Глоссарий по курсу «Машинное обучение и большие данные»

Accuracy (точность) — сочетание или сумма всех правильно предсказанных единичек и ноликов на, собственно, общее число всех возможных объектов.

BERT (*Bidirectional Encoder Representations from Transformers*) — это нейронная сеть от *Google*, показавшая с большим отрывом state-of-the-art результаты на целом ряде задач. С помощью BERT можно создавать программы с ИИ для обработки естественного языка: отвечать на вопросы, заданные в произвольной форме, создавать чат-ботов, автоматические переводчики, анализировать текст и так далее. Источник: *https://habr.com*

Baseline (исходные условия) — это согласованное описание атрибутов продукта в определенный момент времени, которое служит основой для определения изменений.

CatBoost — открытая программная библиотека, разработанная компанией Яндекс. Реализует уникальный патентованный алгоритм построения моделей машинного обучения, использующий одну из оригинальных схем градиентного бустинга.

Confusion matrix — матрица с распределением объектов по истинным классам и предсказанным классам.

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

DBSCAN (*Density-Based Spatial Clustering*) — плотностный алгоритм кластеризации. Позволяет находить кластеры произвольной формы в метрическом пространстве.

DBSCAN — это алгоритм кластеризации, основанной на плотности. Если дан набор точек в некотором пространстве, алгоритм группирует вместе точки, которые тесно расположены (точки со многими близкими соседями]), помечая как выбросы точки, которые находятся одиноко в областях с малой плотностью (ближайшие соседи которых лежат далеко).

DeepPavlov — это общедоступная программная библиотека, которая содержит набор компонентов для быстрого прототипирования диалоговых систем.

F-мера — производная метрика от *Precision* и *Recall*, которая уравновешивает точность предсказания первого класса и его полноту.

FOREL (формальный элемент) — алгоритм кластеризации, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения.

fuzzy k-means — метод нечёткой кластеризации средних (англ. *fuzzy clustering*, *soft k-means*, *c-means*). Позволяет разбить имеющееся множество элементов мощностью N на заданное число нечётких множеств k.

Hadoop — проект фонда *Apache Software Foundation*, свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов.

k-means — наиболее популярный метод кластеризации. Стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

LSTM сети (долгая краткосрочная память) — разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям.

MAE — среднее абсолютное отклонение. $\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | y_i - y_i |$.

MAPE — средняя абсолютная процентная ошибка. $\mathrm{MAPE} = 100\ \% \cdot \frac{1}{n}\sum_{i=1}^{n} \frac{| y_i - y_i |}{| y_i |}$.

MSE — среднеквадратическая ошибка прогноза. $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - y_i)^2$.

NLTK (*natural language toolkit*) — зарубежная открытая библиотека, созданная в Python. Позволяет решать задачи обработки естественного языка (токенизация, лемматизация, стемминг и др.)

N-граммы — это альтернатива морфологическому разбору и удалению стоп-слов. *N*-грамма — это часть строки, состоящая из N символов. Например, слово «дата» может быть представлено 3-граммой «_да», «дат», «ата», «та_» или 4-граммой «_дат», «дата», «ата_», где символ подчеркивания заменяет предшествующий или замыкающий слово пробел.

Precision (точность) — доля правильно найденных объектов класса.

Pymorphy — морфологический анализатор, разработанный на языке программирования Python. Выполняет лемматизацию и анализ слов, способен осуществлять склонение по заданным грамматическим характеристикам слов.

Recall (полнота) — доля найденных объектов класса.

re.findall — этот метод возвращает список всех найденных совпадений. Если мы будем искать «москв» в нашей строке, он вернет все вхождения «москв».

TF-IDF — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

U-критерий Манна–Уитни — непараметрический статистический критерий, используемый для сравнения двух независимых выборок по уровню какого-либо признака, измеренного количественно.

Word2vec — способ построения сжатого пространства векторов слов, использующий нейронные сети. Принимает на вход большой текстовый корпус и сопоставляет каждому слову вектор. Сначала он создает словарь, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожий смысл), в векторном представлении имеют высокое косинусное сходство.

