

Введение в анализ данных

25 июля 2019 г.

Понятие интеллектуального анализа данных соответствует широко распространенному термину Data Mining, который часто переводится как добыча данных, глубинный анализ данных, извлечение знаний, раскопка знаний в базах данных.

Вот популярные слова из описания вакансий специалиста Data Science: *SQL, Python, R, SAS, Hadoop, Java, optimization, C++, visualization, MATLAB, Business Intelligence, distributed computation, machine learning, unstructured, Hive.*

Видно, насколько разносторонней (междисциплинарной) должна быть эта квалификация с точки зрения современного работодателя.

Data Mining -- мультидисциплинарная область, возникшая и развивающаяся на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др.

Data Mining это процесс необходимый для поддержки принятия решений, основанный на поиске в данных скрытых закономерностей или аномалий.

Основы обработки данных и основы алгоритмизации

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты. Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций.

Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки. Иными словами, данные -- это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных.

Набор данных представляют в простом случае двухмерной таблицей, в которой строки это объекты, а столбцы это атрибуты или еще их называют измерения

Атрибут - свойство, характеризующее объект: цвет глаз человека, температура воды и т.д. Атрибут также называют полем таблицы.

Адекватность информации - это уровень соответствия образа, создаваемого с помощью информации, реальному объекту, процессу, явлению.

От степени адекватности информации зависит правильность принятия решения.

Адекватность информации может выражаться в трех формах: синтаксической, семантической и прагматической.

Синтаксическая адекватность отображает формально-структурные характеристики информации, не затрагивая ее смыслового содержания. На синтаксическом уровне учитываются тип носителя и способ представления информации, скорость ее передачи и обработки, размеры кодов представления информации, надежность и точность преобразования этих кодов и т. д.

Информацию, рассматриваемую с таких позиций, обычно называют данными.

Семантическая адекватность определяет степень соответствия образа объекта самому объекту. Здесь учитывается смысловое содержание информации. На этом уровне анализируются сведения, отражаемые информацией, рассматриваются смысловые связи. Таким образом, семантическая адекватность проявляется при наличии единства информации и пользователя. Эта форма служит для формирования понятий и представлений, выявления смысла, содержания информации и ее обобщения.

Прагматическая адекватность отражает соответствие информации цели управления, реализуемой на ее основе. Прагматические свойства информации проявляются при наличии единства информации, пользователя и цели управления. На этом уровне анализируются потребительские свойства информации, связанные с практическим использованием информации, с соответствием ее целевой функции деятельности системы.

Информационный процесс

- Поиск информации — это извлечение информации для сбора и хранения
- Сбор информации предназначен для придания информации адекватности. Для этого необходима обработка информации, то есть преобразование информации из одного вида в другой, осуществляемое по формальным правилам.
- Хранение информации — это способ распространения информации в пространстве и времени в доступном виде. Другая цель это многократное использование информация.
- Передача. В процессе передачи информации обязательно участвуют источник и приемник информации: первый передает информацию, второй ее получает. Между ними действует канал передачи информации — канал связи.

Временной ряд — совокупность наблюдений одного и того же показателя в различные моменты времени (обычно последовательные: дни, недели, месяцы, кварталы, годы и т.д.).

Показатели: цены акций, уровень безработицы, обменный курс евро/рубль и т.д. Характерной особенностью временных рядов является естественным образом зафиксированный порядок наблюдений.

"Перекрёстные данные" — это тип данных, собранный путем наблюдения за многими объектами (такими как физические лица, фирмы, страны или регионы) в один и тот же период времени.

Для перекрестных данных (дословный перевод термина cross-sectional data, иногда он переводится как «пространственные данные» в противопоставление временным, или «временной срез») момент времени зафиксирован, и имеются наблюдения для различных однородных объектов (в этот момент).

Панельные данные — это набор наблюдений за некоторыми однородными объектами в различные моменты времени. Если объекты в разные моменты времени одни и те же, то панель называется сбалансированной (например, известны данные о потреблении 10 индивидов в течение 5 лет без пропусков), а если различаются — то несбалансированной (например, известны данные о потреблении 4 индивидов в течение 5 лет без пропусков, 5 индивидов в течение первых 4 лет и 1 индивида в течение последних 3 лет).

