. Телефон делает попытки отправить СМС до тех пор, пока это не удастся. Известно, что вероятность удачной попытки равна 0,05 и не зависит от предыдущих попыток. Чему равно математическое ожидание числа сделанных попыток?

Выберите один ответ:

- ☐ 5
- ☐ 1
- ☒ 20 ✓
- ☐ 10

Ваш ответ верный.

Правильный ответ: 20

Пусть объекты в данных имеют два числовых признака. Тогда эти объекты можно изобразить на двумерной плоскости. В нашей задаче дано 1000 объектов, каждый из которых описывается парой признаков (x_1, x_2) равномерно распределенных на единичной окружности.

Среди перечисленных ниже утверждений найдите ошибочные:

Выберите один или несколько ответов:

- ☐ Величины x_1 и x_2 зависимы, но не линейно
- ☒ Величины x_1 и x_2 линейно зависимы ✓
- ☒ Величины x_1 и x_2 зависимы, поэтому при обучении любого алгоритма машинного обучения на наших данных один из признаков: x_1 или x_2 можно удалить. ✓
- ☐ Значение коэффициента корреляции Пирсона между величинами x_1 и x_2 мало
- ☒ Величины x_1 и x_2 независимы ✓

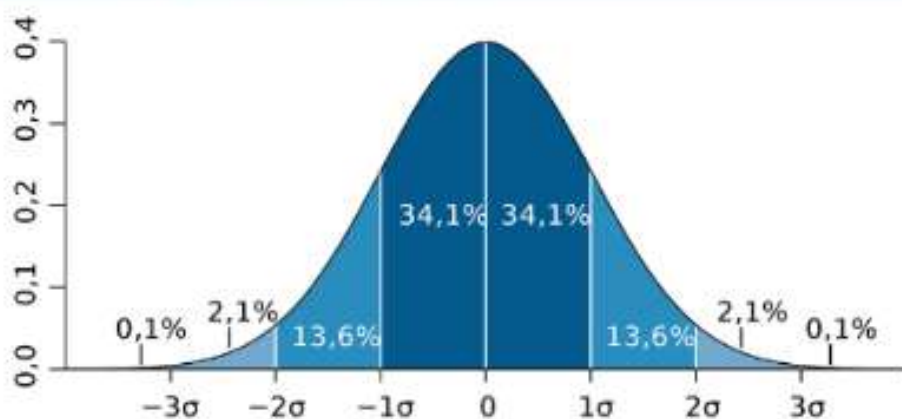
Ваш ответ верный.

Правильные ответы: Величины x_1 и x_2 линейно зависимы

, Величины x_1 и x_2 независимы

, Величины x_1 и x_2 зависимы, поэтому при обучении любого алгоритма машинного обучения на наших данных один из признаков: x_1 или x_2 можно удалить.

Известно распределение веса африканских слонов. Какие из перечисленных ниже утверждений верны, если распределение является нормальным со средним значением 6 тонн и среднеквадратичным отклонением 500 кг?



- ☐ 47.7 % африканских слонов весит от 6 до 7 тонн
- ☐ 13.6 % африканских слонов весит от 5 до 6 тонн
- ☐ 0.1 % африканских слонов весит больше 7.5 тонн
- ☒ Вес 68.2 % африканских слонов находится между 5.5 и 7 тонн ❌

Ваш ответ неправильный.

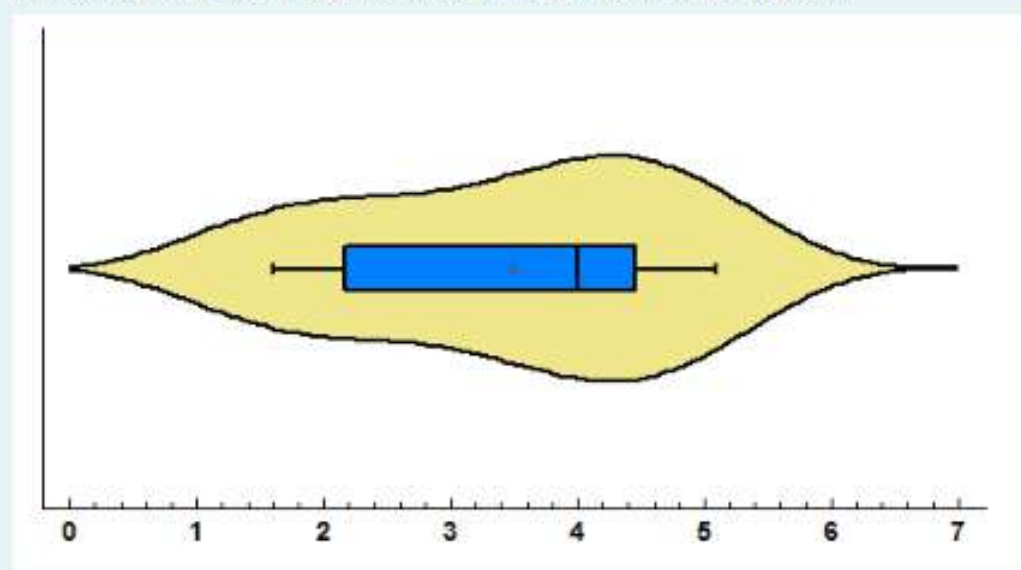
Правильные ответы:

