

Теория

Вопрос	Ответ
Верно ли утверждение, что чем больше данных мы имеем тем лучше качество принимаемых решений/качество модели?	Неверно
Перечислите требования к новым знаниям.	Знания должны быть новые, Знания должны быть не тривиальны, Знания должны иметь практическую пользу, Знания должны быть доступны для понимания человеку
Методы машинного обучения и анализа данных имеют высокую важность т.к. в настоящее время:	наблюдается стремительный рост объема собираемых данных, происходит быстрое накопление данных о протекающих в мире процессах
Какие основные проблемы при работе с "сырыми" данными вам известны?	Пробелы в данных, Недостоверные значения, Неверные данные, Возможно наличие взаимозависимых/взаимосвязанных показателей
Если в данных присутствует дата рождения то замена её на вычисляемый показатель возраста может привести к повышению качества разрабатываемой модели?	Верно
Метрика использующая среднеквадратичное отклонение получается на основе	евклидова расстояния
Метод наименьших квадратов удастся применить когда:	существует обратная от матрицы A , есть гипотеза о виде функции для описания статистических данных, имеется множество статистических данных
Временной ряд у которого такие статистические характеристики, как его математическое ожидание (среднее), дисперсия (ср. кв. отклонение) и ковариация, не зависят от момента времени называют	стационарным в слабом смысле
Для проверки стационарности ряда необходимо	выполнить группу тестов и оценок и сделать общий вывод

Вопрос	Ответ
Использование базисных функций необходимо при использовании	метода наименьших квадратов метода наименьших квадратов для локальной регрессии
Задачей классификации называют	отнесение данных из заданного набора к одному из заранее определенных классов
Ассоциативные правила уазывают на	взаимосвязанные события/факторы
Рассмотрим в качестве примера вещественные числа и два класса — t_1 и t_2 . Пусть $t_1=\{1\}$, $t_2=\{10\}$. Будем использовать метод kNN со значением $k=1$. К какому классу будет отнесено значение 7?	t_2
Рассмотрим в качестве примера вещественные числа и два класса — t_1 и t_2 . Пусть $t_1=\{1\}$, $t_2=\{10\}$. Предположим теперь, что мы получили набор неразмеченных данных $\{2, 3, 4, 5, 6, 11, 12, 13, 14, 15, 16\}$ и использовали самообучение. Будем использовать метод kNN со значением $k=1$. К какому классу будет отнесено значение 7?	t_1
Для задания лингвистической переменной (или нечеткого числа) используют	функцию принадлежности
Если при проверке качества модели классификации посчитать процент правильно отгаданных классов то это	приведет к неверной оценке качества модели на несбалансированных данных
Кросс-валидация используется	повышения точности моделей на несбалансированных данных
Оценкой качества моделей классификации является	площадь под ROC кривой
Деревья решений хорошо работают когда	имеется большое количество параметров и зависимостей между ними
Расщепление это	условие деления/ветвления полученное на основе значений одного из анализируемых параметров
Дерево решений это	способ представления правил в иерархической, последовательной структуре
Работа метода логистической регрессии основана на использовании	функции – сигмоид

Вопрос	Ответ
Верно ли, что метод логистической регрессии может применяться в задачах регрессии?	Верно
Распределением Бернулли называют	распределение принимающее только два значения с вероятностями p и q
Градиентом называют	нормаль к заданной поверхности
Метод градиентного спуска имеет ограничения связанные с тем, что	может попадать в локальные экстремумы
Метод логистической регрессии	при использовании в задачах классификации применяется при только при бинарной классификации, может решить только задачу линейного разделения на классы.
Фоносемантика позволяет	определить ореол (например, цветовой окрас) слова/текста
Токенизация это	разделение текста на слова/подслова
Bag of words это	упрощенное представление текста, которое используется в обработке естественных языков и информационном поиске
При анализе текста классификацию используют для	определения класса текста, определения типа отношения
Оценка качества систем анализа текста основана на использовании	на сравнении множества полученных моделью ответов с множеством правильных ответов
Расставьте подходы/модели по мере их развития (по мере их появления)	KDD процесс, модель CRISP, модель ASUME
Для каких целей используются KDD процесс; модели SEMMA, CRISP, ASUM?	Для разработки информационных решений основанных на данных
В чем разница между подходами основанными на данных и моделях ИИ?	Хранением/не хранением статистических данных
В каких прикладных областях находят наиболее часто применение методы статистики и машинного обучения?	Обработка изображений, Обработка текстов, Поисковые системы, Обработка сигналов, Обработка звука

Вопрос	Ответ
Данные с которыми работают методы их обработки содержат:	наблюдения, измерения, события, транзакции, записи, структурированные и не структурированные данные, содержат атрибуты (колонки таблиц, длина, тип данных и т.д);
Если задана последовательность данных $X=\{100,101,98,103,95,101,100,102,2,102\}$, и Вы знаете, что среди данных есть ошибочное значение, то какое из приведённых значений Вы примете за неверное?	2
Из наблюдений известно, что изменение значения возраста Мисс Америка и числа смертельных случаев от пара и горячих предметов в США очень похожи. Если Вы разрабатываете модель для предсказания победителя нового конкурса Мисс Америка, то, как Вы думаете, для построения более эффективной модели следует использовать оба параметра?	Неверно
Аппроксимацией называют	процедуру замены одной функции на другую функцию
Необходимость в решении задачи аппроксимации возникает когда	при заданном значении аргумента значение функции может быть найдено только из эксперимента, функция задана таблично, а необходимые значения не совпадают с табличными, вычисление заданной функции связано с некоторыми сложностями (например, требует большого времени)
При построении интерполяционной функции с использованием многочлена Лагранжа по 10 точкам, многочлен какой степени мы получим?	9 (на один меньше чем точек)
Сплаины используются:	для снижения степени интерполяционной функции
Конечными разностями первого порядка называют -	разности между значениями функции в соседних узлах
Для использования второй формулы Ньютона значения узлов должны быть -	равноотстоящими
Задача обратного интерполирования заключается в том, чтобы	по значению функции (y) найти значение аргумента (x)
Основная идея метода PLS состоит в	введении дополнительных внутренних переменных