Программное обеспечение под названием «*word2vec*» было разработано группой исследователей *Google* в 2013 году.

Автокорреляция — статистическая взаимосвязь между последовательностями величин одного ряда, взятыми со сдвигом, например, для случайного процесса — со сдвигом по времени.

Адаптивная резонансная теория (сети адаптивного резонанса, ART) — разновидность искусственных нейронных сетей, основанная на теории адаптивного резонанса Стивена Гроссберга и Гейла Карпентера. Включает в себя модели обучения с учителем и без учителя, которые используются при решении задач распознавания образов и предсказания. Основная идея заключается в том, что распознавание образов является

результатом нисходящих ожиданий и восходящей сенсорной информации. Система предлагает решение проблемы пластичности/стабильности, то есть проблемы приобретения нового знания без нарушения уже существующего.

Алгоритм *Apriori* занимается полным перебором всех возможных комбинаций наших продуктов и поиском их частот.

В алгоритме используется определенный фактор, который позволяет сократить полный перебор и снизить вычислительную сложность алгоритма.

Алгоритм t-SNE (*t-distributed stochastic neighbor embedding*) —техника нелинейного снижения размерности и визуализации многомерных переменных. Этот алгоритм может свернуть сотни измерений к меньшему количеству, сохраняя важные отношения между данными: чем ближе объекты располагаются в исходном пространстве, тем меньше расстояние между этими объектами в пространстве сокращенной размерности.

Алгоритмы извлечения ключевых слов:

- RAKE (*rapid automatic keyword extraction*)
- Графовые алгоритмы
- Нейросетевой подход

Анализ статистических связей — использование различных методов, позволяющее проверить гипотезу о наличии или отсутствии взаимосвязей между теми или иными признаками, выдвигаемую на основе содержательного анализа. Лишь посредством математических методов можно установить тесноту и характер взаимосвязей или выявить силу (степень) воздействия различных факторов на результат.

Анализ рыночной корзины (*market basket analysis*) — эта задача была предложена для поиска ассоциативных правил для нахождения типичных шаблонов покупок, совершаемых в супермаркетах.

Асимметрия — третий центральный момент. Позволяет установить симметричность распределения случайной величины относительно математического ожидания.

Ассоциативные правила позволяют находить закономерности между связанными событиями.

Большие данные (*Big Data*) представляют собой новое качество обычных данных в электронном виде, накопленных в большом объёме в разнообразных информационных системах, корпоративных или государственных, сайтах, блогах.

Векторизация — представление текста вектором чисел, где измерение — это токен, а число, которое соответствует этому измерению, — частота, с которой данный токен встречается в нашем тексте.

Визуализация (от лат. *visualis*, «зрительный») — общее название приёмов представления числовой информации или физического явления в виде, удобном для зрительного наблюдения и анализа.

Визуализация данных относится к графическому представлению данных в поисковом анализе данных, включая отображение статистических данных, структур, и т. д.

Визуализация информации работает с компьютерным интерактивным визуальным представлением абстрактных данных для усиления познавательных процессов и связана с созданием подходов для преобразования произвольной информации в интуитивно понятные формы.

Визуальная аналитика — сочетает техники автоматизированного анализа, интерактивную визуализацию для эффективного понимания, объяснения и принятия решений на базе сверхбольших и сложных наборов данных.

Вращение — это способ превращения факторов, полученных на предыдущем этапе, в более осмысленные.

Выброс — это аномально маленькое или аномально большое значение по сравнению с остальными значениями выборки.

Генеральная совокупность — вся совокупность изучаемых объектов, интересующая исследователя.

Геометрические техники визуализации — техники визуализации данных, в которых используются геометрические фигуры и различные закономерности для представления данных:

- матрица диаграмм рассеяния
- ящик-с-усами
- техники поиска проекций
- срезы (*Prosection Views*)
- многомерные фрагменты (*Hyperslice*)
- параллельные координаты

- паутинная диаграмма

Гибридные техники — одновременное использование различных техник, для большей выразительности визуализаций.

