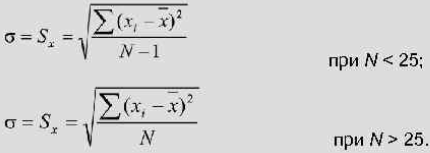
Алгоритм взаимосвязей или ассоциативных правил позволяет выявить часто встречающиеся сочетания элементов данных и использовать обнаруженные закономерности для построения прогноза.

Кластеризация последовательностей – выявить часто встречающиеся последовательности событий. В данном случае учитывается, в какой очередности события происходят, схожие последовательности объединяются в кластер. Кроме анализа характеристик кластеров.

**Среднее квадратичное отклонение (СКО)**

Выборочную оценку среднеквадратичного отклонения генеральной совокупности описывает следующими способами: если выполняется нормальный закон распределения, то величину среднеквадратичного отклонения для рада x вычисляют:

Связь с дисперсии (момент второго порядка) и СКО:

**Интерпретация величины среднеквадратического отклонения**

Большее значение среднеквадратического отклонения показывает больший разброс значений в представленном множестве со средней величиной множества; меньшее значение, соответственно, показывает, что значения в множестве сгруппированы вокруг среднего значения.

В общем смысле среднеквадратическое отклонение можно считать мерой неопределенности. К примеру, в физике среднеквадратическое отклонение используется для определения погрешности серии последовательных измерений какой-либо величины. Это значение очень важно для определения правдоподобности изучаемого явления в сравнении с предсказанным теорией значением: если среднее значение измерений сильно от предсказанных теорией значений (большое значение среднеквадратического отклонения), то полученные значений или метод их получения следует перепроверить.

**Кодирование категориальных признаков**

Иногда полезно разбить объекты на категории не по количеству, а по качеству. Это качественная информация представляется как принадлежность наблюдений к отдельной категории. Набор категорий без внутреннего упорядочивания называется номинальным. Набор категорий, имеющий некоторое естественное упорядочивание называется порядковым.

Кодирование категориальных признаков

df\_dummies=pd.get\_dummies(df[“type”])

Кодирование порядковых категориальных признаков (с 111 в учебнике)

**Машинное обучение. Метод ближайших соседей**

Метрический классификатор и метод ближайших соседей — ВикиконспектыМетод ближайшего соседа основан на хранении известных данных для сравнения с новыми элементами

При классификации неизвестного объекта находится отклонение между ним и известными объектами, и идентифицируется ближайший сосед

Метод k-ближайших соседей – для повышения надежности объект относится к тому классу, которому принадлежит большинство из k его соседей

В задачах с двумя классами число соседей k берут нечётным, чтобы не возникало ситуаций неоднозначности.

Метод взвешенных ближайших соседей – в задачах с числом классов 3 и более нечётность уже не помогает и ситуации неоднозначности все равно могут возникать

Тогда i-му соседу приписывается вес wh убывающей с ростом ранга соседа

Объект относится к тому классу, который набирает больший суммарный вес среди k ближайших соседей.

**Проблемы и решения**

Выбор метрики.

Проблема сложная, однозначного решения нет. Если используется евклидово расстояние, признаки должны быть нормированы, иначе признаки с наибольшими числовыми значениями будут доминировать

**Деревья решений. Случайный лес**

CART (Classification and Regression Tree) - это алгоритм построения бинарного дерева решений. Каждый узел дерева при разбиении имеет только двух потомков. Как видно из названия алгоритма, решает задачи классификации и регрессии. Используется в scikit-learn.

C4.5 – Алгоритм построения дерева решений, количество потомков у узла не ограничено. Решает только задачи классификации

|  |  |  |  |
| --- | --- | --- | --- |
| Проверка | Узел | Лист | Глубина |
| Условие в узле | Внутренний узел дерева, узел проверки | Конечный узел дерева, узел решения | Самое большое количество родителей |



Большинство из известных алгоритмов являются «жадными алгоритмами»

Перебирают все признаки и все возможные значения в выборке

Если один раз был выбран атрибут, и по нему было произведено разбиение на подмножества, то алгоритм не может вернуться назад и выбрать другой атрибут, который дал бы лучшее разбиение

И поэтому на этапе построения нельзя сказать даст ли выбранный атрибут, в конечном итоге, оптимальное разбиение

Первое множество (корень): есть ли такие два подмножества, которые улучшат предсказания? Подмножества определим по какому-то признаки и его значению. Делим первое множество.

И так поступаем далее. Однако, если улучшения нет, или достигли порог ошибки, или решили остаться на этой глубине дерева, то это лист.

**Достоинства и недостатки деревьев решений**

Достоинства:

* Поддерживают работу с входными переменными разных (смешанных типов) типов
* Возможность обрабатывать данные с пропущенными значениями
* Устойчивы к выбросам
* Нечувствительные к монотонным преобразваниям входных переменных
* Поддерживают работу с большими данными
* Возможность интерпретации построенного решающего правила

Недостатки:

* Основной недостаток – плохая предсказательные (обобщающая) способность

**Работа с числовыми данными. Шкалирование, стандартизация, нормализация признака**

1. Прошкалровать числовой признак в диапазоне между двумя значениями

2. Стандартизация – требуется преобразовать признак, чтобы он имел среднее значение – и стандартное отклонение 1. Если выбросов много, то лучше не использовать.

3. Класс normalizer шкалирует значения в отдельных наблюдениях, приводя их к единичной норме. Этот тип шкалирования часто используется, когда имеется много эквивалентных признаков. Каждое слово или группа слов является признаком.

**Обработка выбросов**

При обработке выбросов используют три стратегии:

1. Отбросить их

Логистическая регрессией называется регрессией, хотя на самом деле является методом контролируемой классификации. Вектор целей в логистической регрессии может принимать только два значения

**Работа с данными и временем**

**Опорновекторные машины**

Опорновекторные машины классифицируют данные путём нахождения гиперплоскости, которые максимизируют допустимый промежуток между классами в тренировочных данных.  
From sklearn.svm import linearSVC [#опорновекторная](https://vk.com/im?sel=174651667&st=%23%D0%BE%D0%BF%D0%BE%D1%80%D0%BD%D0%BE%D0%B2%D0%B5%D0%BA%D1%82%D0%BE%D1%80%D0%BD%D0%B0%D1%8F) классификация