Вопрос	Ответ
??	с разрешимостью оптимизационной задачи точными методами, с требованиями наличия единственного решения.
Задачей множественной регрессии называют	оба варианта
Использование базисных функций необходимо при реализации	метода наименьших квадратов
При использовании многопараметрической модели прогнозирования на заданный горизонт планирования помимо прогнозирования значений целевого параметра также необходимо	прогнозировать значения других используемых параметров
Модель Лотки-Вольтеры используется для описания:	экономических процессов, процессов в биологических сообществах
При декомпозиции временных рядов могут выделяться составляющие:	уровня, тренда, сезонная, шумовая
Значения ковариации?	от -1 до 1
Коэффициент корреляции Пирсона	оценивает линейную связь переменных
??	дифференциальных уравнений и/или системы дифференциальных уравнений, передаточных функций
Не центральные распределения отличаются от центральных	скошенностью
Использование BoxPlot диаграмм позволяет качественным образом определить	совпадения средних значений и скошенность
Закон распределения является	характеристикой данных (измеряемые значения)
Количественные критерии оценки адекватности отличаются от качественных тем, что	величину оценки вероятности принятия/отклонения гипотезы можно использовать для оценки качества экстраполяционных моделей
Коэффициент Джинни отвечает на вопрос об	относительной концентрации значений
Вероятность Лапласа это значение равное	отношению числа выбранного повторяющегося значения в статистической выборке к общему числу значений в выборке
Энтропией называют	меру неопределенности или непредсказуемости некоторой системы

Вопрос	Ответ
Перечислите аксиомы Колмогорова	$P(A) \geq 0$ Если пересечение множеств A и B дает пустое множество то $P(\text{объединения } A \text{ и } B) = P(A) + P(B)$ вопрос $P(\text{множество возможных значений}) = 1$
Дисперсия для независимых событий может быть определена по формуле:	$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$
Теорема Байеса позволяет (Теорема Байера позволяет)	определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие
Временным рядом называется	расположенная в хронологическом порядке последовательность значений каких-либо параметров исследуемого процесса
SlopeOne позволяет учитывать	множество видов анализируемых действий
ограничения метода SVM/SVR связаны с	<ul style="list-style-type: none"> - разрешимостью оптимизационной задачи точными методами - необходимостью наличия единственного решения
принцип работы PageRank алгоритма ориентируется на	число ссылок на страницу
ассоциативные правила нашли применение в задачах	<ul style="list-style-type: none"> - раскладки товаров на полки магазина - выдачи списка рекомендуемых товаров в интернет-магазинах - наилучшего подбора ответов на поисковые запросы - разработки диагностических систем
ассоциативные правила	стараяются применить в первую очередь
Информационную систему можно считать системой искусственного интеллекта если она -	проходит тест Тьюринга

Вопрос	Ответ
Какие задачи из нижеприведенных решаются методами машинного обучения?	1) задача классификации 2) задача регрессии 3) задача кластеризации
Какие виды задачи классификации бывают?	1) бинарной классификации 2) множественной классификации
Какие подходы в разработке методов искусственного интеллекта и машинного обучения вам известны?	1) построение систем подобных живым 2) построение систем/методов имитирующих процессы протекающие при решении задач человеком
Когнитивная карта это -	образ знакомого пространства окружения.
Генераторы случайных чисел используют для:	1) генерации обучающих выборок; 2) заполнения пропущенных значений;
На точность прогноза методами экстраполяции влияют:	1) выбор параметров; 2) используемый метод; 3) горизонт прогнозирования;
Количественные критерии оценки адекватности отличаются от качественных тем, что	величину оценки вероятности принятия/отклонения гипотезы можно использовать для оценки качества экстраполяционных моделей.
PageRank алгоритм разрабатывался для	ранжирования интернет страниц для формирования выдачи в поисковых системах.
Какой метод для генерации равномерно распределённых случайных чисел разработал Дж. фон Нейман?	Метод среднего квадрата
К качественным критериям относят:	1) критерий Колмогорова 2) критерий Пирсона;
Наивный Байесовский классификатор опирается в своей работе на	теорему Байеса.

Вопрос	Ответ
Для решения задачи классификации с использованием теории графов необходимо	1) решить задачу о нахождении минимального разреза;, 2) присвоить ребрам значения показывающие степень сходства между вершинами;, 3) сопоставить узлам классифицируемые элементы.
К методам построения композиций моделей относятся:	1) бэггинг; 2) бустинг; 3) стекинг;
Начальные центры кластеров могут определяться:	1) случайным образом; 2) на основе данных полученных при работе другого алгоритма. 3) экспертным методом;
Алгоритмы кластеризации как правило применяются для решения следующих задач:	1) понимание данных — под пониманием в методах кластеризации подразумевают разбиение данных на кластеры, к которым могут применяться индивидуальные методы; 2) обнаружение новых объектов, которые не удается присоединить ни к одному из известных кластеров.
Идея бустинга состоит	в том, чтобы для неверно классифицированных данных использовать другую модель.
Кросс-валидация используется	повышения точности моделей на несбалансированных данных.
Самоорганизующаяся карта Кохонена это	нейронная сеть.
Если какое-то значение встречается намного чаще остальных, то это может говорить о	1) Несбалансированности данных 2) Поломке/не стабильной работе датчика снимающего данный параметр 3) О наиболее вероятном значении параметра