1. Описательная статистика
2. Фильтрация
3. Заполнение пропусков
4. Дубликаты и противоречия
5. Редактирование выбросов
6. Спектральная обработка предназначена для очистки от шумовой составляющей и сглаживания рядов данных. Сглаживание необходимо в том случае, когда ряд данных оказывается неравномерным, содержит большое количество мелких структур, препятствующих исследованию более значительных объектов и закономерностей.

MapReduce предполагает, что обработка данных происходит в 3 стадии:

1. Стадия Map.
2. Стадия Shuffle.
3. Стадия Reduce.

Схема map-reduce. Лямбда-архитектура

Лямбда-архитектура - это архитектура обработки данных, основанная на использовании схемы map-reduce, разработанная для обработки огромных объемов данных с использованием методов пакетной и потоковой обработки. Целью такого подхода является балансировка: задержки, пропускной способности и отказоустойчивости.

Схема map-reduce. Стадия Map

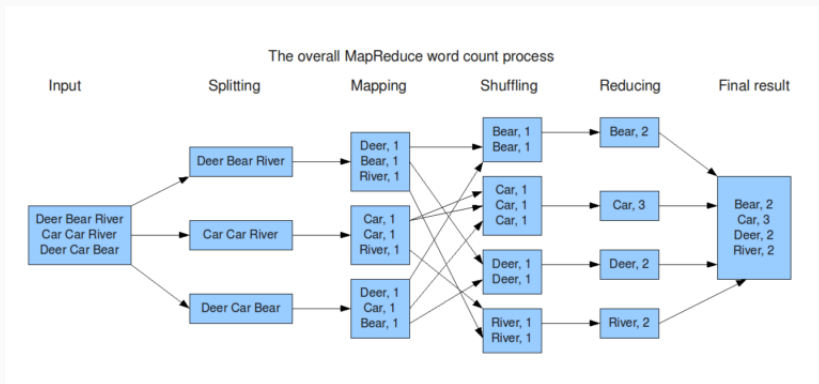
На этой стадии данные преобразуются при помощи функции `map()`, которую определяет пользователь. Работа этой стадии заключается в преобразовке и фильтрации данных. Работа очень похожа на операцию `map` в функциональных языках программирования – пользовательская функция применяется к каждой входной записи.

Функция `map()` примененная к одной входной записи и выдаёт множество пар ключ-значение. Множество – т.е. может выдать только одну запись, может не выдать ничего, а может выдать несколько пар ключ-значение. Что будет находится в ключе и в значении – решать пользователю, но ключ – очень важная вещь, так как данные с одним ключом в будущем попадут в один экземпляр функции `reduce`.

Схема map-reduce. Стадия shuffle и reduce.

- Стадия Shuffle. Проходит незаметно для пользователя. В этой стадии вывод функции map «разбирается по корзинам» – каждая корзина соответствует одному ключу вывода стадии map. В дальнейшем эти корзины послужат входом для reduce.
- Стадия Reduce. Каждая «корзина» со значениями, сформированная на стадии shuffle, попадает на вход функции reduce().
Функция reduce задаётся пользователем и вычисляет финальный результат для отдельной «корзины». Множество всех значений, возвращённых функцией reduce(), является финальным результатом MapReduce-задачи.

Схема map-reduce



[http : //www.igfasouza.com/blog/](http://www.igfasouza.com/blog/)

Существующие, наборы данных, визуализация модели классификации

Работа с «грязными» данными: редактирование, форматирование, удаление недопустимых значений и дубликатов. Агрегирование данных. Основные модели классификации

Рассматривать мы будем программный продукт Orange3, позволяющий интерактивно работать с данными. Для знакомства выполним несколько заданий

Противоречивость информации;

Пропуски в данных;

Аномальные значения;

Шум;

Ошибки ввода данных.