Гипотеза — частично обоснованная закономерность знаний, служащая либо для связи между различными эмпирическими фактами, либо для объяснения факта или группы фактов.

Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений.

Данные — это факты, которые мы еще не проанализировали.

Дерево решений — это набор правил («Если — то»), который можно объединить в некоторую иерархию.

Диаграмма Ганта — это тип столбчатых диаграмм, который используется для иллюстрации плана, графика работ по какому-либо проекту.

Диаграмма рассеяния (разброса) показывает взаимосвязь между двумя видами связанных данных и подтверждает их зависимость.

Динамические техники — техники визуализации данных, в которых применяются специальные инструменты для возможности интерактивной работы с данными:

- динамические проекции
- динамическое окружение
- динамическая детализация
- динамическое увеличение
- динамическое связывание

Дисперсионный анализ — основной целью дисперсионного анализа (ANOVA) является исследование значимости различия между средними с помощью сравнения (анализа) дисперсий.

Длиной последовательности называется количество предметов в этой последовательности, последовательность длины k — k-последовательностью.

Доверительный интервал — оценка какого-то истинного значения распределения, которое мы изучаем. Такой интервал представляет из себя некий диапазон от левой до правой границы, который накрывает то, что нам нужно с некоторой заранее заданной вероятностью.

Достоверность (*конфиденс, уверенность*) — это число транзакций, содержащих комбинацию наших товаров *A*, *B*, *C* и *D*, по сравнению числом транзакций, которые содержат только товар *A*.

То есть, на сколько часто в транзакции с содержанием товара *A* встречаются товары *A*, *B*, *C*, *D* одновременно.

Задача кластеризации, т. е. разбиения на группы, является субъективной и не всегда имеет «истинное решение», таким образом количество, размер (объём), состав и форма кластеров может меняться для одной и той же задачи при условии применения разных методов, или одних и тех же методов, но с разными параметрами.

Иерархические техники — техники визуализации данных, в которых данные представлены в виде неких древовидных структур, иерархий:

- многомерное наложение
- тепловые карты
- построение иерархических графиков
- Миры-внутри-Миров (*Worlds-within-Worlds*)
- древовидные графики
- «хвойные» деревья

Изменение — это движение от этого базового состояния к следующему состоянию.

Измерение — процесс присвоения чисел характеристикам изучаемых объектов согласно определенному правилу (шкале).

Инбридинг — это уже не просто набор слов, а набор всех возможных слов, которые связаны с исходными.

Интервал — это разность между максимальным и минимальным значениями переменной в наборе данных, которая показывает, в каком диапазоне распределены данные.

Информационный дизайн предназначен для ясной, прозрачной и недвусмысленной представлением информации для улучшения понимания и процесса коммуникации.

Информация — это факты, которые мы уже рассмотрели и проанализировали.

Искусственные нейронные сети (ИНС) — математические модели, а также их программные или аппаратные реализации, построенные по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма.

Искусственный интеллект (*Artificial intelligence* (*AI*)) — раздел информатики, изучающий возможность обеспечения разумных рассуждений и действий с помощью вычислительных систем и иных искусственных устройств.

Исходные данные кластерного анализа представляют собой матрицу N **объектов**, измеренных по M **показателям** ($M << N$) или матрицу расстояний (сходств) между N объектами.

Каждый объект можно рассматривать как M-мерный вектор значений характеристик.

Квантиль — это значение нашей переменной, соответствующее некой вероятности или частоте.

Квантиль — значение выборки, которое выпадает с определенной частотой или вероятностью.

Квантование — процедура преобразования данных, состоящая из 2-х шагов.

- На первом шаге диапазон значений переменной разбивается на заданное число интервалов, каждому из которых присваивается некоторый номер (уровень квантования).
- На втором шаге каждое значение заменяется номером интервала квантования.

Классификация — отнесение объектов (наблюдений, событий) к одному из заранее известных классов по некоторым правилам.