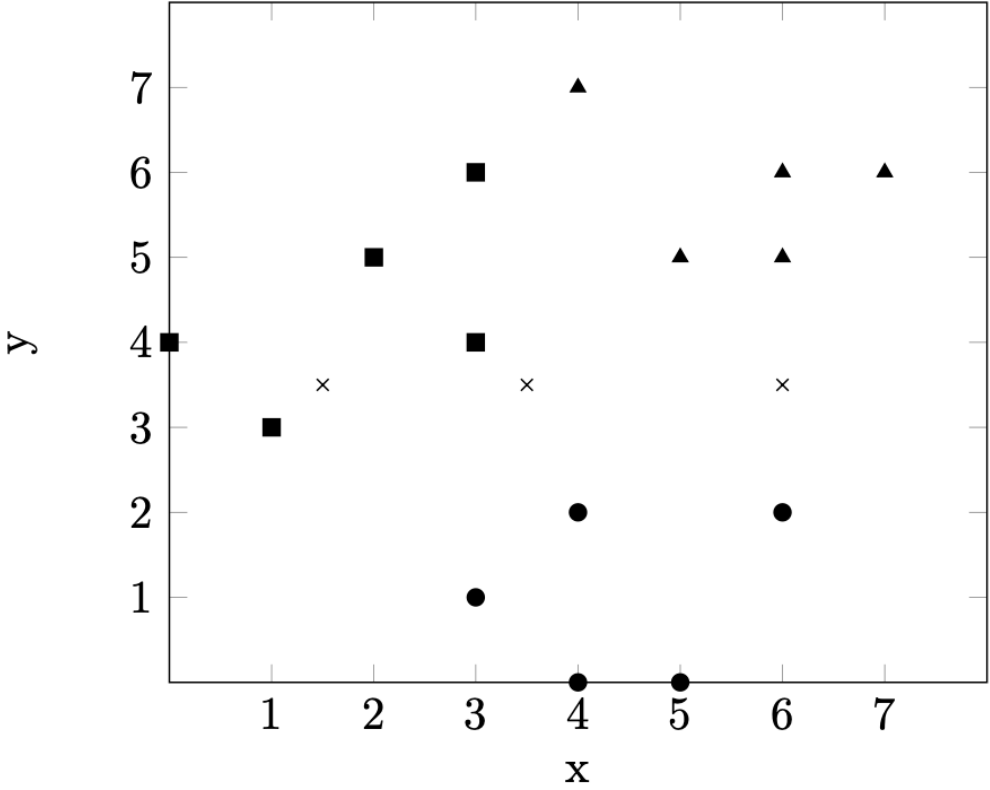
Вопрос	Ответ
На графике приведена зависимость между возрастом Мисс Америка и числом смертельных случаев в США от пара и горячих предметов...	Нет
Верно ли утверждение, что чем больше данных, тем лучшую по качеству модель мы получим?	Нет
Проклятье размерности это	невозможность обучения модели или получения ответа на обученной модели за конечное/доступное для принятия решения время
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?	MAE
Верно ли утверждение, что изменение набора показателей для обучения модели будет влиять на качество её работы?	Верно
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):	$TN/(FP+TN)$

Вопрос	Ответ						
<p>Найдите значения коэффициентов c_1 и c_2 для функции $y=c_1+c_2*x$, если известно, что функция экстраполирует следующие данные</p> <p>-----</p> <table><tr><td>y</td><td>1</td><td>10</td></tr></table> <p>-----</p> <table><tr><td>x</td><td>1</td><td>10</td></tr></table> <p>-----</p>	y	1	10	x	1	10	$c_1=0; c_2=1$
y	1	10					
x	1	10					
<p>При использовании интерполяционной формулы Лагранжа на данных</p> <p>-----</p> <table><tr><td>y</td><td>1</td><td>10</td></tr></table> <p>-----</p> <table><tr><td>x</td><td>1</td><td>10</td></tr></table> <p>-----</p> <p>первое слагаемое многочлена примет вид</p>	y	1	10	x	1	10	$(10-x)/9$
y	1	10					
x	1	10					

Вопрос	Ответ
<p>К алгоритмам кластеризации предъявляют следующие требования:</p>	<p>алгоритм должен работать в режиме, позволяющем иметь доступ к данным другим задачам и приложениям (часто только в режиме чтения).</p> <p>работа в условиях ограниченной памяти компьютера;</p> <p>возможность прервать работу алгоритма с сохранением промежуточных результатов;</p> <p>минимально возможное количество проходов по базе данных;</p>
<p>Рассмотрим регулярные деревья решений глубиной 2 над n булевыми переменными! Регулярные деревья решений глубины 2 - это дерево решений глубины 2 (дерево с четырьмя листьями на расстоянии 2 от корня), в котором левый и правый дочерний элемент корня должны содержать одну и ту же переменную. Пример приведен на рисунке. Каков размер пространства гипотез в зависимости от числа переменных n?</p>	<p>4</p>

Вопрос	Ответ
<p>Пожалуйста расставьте соответствия.</p>	<p>Поиск функции отображения при отсутствии заранее заданных классов в размечаемых данных при наличии для обучения как размеченных данных так и не размеченных данных называется - полуконтролируемая кластеризация.</p> <p>Поиск функции отображения при наличии заданных классов и размеченных данных называется - классификация.</p> <p>Поиск функции отображения при отсутствии заранее заданных классов и размеченных данных называется - кластеризация.</p> <p>Поиск функции отображения при наличии заданных классов и допускающей наличие других классов в размечаемых данных при наличии для обучения как размеченных данных так и не размеченных данных называется - полуконтролируемая классификация.</p>
<p>К количественным критериям согласия относят:</p>	<p>критерий Стьюдента;</p> <p>критерий Фишера;</p>

Вопрос	Ответ
<p>При использовании для поиска значений коэффициентов модели методом Lasso на данных</p> <p>-----</p> <p>$y \ 1 \ 10 \$</p> <p>-----</p> <p>$x \ 1 \ 10 \$</p> <p>-----</p> <p>обнулятся</p>	<p>значение коэффициента c_1</p>