Агрегирование позволяет преобразовать группы наблюдений в наблюдения, содержащие агрегированную информацию по соответствующей группе, и создавать новый, агрегированный, набор данных. Наблюдения агрегируются на основе значений нуля или большего числа группирующих переменных. Если группирующие переменные не заданы, то весь набор данных представляет собой единую группу для агрегирования.

Агрегация должна содержать по одному наблюдению на каждую группу, определяемую группирующей переменной.

Группирующие переменные. Наблюдения разбиваются на группы, на основании значений этих переменных. Каждая уникальная комбинация значений группирующих переменных определяет группу.

Агрегируемые переменные. Для создания новых переменных используются исходные переменные с групповыми функциями.

Классификатор - систематизированный перечень наименованных объектов, каждому из которых в соответствие дан уникальный код.

В классификаторах применяется три метода классификации: иерархический, фасетный и дескрипторный. Выбор между этими тремя методами зависит от особенностей конкретной предметной области. Существуют следующие требования для выбранной системы классификации:

- достаточная емкость и необходимая полнота, которые гарантируют охват всех объектов классификации в заданных границах;
- оправданная глубина;
- обеспечение возможности решения комплекса задач различного уровня;

Иерархическим методом классификации понимается метод, при котором объекты делятся на вложенные или подчиненные подмножества, движение идет в сторону конкретизации объектов. При этом основанием деления служит одно из свойств объекта. Совокупность получившихся группировок при этом образует иерархическую древовидную структуру, узлами которого являются группировки.

Выбор последовательности признаков зависит, прежде всего, от характера информации. При построении классификации выбор последовательности признаков зависит от вероятности обращения к тому или иному признаку. При этом наиболее вероятным обращениям должны соответствовать высшие уровни классификации.

Требования к классификатору, построенному на иерархическом методе классификации: Классификационные группировки, расположенные на одной ступени классификатора, не должны пересекаться, то есть не должны включать в себя аналогичных понятий. На каждой ступени классификатора для разделения вышестоящей группировки должен использоваться только один признак. Сумма подмножества всегда должна давать делимое множество объектов; не должна оставаться часть объектов, не вошедших в состав классификационной группировки.

Основными преимуществами иерархического метода является большая информационная ёмкость, традиционность и привычность применения, возможность создания для объектов классификации мнемонических кодов, несущих смысловую нагрузку.

Значительным недостатком иерархической классификации является слабая гибкость структуры, обусловленная фиксированным основанием деления и заранее установленным порядком следования, не допускающим включение новых объектов и классификационных группировок.

Фасетный метод классификации подразумевает параллельное разделение множества объектов на независимые классификационные группировки. При этом не предполагается жёсткой классификационной структуры и заранее построенных конечных группировок. Классификационные группировки образуются путём комбинации значений, взятых из соответствующих фасетов. Последовательность расположения фасетов при образовании классификационной группировки задается "фасетной формулой". Количество фасетных формул определяется возможными сочетаниями признаков.

К классификатору, построенному на фасетном методе классификации, предъявляются следующие требования:

1. Должен соблюдаться принцип непересекаемости фасета, то есть состав признаков одного фасета не должен повторяться в других фасетах этого же класса;
2. В состав классификатора должны быть включены только такие фасеты и признаки, которые необходимы для решения конкретных задач.

Основным "преимуществом" классификации с использованием фасетного метода является гибкость структуры её построения. Изменения в любом из фасетов не оказывают существенного влияния на все остальные. При фасетной классификации появляется возможность объектов и осуществления информационного поиска по любому сочетанию фасетов.

"Недостатками" фасетного метода классификации являются неполное использование ёмкости, нетрадиционность и иногда сложность применения.

Для организации поиска информации, для ведения тезаурусов (словарей) эффективно используется дескрипторная (описательная) система классификации, язык которой приближается к естественному языку описания информационных объектов:

Методы классификации: Дескрипторная классификация

1. отбирается совокупность ключевых слов или словосочетаний, описывающих определенную предметную область или совокупность однородных объектов.
2. выбранные ключевые слова и словосочетания подвергаются нормализации, т.е. из совокупности синонимов выбирается один или несколько наиболее употребимых;
3. создается словарь дескрипторов, т.е. словарь ключевых слов и словосочетаний, отобранных в результате процедуры нормализации.
4. между дескрипторами устанавливаются связи, которые позволяют расширить область поиска информации.