0.1 % африканских слонов весит больше 7.5 тонн,

47.7 % африканских слонов весит от 6 до 7 тонн.

Скрипичный график (violin plot) - это визуализация, представляющая собой комбинацию "ящика с усами" и ядерной оценки плотности. Из скрипичного графика мы можем извлечь ту же информацию, что и из "ящика с усами": медиана - вертикальный отрезок внутри прямоугольника; интерквартильный диапазон - задается вертикальными сторонами прямоугольника; усы (отрезки, идущие в обе стороны от прямоугольника) - задают величины $Q_1 - 1.5 \cdot IQR$ и $Q_3 + 1.5 \cdot IQR$, где $IQR = Q_3 - Q_1$, Q_1, Q_3 - первая и третья квартили. Наблюдения, выходящие за пределы отрезка $(Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR)$, объявляются выбросами.

Ниже изображен скрипичный график для некоторого набора данных. Выберите два верных утверждения относительно этих данных, анализируя график.



- ☒ медиана набора данных равна 4 ✓
- ☐ Ядерная оценка плотности имеет более двух мод
- ☐ Интерквартильный размах равен 7
- ☒ В данных есть выбросы ✓

Ваш ответ верный.

Правильные ответы:

медиана набора данных равна 4,

В данных есть выбросы

2) Сколько кошек среди не прошедших полосу препятствий имели инструктора с уровнем образования "high school"?

Ответ: 35



Правильный ответ: 35

3) Сколько диких кошек среди прошедших полосу препятствий не проходили специальный курс подготовки?

Ответ: 152



Правильный ответ: 152

4) Чему равна медиана баллов, выставленных первым судьей?

Ответ: 66



Правильный ответ: 66

5) Найдите межквартильный размах баллов третьего судьи (третья квартиль минус первая квартиль) для домашних кошек, не проходивших специальный курс подготовки.

Комментарий: для вычисления квартилей дискретного распределения используйте интерполяцию меньшим значением (lower interpolation). Это означает, что если искомая квартиль лежит между двумя измерениями i и j , то значение квартили равно i .

Ответ: 20



Правильный ответ: 20

Задание 4 (0.7 баллов).

а) (0.3 балла). Далее используйте только категориальные столбцы. Закодируйте их с помощью One-hot encoding с учетом того, что мы не хотим получить мультиколлинеарности в новых данных. Сколько получилось числовых столбцов из исходных категориальных? Кодировать и `df_train`, и `df_test`.

Ответ:



Правильный ответ: 13

б) (0.4 балла). Попробуем по характеристикам кошки (бывшие категориальные, а теперь - числовые столбцы) предсказать, прошла она полосу препятствий или нет.

Сформируйте из `df_train` матрицу объект-признак `X` и вектор ответов `y`.

Обучите решающее дерево (`DecisionTreeClassifier` из библиотеки `sklearn.tree`) глубины 5 с энтропийным критерием информативности на закодированных в пункте а) тренировочных данных по кросс-валидации с тремя фолдами, метрика качества - `roc_auc`.

Чему равен `roc_auc`, усредненный по фолдам? Ответ округлите до десятых.

Комментарий: остальные гиперпараметры дерева оставьте дефолтными (`splitter='best'`, `min_samples_split=2`, `min_samples_leaf=1`, `min_weight_fraction_leaf=0.0`, `max_features=None`, `random_state=None`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `class_weight=None`, `ccp_alpha=0.0`)

Ответ:



Правильный ответ: 0.7

а) (0.25 балла). Подберите глубину решающего дерева (`max_depth`), перебирая глубину от 2 до 20 с шагом 1 и используя перебор по сетке (`GridSearchCV` из библиотеки `sklearn.model_selection`) с тремя фолдами и метрикой качества - `roc_auc`. В ответ запишите наилучшее среди искомых значение `max_depth`.