Вопрос	Ответ
<p>Рассмотрим заданные размеченные и не размеченные данные (см. рис.), а также соответствующие метки $C = \{\text{треугольник, квадрат, круг}\}$! Далее предположим, что вы используете классификатор 3-NN (в случае равенства используются все соседи с одинаковым расстоянием) в условиях активного обучения. Определите классы всех незамеченных данных.</p> 	<p>квадрат, квадрат, треугольник</p>

Вопрос	Ответ
Для реализации нейронных сетей с памятью необходимо	наличие обратных связей
Какие автоматы существуют в рамках теории автоматов	автомат Мура абстрактный автомат автомат Миля автомат Кринского
В рамках какой теории может быть описана машина Тьюринга	теории автоматов
Выберите этапы построения нейронных сетей	выбрать метод обучения корректирующих связей выбрать правило распространения сигналов в сети определить множество входных и выходных связей выбрать правило вычисления сигнала активности определить множество нейронов выбрать правило комбинирования входных сигналов

Вопрос	Ответ
Какие нейронные сети вы знаете (Какие нейронные сети вы знаете)	LSTM GRU FNN Сеть Элмана RNN
Какие из приведенных ниже методов обучения нейронных сетей вам известны?	Метод Хебба Метод градиентного спуска Метод обратного распространения
Фоносемантика позволяет	определить ореол (например, цветовой окрас) слова/текста
Токенизация это	разделение текста на слова/подслова
Bag of words это	упрощенное представление текста, которое используется в обработке естественных языков и информационном поиске
При анализе текста классификацию используют для	определения класса текста определения типа отношения
Оценка качества систем анализа текста основана на использовании	на сравнении множества полученных моделью ответов с множеством правильных ответов

Вопрос	Ответ
Какие методы распознавания изображений существуют в машинном обучении?	Метод Далала-Тригса Метод Виолы-Джонса
Какие уровни аннотации применяются при решении задач распознавания изображений? (Какие уровни аннотации применяются при решении задач распознавания изображений?)	Изображения "Окна" Пикселы
Какие задачи распознавания изображений рассматриваются?	Оценка конфигурации Контроль выполнения операций Нахождение объектов
Градиент - это	вектор, своим направлением указывающий направление наискорейшего роста некоторой скалярной величины (значение которой меняется от одной точки пространства к другой, образуя скалярное поле)
Визуальное слово это -	элемент изображения используемый как один из идентификаторов распознаваемого объекта
Равновесие Нэша это	набор стратегий в игре для двух и более игроков, в котором ни один участник не может увеличить выигрыш.
Интеллектуальный агент это	программа, самостоятельно выполняющая задание сущность, получающая информацию через систему сенсоров о состоянии управляемых ими процессов и осуществляющие влияние на них через систему актуаторов

Вопрос	Ответ
Предикат это	это высказывание, в которое можно подставлять аргументы
Какие типы экспертных систем существуют	продукционные Нейлоровская диагностирующая система фреймовые
Для построения экспертных систем применяются	фреймы матрицы предикаты

Формулы

Название	Обозначение	Формула	Пример
Среднее значение	$M(x), \bar{x}$	$\frac{x_1 + \dots + x_n}{n}$ (сумма значений поделить на число значений)	$\{1, 5, 3, 2\}$ $(1+5+3+2)/4=2.75$
Медиана		Значение посередине отсортированной выборки (нечётное число значений), или среднее значений значений рядом с серединой отсортированной выборке (чётное число значений)	$\{2, 5, 4, 2, 4\} \rightarrow \{2, 2, 4, 4, 5\} \rightarrow 4$ $\{2, 5, 4, 2\} \rightarrow \{2, 2, 4, 5\} \rightarrow (2+4)/2 = 3$
Мода		Среднее значение наиболее часто встречающихся в выборке значений	$\{2, 2, 4, 5, 2\} \rightarrow 2$ $\{2, 4, 5, 2, 4\} \rightarrow (2+4)/2 = 3$

Название	Обозначение	Формула	Пример
Дисперсия	$D(x)$	$\frac{(x_1 - M(x))^2 + \dots + (x_n - M(x))^2}{n}$ (сумма квадратов разностей значений выборки и среднего значения выборки поделить на число значений) ИЛИ $M(x^2) - M(x)^2$ (разность среднего значения выборки, где все значения возведены в квадрат, и квадрата среднего ожидания выборки)	{1, 5, 3, 2} $M(x) = 2.75$ $M(x^2) = 9.75$ $((1-2.75)^2 + (5-2.75)^2 + (3-2.75)^2 + (2-2.75)^2)/4 = 2.1875$ $9.75 - 2.75^2 = 2.1875$
Среднеквадратичное отклонение	σ	$\sqrt{D(x)}$	{1, 5, 3, 2} $\text{sqrt}(2.1875) = 1.47901995$
Частота значения	p_i	$p_i = \frac{n_i}{n}$, где n_i - сколько раз значение встречается в выборке, n - размер выборки	{1, 2, 2, 4, 2, 5, 1} Для 2: $3/7 = 0.428571429$
Количество интервалов для интервального ряда (Формула Стёрджиса)	k	$1 + \lceil 3.32 * \log_{10}(n) \rceil$, где $[a]$ - целая часть a, n - размер выборки.	{1, 2, 3, 2, 3, 2, 1} $1 + \lceil 3.32 * \log_{10}(7) \rceil = 1 + \lceil 2.80572549 \rceil = 3$
Ширина интервала для интервального ряда	h	$\frac{\max(x_i) - \min(x_i)}{k}$, где k - количество интервалов	{1, 2, 3, 2, 3, 2, 1} k=3 $h = (3-1)/3 = 1.33333333$