Классификация «без учителя» — разбиение множества объектов на группы, сходные по свойствам, не имея исходных представлений о структуре таких групп.

Кластерный анализ — это отдельный анализ, который делит объекты на группы по сходным значениям переменных.

Кластерный анализ (*кластер* в переводе с лат. — *скопление* или *гроздь*) является совокупностью методов, позволяющих исследователю производить классификацию «без учителя».

Корреляционный анализ — проверка наличия связи между некоторыми явлениями и определение количества силы этой связи.

Коэффициент ассоциации Юла $\mathrm{K}_c = \dfrac{ad - bc}{ad + bc}$.

Коэффициент вариации — это отношение стандартного отклонения к средней, выраженное в процентах: $V = \dfrac{\sigma}{x} \cdot 100\ \%$.

Коэффициент взаимной сопряженности Пирсона

$$\mathrm{K}_{\Pi} = \sqrt{\frac{\varphi^2}{1+\varphi^2}};\quad \varphi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\frac{n_{ij}}{n_i n_j} - 1.$$

Коэффициент взаимной сопряженности Чупрова

$$\mathrm{K}_{\Pi} = \sqrt{\frac{\varphi^2}{\sqrt{(k_1-1)(k_2-1)}}};\quad \varphi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\frac{n_{ij}}{n_i n_j} - 1.$$

Коэффициент контингенции Пирсона $\mathrm{K}_c = \dfrac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$.

Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами.

Коэффициент корреляции Спирмена — мера линейной связи между случайными величинами. Для оценки силы связи используются не численные значения, а соответствующие им ранги. Коэффициент инвариантен по отношению к любому монотонному преобразованию шкалы измерения. Вычисление коэффициента корреляции Спирмена происходит по формуле:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n^3 - n}.$$

Коэффициент ранговой корреляции Кендалла (несвязные ранги) — альтернатива методу определения корреляции Спирмана. Он предназначен для определения взаимосвязи между двумя ранговыми переменными. Вычисляется по формуле:

$$\tau = 1 - \frac{4\kappa}{n^2 - n}.$$

Коэффициент Фехнера для нелинейной функциональной связи — определение тесноты нелинейной связи.

Для расчета коэффициента Фехнера вычисляют средние выборочные значения $\bar{x}$, $\bar{y}$, затем определяют знаки совпадений $x - \bar{x}$, $y - \bar{y}$, при этом возможны сочетания:

(+, +), (–, –), (+, –), (–, +), (0, +), (0, –), (–, 0), (+, 0), (0, 0).

Далее вводятся обозначения:

- V — количество совпадений знаков
- W — количество несовпадений знаков

$$i = \frac{V - W}{V + W}.$$

Критерий доли воспроизводимой дисперсии. Факторы ранжируются по доле детерминируемой дисперсии; когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса.

Критерий значимости. Он особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Но критерий непригоден для поиска изменений в модели и реализуем только в факторном анализе по методу наименьших квадратов или максимального правдоподобия.

Критерий интерпретируемости и инвариантности. Данный критерий сочетает статистическую точность с субъективными интересами. Согласно ему, главные факторы можно выделять до тех пор, пока будет возможна их ясная интерпретация, которая зависит от величины факторных нагрузок.

Критерий Кайзера или критерий собственных чисел. Отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

Критерий каменистой осыпи (англ. *scree*) или критерий отсеивания (графический метод). Предлагается найти такое место на графике, где убывание собственных значений слева направо максимально замедляется.

Кросс-валидация — метод оценки аналитической модели и её поведения на независимых данных.

Кросс-табуляция — статистический метод, который одновременно характеризует две или более переменных и заключается в создании таблиц сопряженности признаков, отражающих совместное распределение двух или более переменных с ограниченным числом категорий или определенными значениями.

Лемматизация — процедура, в результате которой удаляются только флективные окончания и возвращается основная, или словарная, форма слова, называемая леммой.