Комментарий: остальные гиперпараметры дерева оставьте дефолтными (`splitter='best'`, `min_samples_split=2`, `min_samples_leaf=1`, `min_weight_fraction_leaf=0.0`, `max_features=None`, `random_state=None`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `class_weight=None`, `ccp_alpha=0.0`)

Ответ:



Правильный ответ: 2

б) (0.5 балла). Добавьте к данным новый признак `cat_bio`, содержащий в качестве значений пары значений из столбца `type` и столбца `group`. Например, если кошка имеет `type='wild'` и `group='group B'`, то в `cat_bio` будет стоять строка `'(wild, group B)'`. Примените `OneHotEncoding` (с учетом того, что мы не хотим получить мультиколлинеарности в новых данных) к столбцам `'cat_bio'`, `'education'`, `'meal'`, `'preparation course'`, а затем обучите решающее дерево глубины 5 с энтропийным критерием информативности на полученных после кодирования данных. Чему равен `roc_auc`? Ответ округлите до сотых.

Комментарий: остальные гиперпараметры дерева оставьте дефолтными (`splitter='best'`, `min_samples_split=2`, `min_samples_leaf=1`, `min_weight_fraction_leaf=0.0`, `max_features=None`, `random_state=None`, `max_leaf_nodes=None`, `min_impurity_decrease=0.0`, `min_impurity_split=None`, `class_weight=None`, `ccp_alpha=0.0`)

Ответ:



Правильный ответ: 0.68

На собеседование в некоторую компанию кандидаты на должность data scientist'a либо приходят пешком, либо приезжают на автомобиле. У нас имеется информация о 100 кандидатах. Для них мы также знаем, приняли кандидата на должность или нет. Имеющиеся у нас данные представлены в виде матрицы ниже:

	Кандидат принят	Кандидат не принят
Приехал на машине	20	28
Пришёл пешком	35	17

Используем логистическую регрессию без регуляризации для предсказания вероятности того, что кандидата возьмут на должность в зависимости от того, пришёл он пешком или приехал на машине. Какой прогноз вероятности того, что кандидат, пришедший пешком, будет принят на должность даёт логистическая модель? Ответ округлите до сотых.

Ответ: 0.67



Правильный ответ: 0.67

На плоскости даны следующие точки в двумерном пространстве:

$X = [(-1, 1), (1, -1), (1, 1), (0, 0)]$ с соответствующими метками классов $y = [1, 1, 1, -1]$.

С помощью leave-one-out кросс-валидации найдите оптимальное число соседей $k \in [1, 3]$ в методе ближайших соседей.

В качестве меры близости используется евклидово расстояние, метрика качества - ассигасу.

Ответ: 0.75



Правильный ответ: 3

Пусть каждый объект описывается двумерным вектором $x = (x_1, x_2)$.

Дан вектор $w = (2, 3)$ и число $w_0 = 7$.

Найдите ширину полосы между $\langle w, x \rangle = w_0 + 1$ и $\langle w, x \rangle = w_0 - 1$, где $\langle w, x \rangle$ - скалярное произведение вектора w и вектора x .

Ответ округлите до сотых.

Ответ:



Правильный ответ: 0.55

Компания по страхованию автомобилей разделяет водителей по трём классам: класс А (мало рискует), класс В (рискует средне), класс С (рискует сильно).

Компания предполагает, что из всех водителей, застрахованных у неё, 30% принадлежат классу А, 50% – классу В, 20% – классу С. Вероятность того, что в течение года водитель класса А попадёт хотя бы в одну автокатастрофу, равна 0,01; для водителя класса В эта вероятность равна 0,03, а для водителя класса С – 0,1. Мистер Джонс страхует свою машину у этой компании и в течение года попадает в автокатастрофу. Какова вероятность того, что он относится к классу А? Ответ округлите до сотых.