Название	Обозначение	Формула	Пример
Семинар 3 (многочлен лагранжа, многочлен Ньютона)		https://edu.hse.ru/tokenpluginfile.php/2a4513f1d8b9c9d6b70eae9a794839f3/2311707/mod_resource/content/0/%D0%A1%D0%B5%D0%BC%D0%B8%D0%BD%D0%B0%D1%80%203.pdf Формула n-ого слагаемого для k точек: $y(x) = y_n \frac{(x-x_1) \cdot \dots \cdot (x-x_{n-1})(x-x_{n+1}) \cdot \dots \cdot (x-x_k)}{(x_n-x_1) \cdot \dots \cdot (x_n-x_{n-1})(x_n-x_{n+1}) \cdot \dots \cdot (x_n-x_k)}$	$y = \{1, 10\}$ $x = \{1, 10\}$ Для 1 слагаемого: $y_1 \frac{(x-x_2)}{(x_1-x_2)} = 1 \cdot \frac{(x-10)}{(1-10)} = \frac{10-x}{9}$
Матрица коэффициентов уравнения по МНК (метод наименьших квадратов)	c	$((A^T * A)^{-1} * A^T) * y$, где A - матрица значений мономов, y - матрица значений функции, A^T - транспонирование матрицы, A^{-1} - обратная матрица. https://edu.hse.ru/tokenpluginfile.php/2a4513f1d8b9c9d6b70eae9a794839f3/2320674/mod_resource/content/0/%D0%A1%D0%B5%D0%BC%D0%B8%D0%BD%D0%B0%D1%80%204.pdf	Сложно кратко записать. Да и считать долго. Надеюсь, не будет. Если будет - скипай задачу, и вернись с каким-нибудь калькулятором, если будет время.
Динамические регрессионные модели		https://edu.hse.ru/tokenpluginfile.php/2a4513f1d8b9c9d6b70eae9a794839f3/2352617/mod_resource/content/0/%D0%A1%D0%B5%D0%BC%D0%B8%D0%BD%D0%B0%D1%80%206.pdf	Скипай задачу, не трать на неё время.
Среднее геометрическое	$x_{геом}$	$\sqrt[n]{x_1 * \dots * x_n}$, n - размер выборки	$\{1, 2, 3, 2, 3, 2, 1\}$ $\sqrt[7]{1 * 2 * 3 * 2 * 3 * 2 * 1} = 1.84218488139$

Название	Обозначение	Формула	Пример
Среднее гармоническое	x_{har}	$\frac{1}{\frac{1}{n} \left(\frac{1}{x_1} + \dots + \frac{1}{x_n} \right)}$ - не существует, если в выборке есть 0	$\{1, 2, 3, 2, 3, 2, 1\}$ $1 / ((1 + 1/2 + 1/3 + 1/2 + 1/3 + 1/2 + 1) / 7) = 1.68$
Кривая Лоренца, коэффициент Джинни, индекс Херфиндаля		https://edu.hse.ru/tokenpluginfile.php/2a4513f1d8b9c9d6b70eae9a794839f3/2360238/mod_resoucontentrce//0/%D0%A1%D0%B5%D0%BC%D0%B8%D0%BD%D0%B0%D1%80%207.pdf	Лучше посмотри в семинар, но это тоже не быстрая задача.
Mean absolute error Средняя абсолютная ошибка	MAE	$\frac{1}{n} (y_1 - y_1^* + \dots + y_n - y_n^*)$; y_i - значения выборки, y_i^* - соответствующее значение модели	Данные - $\{1, 3, 2\}$ Модель - $\{3, 3, 1\}$ $(1-3 + 3-3 + 2-1) / 3 = 1$
Root mean square error Среднеквадратичная ошибка	RMSE	$\sqrt{\frac{1}{n} ((y_1 - y_1^*)^2 + \dots + (y_n - y_n^*)^2)}$; y_i - значения выборки, y_i^* - соответствующее значение модели	Данные - $\{1, 3, 2\}$ Модель - $\{3, 3, 1\}$ $\text{sqrt}((1-3)^2 + (3-3)^2 + (2-1)^2) / 3 = 1.29099445$

Название	Обозначение	Формула	Пример
Mean percentage absolute error Средняя абсолютная ошибка в процентах Mean absolute scaled error Средняя абсолютная масштабированная оценка	MAPE MASE	https://edu.hse.ru/tokenpluginfile.php/2a4513f1d8b9c9d6b70eae9a794839f3/2360238/mod_resource/content/0/%D0%A1%D0%B5%D0%BC%D0%B8%D0%BD%D0%B0%D1%80%207.pdf $MAPE = \frac{1}{n} \sum_{j=1}^n \frac{ y[i] - y^*[i] }{y[i]} * 100\%$ $MASE = \frac{MAE}{\frac{1}{n-1} \sum_{j=2}^n y[j] - y[j-1] }$	Смотри в семинаре
Энтропия	H(x), H[x]	$- (p(x_1) * \log_{base}(p(x_1)) + \dots + p(x_n) * \log_{base}(p(x_n)))$; где $p(x_i)$ - вероятность значения x_i , base - база логарифма, может быть 2, e, 10.	
Accuracy		$\frac{TP+TN}{TP+TN+FP+FN}$; где TP - true positives, TN - true negatives, FP - false positives, FN - false negatives	
Precision		$\frac{TP}{TP+FP}$	
Recall, True positive rate	TPR	$\frac{TP}{TP+FN}$	
Specificity		$\frac{TN}{TN+FP}$	

Название	Обозначение	Формула	Пример
F-мера		$(1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$	
False positive rate	FPR	$1 - Specificity$	

Пробный тест африканских слонов

Smart LMS

smartedu.hse.ru/mod/quiz/0/814557#moodle=/mod/quiz/attempt.php?attempt=2395641&cmid=814557