Связи могут быть трех видов:

1. синонимические, указывающие на некоторую совокупность ключевых слов как синонимов («студент – учащийся – обучаемый»);
2. родовидовые, отображающие включение некоторого класса объектов в более представительный класс («университет – факультет – кафедра»);
3. ассоциативные, соединяющие дескрипторы, обладающие общими свойствами («студент – экзамен – профессор – аудитория»).

Объекты и признаки. Типы шкал.

**Типы задач: классификация,
регрессия, прогнозирование,
ранжирование.**

Абсолютные статистические величины. Относительные статистические величины. Описательная статистика: медиана, среднее, разница применения и их трактовки. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, ранжирование.

Изучая массовые общественные явления, статистика в своих выводах опирается на числовые данные, полученные в конкретных условиях места и времени. Результаты статистического наблюдения регистрируются прежде всего в форме первичных абсолютных величин. . Абсолютная величина отражает уровень развития явления.

В статистике все абсолютные величины являются именованными, измеряются в конкретных единицах и, в отличие от математического понятия абсолютной величины, могут быть как положительными, так и отрицательными (убытки, убыль, потери и т.п.).

Относительная величина в статистике – это обобщающий показатель, который дает числовую меру соотношения двух сопоставляемых абсолютных величин. Так как многие абсолютные величины взаимосвязаны, то и относительные величины одного типа в ряде случаев могут определяться через относительные величины другого типа.

Относительные статистические величины.

Основное условие правильного расчета относительной величины – сопоставимость сравниваемых показателей и наличие реальных связей между изучаемыми явлениями. Таким образом, по способу получения относительные показатели – всегда величины производные, определяемые в форме коэффициентов, процентов, промилле, продецимилле и т.п. Однако нужно помнить, что этим безразмерным по форме показателям может быть, в сущности, приписана конкретная, и иногда довольно сложная, единица измерения.

Объекты классифицированы, классам присвоены (наименований) словесные наименования или условные номера - коды. То, что номер одного класса больше или меньше другого, еще ничего не говорит о свойствах объектов, относящихся к этим классам, за исключением того, что они различаются.

Объекты классифицированы, а классы обозначены номерами (закодированы). Значения чисел, присваиваемые классам, качественно отражают степень выраженности определенных свойств предметов, принадлежащих этим классам. То есть большим значениям кодов классов соответствует и большая степень выраженности измеряемого свойства, на основании чего классы можно ранжировать.

Существует единица измерения, при помощи которой классы можно не только упорядочить, но и приписать им числа так, чтобы равные разности чисел присвоенных классам, отражали равные различия в количествах измеряемых свойств. Нулевая точка интервальной шкалы произвольна (условна) и не указывает на отсутствие свойства.

Типы задач: классификация, регрессия, прогнозирование, ранжирование

- Классификация
- Регрессионный анализ
- Прогнозирование
- Ранжирование
- Кластеризация

Типы задач: классификация

Классификация — один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделённых некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

Классифицировать объект — значит, указать номер (или наименование класса), к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Типы задач: классификация

В математической статистике задачи классификации называются также задачами дискриминантного анализа.

В машинном обучении задача классификации относится к разделу обучения с учителем. Существует также обучение без учителя, когда разделение объектов обучающей выборки на классы не задаётся, и требуется классифицировать объекты только на основе их сходства друг с другом. В этом случае принято говорить о задачах кластеризации или таксономии, и классы называть, соответственно, кластерами или таксонами.

Типы задач: регрессия

Регрессионный анализ — метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной).

Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной.

Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента.

Прогноз — это процесс или результат предсказания тех или иных фактов, событий, явлений, величин, которые станут известны лишь в будущем по отношению к моменту времени, в котором создается прогноз. Под прогнозом также иногда понимают модель будущего события, явления и т. п.

Прогнозирование — это процесс (часто основанный на научном исследовании) по расчету прогноза или разработке прогнозной модели.