Лифт — во сколько раз совместные продажи товаров происходят *чаще* по сравнению с продажей отдельного товара из этого набора. Лифт определяется, как отношение поддержки и достоверности.

Логистическая регрессия (логит-регрессия) — это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события путём подгонки данных к логистической кривой.

Случайный процесс называется марковским, если вероятность любого состояния в будущем зависит только от его состояния в настоящем и не зависит от того, когда и каким образом процесс оказался в этом состоянии.

Описывающий поведение системы процесс называется цепью Маркова.

Матрица расстояний означает набор значений абстрактных расстояний ρ или сходств φ, являющихся мерой совпадения объектных векторов.

Машинное обучение (*Machine Learning*) — обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Медиана — это значение, которое выпадает с 50-ти процентной вероятностью.

Медиана — это значение, которое разбивает выборку на две равные части. Половина наблюдений лежит ниже медианы, и половина наблюдений лежит выше медианы.

Меры среднего уровня дают усредненную характеристику совокупности объектов по определенному признаку.

Метод «ближайшего соседа» — объединяются кластеры с наименьшим расстоянием между элементами.

Метод главных компонент (МГК) применяется для снижения размерности пространства наблюдаемых векторов, не приводя к существенной потере информативности

Метод главных факторов — цель метода: выявление и интерпретации латентных общих факторов, минимизация их количества и степени зависимости от своих специфических остаточных случайных компонент.

Метод «дальнего соседа» — объединяются кластеры с наибольшим расстоянием между элементами.

Метод деревьев решений (*decision trees*) является одним из наиболее популярных методов решения задач классификации и прогнозирования. Иногда этот метод *Data Mining* также называют деревьями решающих правил, деревьями классификации и регрессии. Если зависимая, т. е. целевая переменная принимает дискретные значения, при помощи метода дерева решений решается задача классификации. Если же зависимая переменная принимает непрерывные значения, то дерево решений устанавливает зависимость этой переменной от независимых переменных, т. е. решает задачу численного прогнозирования.

Метод иерархической агломерации заключается в последовательном объединении N исходных объектов до момента, пока все они не будут объединены в один кластер объёма N. С учётом того, что на каждом шаге подвергаются слиянию только два кластера, процедура содержит $N - 1$ шагов объединения.

Весь процесс объединения изображают в виде дендрограммы — графика, на котором по оси абсцисс нанесены номера объектов, по оси ординат изображено расстояние объединения. Данный график показывает состав кластерного решения на каждом шаге объединения.

Метод Краскела–Уоллиса предназначен для проверки равенства медиан нескольких выборок. Метод является ранговым.

Метод к-средних или метод Мак-Кина заключается в разбиении всего исходного множества объектов вокруг заданных на первом этапе начальных центров k кластеров. Эти центры могут быть получены как с помощью специальных процедур, на основе априорных предположений, так и извлечены из выборки случайным образом.

Метод наивного Байеса — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости.

Метод наименьших квадратов — это графический метод оценивания параметров линейной модели на основе минимизации суммы квадратов отклонений наблюдаемых и модельных i (расчетных) значений зависимой переменной. Метод может быть использован для расчета коэффициентов только линейного уравнения регрессии.

Методы систематизации данных — методы, с помощью которых происходит перегруппировка данных с целью выявления и осмысления важных тенденций и отклонений:

- расчет статистик
- проверка статистических тестов и гипотез
- построение диаграмм

Метод Уорда (Варда, *Ward*) — объединяются кластеры, дающие наименьшую суммарную дисперсию.

Метод центроидов — объединяются кластеры с наименьшим расстоянием между центрами кластеров.

Множественный коэффициент корреляции характеризует степень линейной статистической связи (зависимости) результативного y и линейной комбинации факторных x_1, x_2, …, x_n признаков:

$$r_{yx_i} = \frac{\sum_{j=1}^{n}(x_{ij} - \frac{\sum_{j=1}^{n} x_{ij}}{n})(y_{ij} - \frac{\sum_{j=1}^{n} x_{ij}}{n})}{\sqrt{\sum_{j=1}^{n}(x_{ij} - \frac{n\sum_{j=1}^{n} x_{ij}}{n})^2 (y_{ij} - \frac{\sum_{j=1}^{n} x_{ij}}{n})^2}}.$$

Мода — это значение, которое выпадает в выборке наиболее часто.