Вышка Digital | Smart LMS

Маркетплейс курсов

RU

← Все модули

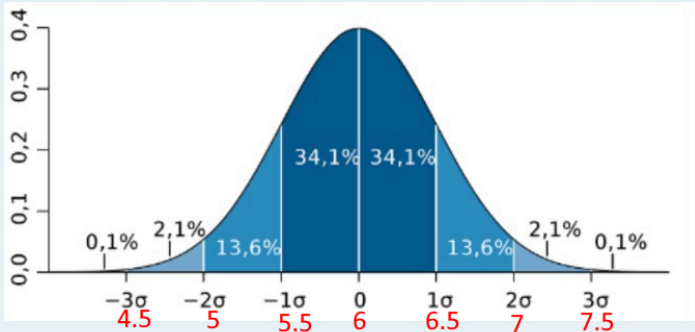
Вопрос 1

Пока нет ответа

Балл: 1,00

Отметить вопрос

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?



Interval (Tons)	Percentage
$-\infty$ to -3σ (4.5)	0.1%
-3σ to -2σ (5)	2.1%
-2σ to -1σ (5.5)	13.6%
-1σ to 0 (6)	34.1%
0 to 1σ (6.5)	34.1%
1σ to 2σ (7)	13.6%
2σ to 3σ (7.5)	2.1%
3σ to $+\infty$	0.1%

- ☒ a. 47.7 % африканских слонов весит от 6 до 7 тонн $34.1+13.6=47.7$
- ☐ b. 68.2 % африканских слонов находится между 5.5 и 7 тонн $34.1+34.1+13.6=81.8$
- ☒ c. 0.1 % африканских слонов весит больше 7.5 тонн $0.1=0.1$
- ☐ d. 13.6 % африканских слонов весит от 5 до 6 тонн $13.6+34.1=47.7$

Следующая страница

Навигация по тесту

1

2

3

4

5

6

7

8

9

10

Закончить попытку...

Оставшееся время 0:15:07

← Все модули

>



Вопрос 2

Ответ сохранен

Балл: 1,00

🚩 Отметить вопрос

Рассмотрим в качестве примера вещественные числа и два класса — $t1$ и $t2$. Пусть $t1=\{1\}$, $t2=\{10\}$. Будем использовать метод kNN со значением $k=1$. К какому классу будет отнесено значение 7?

- ☐ a. $t1$
☒ b. $t2$

Очистить мой выбор

$$|1-7|=6 > |10-7|=3$$

$t1$ $t2$

Предыдущая страница

Следующая страница

Навигация по тесту



Закончить попытку...

Оставшееся время 0:14:54

← Все модули

Вопрос 3

Пока нет ответа

Балл: 1,00

🚩 Отметить вопрос

При использовании интерполяционной формулы Лагранжа на данных

	0	1
+	-----	+
y	1 10	

x	1 10	
+	-----	+

первое слагаемое многочлена примет вид

- ☐ a. $10 \cdot (1-x)/9$
- ☐ b. $10 \cdot (x-1)/9$
- ☐ c. $(x-10)/9$
- ☒ d. $(10-x)/9$

Очистить мой выбор

$$y_0 \frac{x-x_1}{x_0-x_1} = 1 \cdot \frac{x-10}{1-10} = \frac{10-x}{9}$$

Предыдущая страница

Следующая страница

Навигация по тесту



Закончить попытку...

Оставшееся время 0:12:22

← Все модули

Вопрос 4

Ответ сохранен

Балл: 1,00

Отметить вопрос

Средне квадратичное отклонение (ошибка) модели $y=x$ на данных

y 1 10

x 1 10

x 1 10

$$\sqrt{\frac{1}{2}(0+0)} = 0$$

будет равно

☐ a. 0,1

☐ b. 10

☐ c. 1

☒ d. 0

Очистить мой выбор

Предыдущая страница

Следующая страница

Навигация по тесту

1 2 3 4 5 6 7 8 9 10

Закончить попытку...

Оставшееся время 0:11:28

← Все модули

Вопрос 5

Ответ сохранен

Балл: 1,00

🚩 Отметить вопрос

Если проводить проверку на адекватность гипотезы $y = -2 + 4x - x^2$ полученной на данных визуальными методами (по графикам и BoxPlot диаграммам) то какие из утверждений будут верными?

Статистические данные

```
+-----+
| y | 1 | 2 | 1 |
+-----+
| x | 1 | 2 | 3 |
+-----+
```

y' 1 2 1

- ☐ a. Интервалы роста и спада значений не совпадают.
- ☒ b. Средние значения совпадают.
- ☒ c. Значения дисперсии совпадают.
- ☐ d. Функции распределения смещены относительно друг друга.

Распределения совпадают, поэтому у них всё совпадает

Предыдущая страница

Следующая страница

Навигация по тесту

1 2 3 4 5 6 7 8 9 10

Закончить попытку...

Оставшееся время 0:11:20

← Все модули

Вопрос 6
Ответ сохранен
Балл: 1,00
🚩 Отметить вопрос

Задачей классификации называют

- ☐ a. выделение из заданного набора данных групп сходных объектов (заранее не известных классов).
- ☒ b. отнесение данных из заданного набора к одному из заранее определенных классов.

Очистить мой выбор

Предыдущая страница

Следующая страница

Навигация по тесту

1 2 3 4 5 6 7 8 9 10

Закончить попытку...

Оставшееся время 0:17:11

← Все модули

Вопрос 7

Пока нет ответа

Балл: 1,00

🚩 Отметить вопрос

Вероятность Лапласа это значение равное

- ☒ a. отношению числа выбранного повторяющегося значения в статистической выборке к общему числу значений в выборке.
- ☐ b. числу повторов выбранного значения в статистической выборке.
- ☐ c. значению выбранного элемента выборки.