Ответ:



Правильный ответ: 0.08

Задание 1 (0.25 балла). Заполните пропуски в столбце уникальной категорией, если столбец с пропуском категориальный, и средним значением, если столбец числовой. Заполняйте одновременно и `df_train`, и `df_test` - одинаковым образом. В ответе укажите количество различных значений, потребовавшихся для заполнения пропусков (это равно количеству новых уникальных категорий плюс количество средних значений для заполнения пропусков в числовых столбцах).

Ответ: 32



Правильный ответ: 1

Задание 2 (0.3 баллов). Кошка прошла полосу препятствий по мнению судьи, если он поставил ей больше 50 баллов. Кошка считается прошедшей полосу препятствий, если все судьи поставили ей больше 50 баллов. В `df_train` создайте колонку 'Pass' и запишите в неё 1, если кошка прошла полосу препятствий, и 0 иначе. В ответ запишите, сколько кошек из `df_train` не прошли полосу препятствий. В `df_test` от вас скрыта информация о судейских баллах, поэтому неизвестно, прошла кошка полосу препятствия или нет - это и надо будет предсказать в заданиях ниже.

Ответ:

145



Правильный ответ: 145

Задание 3 (каждый пункт - 0.25 балла, 1.25 балла максимум).

Это задание выполняйте по данным `df_train`.

1) Среди всех диких кошек найдите долю кошек, прошедших полосу препятствий. Такую же долю рассчитайте для домашних кошек. В ответе укажите модуль разности этих долей. Ответ округлите до сотых.

Ответ:



Правильный ответ: 0.02

Выберите все корректные утверждения про градиентный спуск:

Выберите один или несколько ответов:

- ☐ На каждом шаге алгоритма считается градиент от одного, случайно выбранного элемента.
- ☒ Правильный подбор шага градиентного спуска позволяет уменьшить количество шагов, необходимых для поиска минимума. ✓
- ☒ Если сделать длину шага градиентного спуска недостаточно маленькой, то алгоритм может разойтись. ✓
- ☐ Градиентный спуск применяется для нахождения максимума функции потерь

Ваш ответ верный.

Правильные ответы: Если сделать длину шага градиентного спуска недостаточно маленькой, то алгоритм может разойтись., Правильный подбор шага градиентного спуска позволяет уменьшить количество шагов, необходимых для поиска минимума.

Какие из перечисленных ниже подходов могут помочь снизить переобучение в градиентном бустинге на решающих деревьях?

Выберите один или несколько ответов:

- ☒ ограничение сверху на количество листьев в дереве ✓
- ☒ ограничение сверху на абсолютную величину прогнозов в листьях дерева в задаче регрессии ✓
- ☒ ограничение сверху на количество деревьев в композиции ✓
- ☐ ограничение сверху на минимальное количество объектов в листе
- ☒ ограничение сверху на глубину дерева ✓

Ваш ответ верный.

Правильные ответы: ограничение сверху на глубину дерева, ограничение сверху на количество деревьев в композиции, ограничение сверху на количество листьев в дереве, ограничение сверху на абсолютную величину прогнозов в листьях дерева в задаче регрессии

Выберите все верные утверждения про алгоритм случайного леса:

Выберите один или несколько ответов:



- ☒ В случайном лесе при выборе наилучшего разбиения в вершине перебирается лишь случайное подмножество признаков ✓
- ☐ Случайный лес имеет меньшее смещение, чем решающее дерево той же глубины
- ☒ Случайный лес не переобучается с ростом числа деревьев ✓
- ☒ Классификация объектов проводится путём голосования деревьев внутри случайного леса. ✓
- ☐ В случайном лесе каждое дерево обучается на подвыборке обучающей выборки, сгенерированной таким образом, чтобы в ней не было повторяющихся объектов (bootstrap)

Ваш ответ верный.

Правильные ответы: Случайный лес не переобучается с ростом числа деревьев, В случайном лесе при выборе наилучшего разбиения в вершине перебирается лишь случайное подмножество признаков, Классификация объектов проводится путём голосования деревьев внутри случайного леса.

Выберите верные утверждения про K-means:

Выберите один или несколько ответов:

- ☐ Метод сам выбирает необходимое число кластеров.
- ☒ Найденная методом кластеризация зависит от выбора начального положения центров кластеров 
- ☒ Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. 
- ☐ Метод подходит для кластеров со сложной геометрией

Ваш ответ верный.

Правильные ответы: Найденная методом кластеризация зависит от выбора начального положения центров кластеров, Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния.

На скрытом слое нейронной сети используется функция активации σ . Выходное значение некоторого нейрона после применения функции активации получилось равным "-0.007". Какая из перечисленных функций активации σ могла быть использована в этой сети?