Модели ART:

- ART1 (от англ. ART — *Adaptive Resonance Theory*) — модель сети адаптивного резонанса для кластеризации, хранения и идентификации образов в форме двоичных сигналов (обучение без учителя)
- ART2 — модель сети адаптивного резонанса для кластеризации, хранения и идентификации образов, представленных как в форме двоичных сигналов, так и в форме аналоговых сигналов, в том числе с использованием обоих типов сигналов в одной структуре (обучение без учителя)
- ART3 — физиологически более реалистичная версия ART2. Она моделирует медиаторную регуляцию синаптической активности
- Fuzzy ART— вариант модели с применением принципов нечеткой логики
- ARTMAP — модель сети адаптивного резонанса для классификации образов (обучение с учителем)

Мультиколлинеарность — тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат, которая затрудняет оценивание регрессионных параметров.

Мягкая кластеризация (англ. *fuzzy clustering* и *soft clustering*) — тип кластеризации, при котором каждая точка может принадлежать одному или нескольким кластерам. Мягкая кластеризация также называется нечёткой кластеризацией и используется при решении задач обработки естественного языка, в том числе в лексической семантике.

Научная визуализация изучает потенциально огромные объемы научных данных, получаемых от сенсоров, моделирований и лабораторных тестов.

Основной задачей данного направления является преобразование информации на различных уровнях сложности в более доступные для широкой аудитории формы.

Неравномерное (однородное) квантование — преобразование, при котором диапазон значений переменной разбивается на интервалы различной длины (асимметричные). Имеет смысл, если в значениях нет пропусков или сгустков.

Несмещенная оценка — это точечная оценка, чье математическое ожидание равно оцениваемому параметру.

Низкая растерянность указывает на то, что распределение вероятностей хорошо подходит для прогнозирования выборки.

«Облако тегов» — это визуальное представление списка категорий (или тегов, также называемых метками, ярлыками, ключевыми словами и т. п.). Ключевые слова чаще всего

представляют собой отдельные слова, и важность каждого ключевого слова обозначается размером шрифта или цветом.

Обогащение данных — процесс насыщения данных новой информацией, которая позволяет сделать их более ценными и значимыми с точки зрения решения той или иной аналитической задачи.

Обучение без учителя (самообучение, спонтанное обучение, *Unsupervised learning*) — один из способов машинного обучения, при котором испытуемая система спонтанно обучается выполнять поставленную задачу без вмешательства со стороны экспериментатора.

Обучение с подкреплением — это метод машинного обучения, при котором происходит обучение модели, которая не имеет сведений о системе, но имеет возможность производить какие-либо действия в ней. Действия переводят систему в новое состояние и модель получает от системы некоторое вознаграждение.

Обучение с учителем (*Supervised learning*) — один из способов машинного обучения, в ходе которого испытуемая система принудительно обучается на *обучающей выборке*. На основе этих данных требуется восстановить зависимость, т. е. построить алгоритм, эффективно работающий на новых данных (пригодный для прогнозирования).

Объем выборки — число случаев, включенных в выборочную совокупность. Выборки можно условно разделить на малые, большие и сверхбольшие.

Ошибкой правила является количество объектов, имеющих данное значение рассматриваемой независимой переменной ($A^j = x_i^j$), но не имеющих наиболее часто встречающееся значение зависимой переменной у данного значения независимой переменной ($C \neq c_r$).

Параметры — числовые характеристики генеральной совокупности.

Переменная — свойство или характеристика, общая для всех изучаемых объектов, проявление которой может изменяться от объекта к объекту.

Значение переменной является проявлением признака.

Переменные могут являться числовыми данными либо символьными.

Перплексия (растерянность) — это измерение того, насколько хорошо распределение вероятностей или модель вероятности предсказывает выборку.