Очистить мой выбор

Предыдущая страница

Следующая страница

Навигация по тесту

1 2 3 4 5 6 7 8 9 10

Закончить попытку...

Оставшееся время 0:16:53

← Все модули

Вопрос 8

Пока нет ответа

Балл: 1,00

🚩 Отметить вопрос

Фоносемантика позволяет

- ☐ a. соответствие слова/текста смысловой задумке автора.
- ☐ b. мелодичность звучания слова/текста;
- ☒ c. определить ореол (например, цветовой окрас) слова/текста;

Очистить мой выбор

Предыдущая страница

Следующая страница

Навигация по тесту

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Закончить попытку...

Оставшееся время 0:16:39

← Все модули

Вопрос 9

Пока нет ответа

Балл: 1,00

🚩 Отметить вопрос

Кросс-валидация используется

- ☐ a. для оптимального определения размера обучающей и тестовых выборок.
- ☒ b. повышения точности моделей на несбалансированных данных.
- ☐ c. для упорядочивания данных.

Очистить мой выбор

Предыдущая страница

Следующая страница

Навигация по тесту

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Закончить попытку...

Оставшееся время 0:16:28

← Все модули

Вопрос **10**
Ответ сохранен
Балл: 1,00
🚩 Отметить вопрос

Перечислите аксиомы Колмогорова

- ☐ a. $P(\text{пустое множество})=0$
- ☒ b. $P(\text{множество возможных значений})=1$
- ☒ c. Если пересечение множеств A и B дает пустое множество то $P(\text{объединения A и B})=P(A)+P(B)$
- ☒ d. $P(A) \geq 0$

Предыдущая страница

Закончить попытку...

Навигация по тесту

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

Закончить попытку...

Оставшееся время **0:15:52**

Вопрос 1
Верно
Баллов: 1,00 из 1,00
[Отметить вопрос](#)

Коэффициент корреляции Кендалла может принимать значения

- ☐ a. >0
- ☒ b. от -1 до 1 ✓
- ☐ c. не менее -1
- ☐ d. не более 1
- ☐ e. <0

Вопрос 2
Верно
Баллов: 1,00 из 1,00
[Отметить вопрос](#)

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?

- ☐ a. Accuracy
- ☐ b. не подходит ни одна из перечисленных
- ☒ c. MAE ✓
- ☐ d. ROC-AUC
- ☐ e. F1

Вопрос 3
Верно
Баллов: 1,00 из 1,00
[Отметить вопрос](#)

Существуют следующие способы поиска границ кластеров:

- ☒ a. на основе плотности распределения вероятности исследуемых значений. ✓
- ☒ b. на основе метрики показывающей удаление от центра кластера; ✓
- ☐ c. на основе предварительно заданной геометрической формы областей;
- ☒ d. на основе заранее заданных правил; ✓

Вопрос 4

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Найдите значения коэффициентов c_1 и c_2 для функции $y=c_1+c_2*x$, если известно, что функция экстраполирует следующие данные

| y | 1 | 10 |

| x | 1 | 10 |

- ☐ a. $c_1=-1; c_2=0$
- ☐ b. $c_1=1; c_2=0$
- ☒ c. $c_1=0; c_2=1$ ✓
- ☐ d. $c_1=0; c_2=-1$

Вопрос 5

Неверно

Баллов: 0,00 из 1,00

[Отметить вопрос](#)

Рассмотрим в качестве примера вещественные числа и два класса — t_1 и t_2 . Пусть $t_1=\{1\}$, $t_2=\{10\}$. Будем использовать метод kNN со значением $k=1$.

К какому классу будет отнесено значение 7?

- ☐ a. t_2 ;
- ☒ b. t_1 ; ✗
- ☐ c. ни к одному из перечисленных.

Вопрос 6

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Теорема Байеса позволяет

- ☒ a. определить вероятность какого-либо события при условии, что произошло другое статистически взаимосвязанное с ним событие. ✓
- ☐ b. позволяет определить вероятность отдельного независимого от других события.
- ☐ c. позволяет определить к какому классу относится событие.

Вопрос 7

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Верно ли утверждение, что при использовании теории графов при полуконтролируемом обучении введение искусственных узлов с "бесконечными" связями с другими узлами может упростить решение?

- ☒ Верно ✓
- ☐ Неверно

Вопрос 8

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Для каких целей используются KDD процесс; модели SEMMA, CRISP, ASUM?

- ☐ a. Для проверки моделей поддержки принятия решений
- ☒ b. Для разработки информационных решений основанных на данных ✓
- ☐ c. Для автоматизированной обработки статистических данных

Вопрос 9

Верно

Баллов: 1,00 из 1,00

[🚩 Ответить](#)
вопрос

Ограничения метода SVM/SVR связаны:

- ☒ a. с требованиями наличия единственного решения. ✓
- ☒ b. с разрешимостью оптимизационной задачи точными методами; ✓
- ☐ c. наличием вычислительного ресурса;
- ☐ d. количеством используемых статистических данных;

Вопрос 10

Верно

Баллов: 1,00 из 1,00

[🚩 Ответить](#)
вопрос

На точность прогноза методами экстраполяции влияют:

- ☒ a. выбор параметров; ✓
- ☒ b. используемый метод; ✓
- ☒ c. горизонт прогнозирования; ✓
- ☐ d. вычислительная мощность используемого вычислительного устройства.

Вопрос 11

Верно

Баллов: 1,00 из 1,00

[🚩 Ответить](#)
вопрос

Коэффициент корреляции Пирсона

- ☒ a. оценивает линейную связь переменных. ✓
- ☐ b. оценивает нелинейную связь переменных.
- ☐ c. оценивает частоту выбросов.