В узком смысле под прогнозированием понимают предсказание будущих значений временного ряда на основе его значений в прошлом, и, возможно, дополнительной информации. Такую дополнительную информацию представляют влияющие на ситуацию внешние факторы. Например, для прогнозирования временного ряда спроса на какой-либо товар это могут быть предложения замещающих товаров, для экспортных грузоперевозок - таможенные пошлины и курс доллара, для цен акций — политические решения и т. д.

Задачи заключающиеся в подборе ранжирующей модели по обучающей выборке, состоящей из множества списков и заданных частичных порядков на элементах внутри каждого списка. Частичный порядок обычно задаётся путём указания оценки для каждого элемента (например, «релевантен» или «не релевантен»; возможно использование и более, чем двух градаций). Цель ранжирующей модели — наилучшим образом (в некотором смысле) приблизить и обобщить способ ранжирования в обучающей выборке на новые данные.

Показатели вариации. Линейные и нелинейные модели регрессии.

Показатели вариации: дисперсия, стандартное квадратичное отклонение. Перцентили, метод расчета, применение, трактовка. Асимметрия, эксцесс, их применение и трактовка. Понятие выброса. Снижение волатильности через монотонное преобразование. Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия. Понятие причинности, корреляция. Получение и трактовка коэффициентов.

Показатели вариации: дисперсия, стандартное квадратичное отклонение.

$$S = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Дисперсию определяют как квадрат стандартного квадратичного отклонения S^2 -дисперсия

Показатели вариации: дисперсия, стандартное квадратичное отклонение.

Дисперсия

$$D[X] = (X - M[X])^2$$

$M[X]$ - Математическое ожидание

Среднеквадратическое отклонение определяется как квадратный корень из дисперсия случайной величины:

$$\sigma = \sqrt{D[X]}$$

Показатели вариации: r^2

Если использовать выборочную оценку значений соответствующих дисперсий, то получим формулу для выборочного коэффициента детерминации (который обычно и подразумевается под коэффициентом детерминации):

$$R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} = 1 - \frac{SS_{res}/n}{SS_{tot}/n} = 1 - \frac{SS_{res}}{SS_{tot}},$$

где SS_{res} - сумма квадратов остатков регрессии:

$$SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

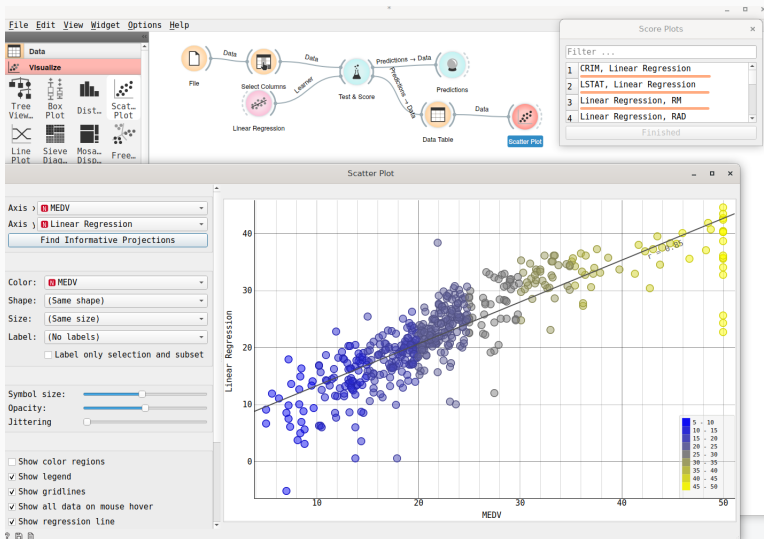
y_i, \hat{y}_i — фактические и расчётные значения объясняемой переменной.