Может использоваться для сравнения вероятностных моделей.

Подтверждающий анализ — процесс направленный на проверку выдвинутых ранее предположений на основе графического представления.

Поисковый анализ — процесс изучения и анализа данных для поиска неявной, но потенциально полезной информации.

Последовательность – это упорядоченный список предметных наборов.

Правило Стерджеса — эмпирическое правило определения оптимального количества интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при построении гистограммы плотности ее распределения.
Количество интервалов n определяется как:

$$n - 1 = [\log_2 N],$$

где N — общее число наблюдений величины.

Предметный набор — это непустой набор предметов (товаров), появившихся в одной транзакции.

Презентация результатов — процесс графического представления результатов анализа в формате, основанном на потребностях заказчика.

Простая линейная регрессия записывается уравнением: $y = a + bx$, где x — независимая переменная, y — зависимая переменная, a — свободный член линии оценки, b — угловой коэффициент или градиент оценённой линии.

Прунинг (*pruning*) — процесс обрезки лишних веток

Путь пользователей — это последовательность страниц, на которых он находится.

Равномерное (однородное) квантование — преобразование, при котором диапазон значений переменной разбивается на интервалы одинаковой длины. Имеет смысл, если значения распределены равномерно по всему диапазону значений.

Ранговая корреляция — определение силы зависимости между случайными величинами по рангам (порядковым номерам). Используется, когда невозможно определить численно силу зависимости при помощи обычного коэффициента корреляции.

Регрессионный анализ — определение характера связи между явлениями, модели связи между явлениями и свойств этой модели.

Регрессионный анализ — техника моделирования данных, направленная на исследование их взаимосвязи. В простейшем случае регрессионный анализ используют для построения моделей прогнозирования новых числовых значений на основе набора известных значений.

Редуцированная матрица — это матрица, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины.

Результатом работы методов кластерного анализа обычно являются вектор принадлежности объектов к кластерам, таблица объёмов кластеров (количества объектов, содержащихся внутри кластеров) и в некоторых случаях значения критериев качества кластеризации.

Рекуррентные нейронные сети (*Recurrent Neural Network,* RNN) — Это класс моделей машинного обучения, основанный на использовании предыдущих состояний сети для вычисления текущего.

Сессия — это период активности, который заканчивается паузой.

Сети Больцмана (ограниченная машина Больцмана) — вид генеративной стохастической нейронной сети, которая определяет распределение вероятности на входных образцах данных. В ней нейроны одного типа не связаны между собой. Ограниченную машину Больцмана можно обучать как FFNN, но с одним нюансом: вместо прямой передачи данных и обратного распространения ошибки нужно передавать данные сперва в прямом направлении, затем в обратном. После этого проходит обучение по методу прямого и обратного распространения ошибки.

Сети Кохонена — класс нейронных сетей, основным элементом которых является слой Кохонена. Слой Кохонена состоит из адаптивных линейных сумматоров («линейных формальных нейронов»). Как правило, выходные сигналы слоя Кохонена обрабатываются по правилу «Победитель получает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль.

Сети Хопфилда — это полносвязная нейронная сеть с симметричной матрицей связей. Во время получения входных данных каждый узел является входом, в процессе обучения он становится скрытым, а затем становится выходом. Сеть обучается так: значения нейронов устанавливаются в соответствии с желаемым шаблоном, после чего вычисляются веса, которые в дальнейшем не меняются. После того, как сеть обучилась на одном или нескольких шаблонах, она всегда будет сводиться к одному из них (но не всегда — к желаемому).

Соревновательные сети — нейронные сети с алгоритмом обучения по методу соревнования. В таких сетях нейроны конкурируют друг с другом за право быть «победителем». К соревновательным сетям относятся нейронные сети группы «победитель получает все» (*winner takes all*), самоорганизующиеся карты Кохонена. Соревновательное обучение применяется также в RBF-сетях вместе с коррекцией ошибки.