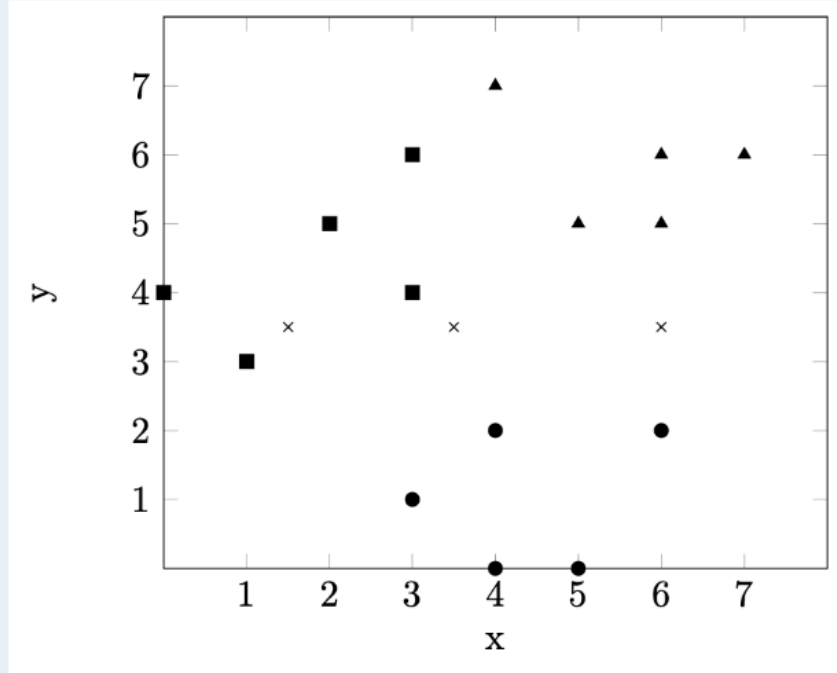
Вопрос 12

Верно

Баллов: 1,00 из 1,00

Отметить вопрос

Рассмотрим заданные размеченные и не размеченные данные (см. рис.), а также соответствующие метки $S = \{\text{треугольник, квадрат, круг}\}$. Далее предположим, что вы используете классификатор 3-NN (в случае равенства используются все соседи с одинаковым расстоянием) в условиях активного обучения. Определите классы всех незамеченных данных.



- ☐ a. квадрат, квадрат, круг
- ☐ b. квадрат, круг, треугольник
- ☐ c. квадрат, круг, круг
- ☒ d. квадрат, квадрат, треугольник ✓

Вопрос 13

Верно

Баллов: 1,00 из 1,00

Отметить вопрос

Выберите верные утверждения про k-means:

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

Вопрос 14

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

К алгоритмам кластеризации предъявляют следующие требования:

- ☒ a. возможность прервать работу алгоритма с сохранением промежуточных результатов; ✓
- ☐ b. возможность ручного ввода количества определяемых кластеров;
- ☒ c. работа в условиях ограниченной памяти компьютера; ✓
- ☒ d. алгоритм должен работать в режиме, позволяющем иметь доступ к данным другим задачам и приложениям (часто только в режиме чтения). ✓
- ☐ e. минимально возможное количество строк кода реализующего алгоритм;
- ☒ f. минимально возможное количество проходов по базе данных; ✓

Вопрос 15

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Чему будет равна составляющая уровня для набора данных

| y | 1 | 10 |

| x | 1 | 10 |

на отрезке $x=[1,10]$ при использовании STL разложения если другие составляющие не выделяются.

- ☐ a. 10
- ☐ b. 0
- ☒ c. 5,5 ✓
- ☐ d. 1

Вопрос 16

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

Алгоритмы кластеризации как правило применяются для решения следующих задач:

- ☐ a. проверки качества разметки данных для задач классификации;
- ☐ b. поиска не достоверных данных (выбросов);
- ☐ c. классификации;
- ☒ d. понимание данных — под пониманием в методах кластеризации подразумевают разбиение данных на кластеры, к которым могут применяться индивидуальные методы; ✓
- ☒ e. обнаружение новых объектов, которые не удается присоединить ни к одному из известных кластеров. ✓

Вопрос 17

Верно

Баллов: 1,00 из 1,00

[Отметить вопрос](#)

К количественным критериям согласия относят:

- ☐ a. критерий Пирсона;
- ☐ b. критерий Колмогорова.
- ☒ c. критерий Фишера; ✓
- ☒ d. критерий Стьюдента; ✓

Вопрос 18

Верно

Баллов: 1,00 из 1,00

[Отметить
вопрос](#)

Рассмотрим в качестве примера вещественные числа и два класса — t_1 и t_2 . Пусть $t_1=\{1\}$, $t_2=\{10\}$. Предположим теперь, что мы получили набор неразмеченных данных $\{2, 3, 4, 5, 6, 11, 12, 13, 14, 15, 16\}$ и использовали самообучение. Будем использовать метод kNN со значением $k=1$. К какому классу будет отнесено значение 7?

- ☐ a. t_2 ;
- ☐ b. ни к одному из перечисленных.
- ☒ c. t_1 ; ✓

Вопрос 19

Верно

Баллов: 1,00 из 1,00

[Отметить
вопрос](#)

Метод динамической трансформации временной шкалы использует

- ☐ a. косинусную меру.
- ☐ b. расстояние Лана.
- ☐ c. манхэттонское расстояние.
- ☒ d. евклидово расстояние. ✓

Вопрос 20

Верно

Баллов: 1,00 из 1,00

[Отметить
вопрос](#)

При использовании для поиска значений коэффициентов модели методом Lasso на данных

.....
 $|y| \ 1 \ 10 \ |$
.....

$|x| \ 1 \ 10 \ |$
.....

обнулятся

- ☐ a. значение коэффициента c_2 .
- ☐ b. значения обоих коэффициентов.
- ☐ c. не одно из значений коэффициентов не обнулится.
- ☒ d. значение коэффициента c_1 . ✓