Выберите один ответ:

- ☐ Tanh
- ☒ ReLU ✖
- ☐ Sigmoid
- ☐ Никакая из перечисленных

Ваш ответ неправильный.

Правильный ответ: Tanh

Алгоритм бинарной классификации выдаёт значения b_i , принадлежащие отрезку $[0, 1]$. Всего имеется 10000 наблюдений. Если ранжировать их по возрастанию b_i , то окажется, что наблюдения с $y_i = 1$ занимают ровно места с 6501 по 6600. Найдите площадь под ROC-кривой. Ответ округлите до сотых.

Ответ:



Правильный ответ: 0.66

Дано сингулярное разложение матрицы X :

$$X = U \cdot \begin{pmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix} \cdot V'$$

Найдите сингулярное разложение $U_{10X} \cdot \Sigma_{10X} \cdot V'_{10X}$ для матрицы $10 \cdot X$. Чему равна сумма всех элементов матрицы Σ_{10X} ?

Выберите один ответ:

- ☐ 10
- ☐ 1000
- ☒ 100 ✓
- ☐ зависит от матрицы X

Ваш ответ верный.

Правильный ответ: 100

Вам нужно с помощью машинного обучения научиться предсказывать для каждой статьи, опубликованной на некотором сайте, число её просмотров. У вас есть следующие признаки: имя автора статьи, рейтинг автора статьи, число статей этого автора на сайте, длина статьи (количество символов) и несколько других характеристик статьи. Целевая переменная используется в алгоритме в исходном виде, без каких-либо изменений. Какую или какие из перечисленных ниже метрик можно использовать для оценки качества алгоритма в этой задаче?

Выберите один ответ:

- ☐ ROC-AUC
- ☐ F1-score
- ☒ MSE ✓
- ☐ не подходит ни одна из перечисленных
- ☐ Accuracy

Ваш ответ верный.

Правильный ответ: MSE

Рассмотрим линейную модель регрессии в задаче предсказания целевой переменной по двум

признакам: $a(x) = w_0 + w_1 x_1 + w_2 x_2$. Функция потерь имеет вид $Q(w) = \sum_{i=1}^l (y_i - a(x_i))^2$

где y_i – значение целевой переменной на i -ом объекте. После оценки качества алгоритма по кросс-валидации выяснилось, что модель переобучилась. Какие из нижеперечисленных подходов корректно описаны и их можно предпринять для уменьшения переобучения?

Выберите один или несколько ответов:

- ☐ Добавим к модели регуляризатор $w_0^2 + w_1^2 + w_2^2$, так как l2-регуляризация может уменьшить переобучение
- ☐ Добавим полиномиальных признаков второй степени, чтобы увеличить обобщающую способность модели
- ☐ Добавим к модели регуляризатор вида $[w_1 \neq 0] + [w_2 \neq 0]$, так как l0-регуляризация может уменьшить переобучение (здесь $[x] = 1$, если выражение x верно, а иначе 0)
- ☐ Уберём константный коэффициент w_0 , так как он увеличивает сложность модели и при этом не влияет на обобщающую способность модели
- ☐ Добавим к модели регуляризатор $|w_1| + |w_2|$, так как l1-регуляризация может уменьшить переобучение

Ваш ответ неправильный.

Правильные ответы: Добавим к модели регуляризатор $|w_1| + |w_2|$, так как l1-регуляризация может уменьшить переобучение

, Добавим к модели регуляризатор вида $[w_1 \neq 0] + [w_2 \neq 0]$, так как l0-регуляризация может уменьшить переобучение (здесь $[x] = 1$, если выражение x верно, а иначе 0)

Мы решаем задачу классификации для идентификации человека по голосу (1 — голос принадлежит пользователю, 0 — голос не принадлежит пользователю). Какую метрику качества следует выбрать, если мы хотим штрафовать только некорректное распознавание чужого голоса как голоса пользователя? (все метрики показывают качество работы алгоритма, т.е. чем больше значение метрики, тем выше качество алгоритма):

Выберите один ответ:

- ☐ $TP/(TP+FN)$
- ☐ $TP/(TP+FP)$
- ☐ $(TP+TN)/(TP+FP+TN+FN)$
- ☒ $TN/(FP+TN)$ ✓

Ваш ответ верный.

Правильный ответ: $TN/(FP+TN)$