Состоятельность — это свойство, которое означает, что по вероятности оценка приближается к истинному значению величины.

Среднее значение — характеристика положения, некоторое число, заключённое между наименьшим и наибольшим из их значений. Рассчитывается как среднее арифметическое или среднее геометрическое.

Три стандарта сферы *data science* (науки о данных):

- KDD
- SEMMA
- CRISP-DM

Стандартная ошибка — характеристика рассеивания, отклонение величины относительно своего среднего.

Стандартное отклонение — в теории вероятностей и статистике наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания.

Статистики — числовые характеристики выборки.

Статистическая связь — это объективная количественная закономерность изменения массовых явлений и процессов, то есть статистическая закономерность является количественной формой проявления причинной связи.

Основные статистические показатели:

- меры среднего уровня
- меры рассеяния

Стемминг — это процесс нахождения основы слова для заданного исходного слова. Основа слова необязательно совпадает с морфологическим корнем слова.

Стоп-слова — это слова, не несущие какой-либо самостоятельной смысловой нагрузки.

Счет — сумма частот значений или просто объем выборки. Показывает качество данных в выборке.

Текстовый анализ — это процесс обнаружения потенциально полезных и понятных шаблонов в неструктурированных текстовых данных.

Текстовый анализ — это нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных.

Теорема Колмогорова — любая непрерывная функция n переменных на единичном отрезке $[0; 1]$ может быть представлена в виде суммы конечного числа одномерных функций:

$$f(x_1, x_2, \dots, x_n) = \sum_{p=1}^{2n+1} g\left(\sum_{i=1}^{n} \lambda_i \varphi_p(x_i)\right),$$

где функции g и φ_p являются одномерными и непрерывными, $\lambda_i = const$ для всех i.

Тест Стьюдента — общее название для статистических тестов, в которых статистика критерия имеет распределение Стьюдента. Такие тесты наиболее часто применяются для проверки равенства средних значений в двух выборках.

Тест хи-квадрат — любой статистический тест гипотезы, где распределение выборки тестовой статистики является хи-квадрат распределение, когда нулевая гипотеза верна. Тест хи-квадрат используется для определения того, есть ли существенная разница между ожидаемыми частотами и наблюдаемыми частотами в одной или нескольких категорий.

При проведении теста хи-квадрат проверяется взаимная независимость **двух переменных** таблицы сопряженности и благодаря этому косвенно выясняется зависимость обоих переменных.

Токенизация заключается в разбиении длинных участков текста на более мелкие (токены), имеющие определенное значение.

Трансформация (преобразование) данных — комплекс методов и алгоритмов, направленных на оптимизацию представления и форматов данных с точки зрения решаемых задач и целей анализа.

Факторный анализ (ФА) представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов.

Функция нормального распределения (распределения Гаусса): $N(\mu, \sigma^2)$, где μ — математическое ожидание (среднее значение), медиана и мода распределения, а параметр σ — среднеквадратическое отклонение (σ^2 — дисперсия) распределения.

Цель описательной статистики — обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, а также их количественное описание посредством основных статистических показателей.

Цель трансформации данных — представить информацию в таком виде, чтобы она могла быть использована наиболее эффективно.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии.

Частные коэффициенты корреляции, измеряющие влияние на y фактора x_i при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_1x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}}.$$

Частотная диаграмма показывает частоту или количество элементов, которые попадают в различные диапазоны значений выборки, если это непрерывные значения.

Частый набор — это набор из наших продуктов, которые наиболее часто встречаются во всех наших чеках.

Последовательности, удовлетворяющие ограничению минимальной поддержки, называются частыми последовательностями.

Чек — это все продажи, все цепочки, которые сделали клиенты (один клиент), за какой-то временной интервал (день).

Шкалы определяют способность наших элементов взаимодействовать друг с другом совершенно различными способами.

Эксцесс характеризует степень концентрации случаев вокруг среднего значения и является своеобразной мерой крутизны кривой.

Эффективная несмещенная оценка определяется как оценка с минимальной дисперсией.