Число, характеризующее выборку (например, набор чисел). Если все элементы выборки различны, то медиана — это такое число выборки, что ровно половина из элементов выборки больше него, а другая половина меньше него. В более общем случае медиану можно найти, упорядочив элементы выборки по возрастанию или убыванию и взяв средний элемент. Например, выборка 11,9,3,5,5 после упорядочивания превращается в 3,5,5,9,11 и её медианой является число 5. Если в выборке чётное число элементов, медиана может быть не определена однозначно: для числовых данных чаще всего используют полусумму двух соседних значений (то есть медиану набора 1,3,5,7 принимают равной 4),

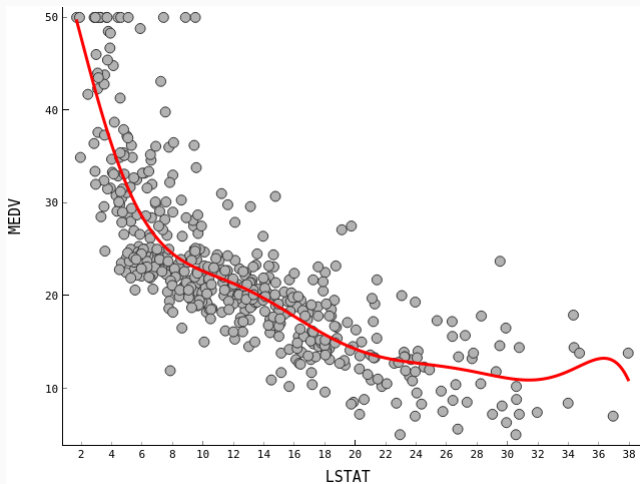
t-критерий для двух независимых выборок (двухвыборочный t-критерий) проверяет гипотезу о равенстве средних в двух выборках (предполагается нормальность распределения переменных, а также равенство дисперсий выборок). Критерий применяется, например, если необходимо сравнить результаты баллов ЕГЭ в двух разных школах.

P-value, p-уровень значимости, p-критерий — вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики (среднего арифметического, медианы и др.), по сравнению с ранее наблюдаемым, при условии, что нулевая гипотеза верна.

Регрессия



Полиномиальная регрессия



Построено по таблице housing.tab, входит в базовый набор таблиц Orange3.

Понижение размерности.

Отбор признаков. Выявление избыточных признаков с помощью фильтров. Выделение признаков. Анализ главных компонент. Ограничения PCA. Применение оберток для задания модели вопросов о признаках. Другие методы обработки признаков.

Многомерное шкалирование.

Отбор признаков, известный также как отбор переменных, отбор атрибутов или отбор поднабора переменных, это процесс отбора подмножества значимых признаков (переменных зависимых и независимых) для использования в построении модели. Техники отбора признаков используются по четырём причинам:

- упрощение моделей для того, чтобы сделать их проще для интерпретации исследователями/пользователями,
- более короткое время тренировки,
- чтобы избежать проклятие размерности,
- улучшенное обобщение путём сокращения переобучения.

Центральный посыл использования техники отбора признаков — что данные содержат некоторые признаки, которые либо излишни, либо не значимы, а потому могут быть удалены без существенной потери информации. Излишни и не значимы являются двумя различными понятиями, поскольку один значимый признак может быть излишним при присутствии другого существенного признака, с которым он сильно коррелирует

Снижение размерности. Важные моменты

- Переменные, значения которых можно измерять в эксперименте, имеют для исследуемого объекта или явления нередко достаточно условный характер, лишь опосредовано отражая его внутреннюю структуру, движущие силы (механизмы) или действующие на объект факторы.
- Когда неизвестный фактор проявляется в изменении нескольких переменных, в процессе анализа можно наблюдать существенную корреляцию или связь между этими переменными. Тем самым глубинных (скрытых, латентных) факторов может быть существенно меньше, чем измеряемых переменных, даже само число которых выбирается исследователем достаточно субъективно.

- Степень влияния фактора на некоторый показатель (переменную) проявляется в величине дисперсии (в разбросе или диапазоне изменения значений) этого показателя при изменении значений фактора.
Действительно, изменение более мощного фактора будет приводить к большим изменениям показателя, на который он оказывает влияние.

- Если расположить оси исходных переменных ортогонально друг к другу (что также является определенным субъективным произволом, поскольку многие переменные связаны, коррелированы между собой, т. е. не могут считаться независимыми или ортогональными друг к другу), то можно обнаружить, что в этом (многомерном) пространстве объекты группируются (в соответствии со своими координатами) в виде некоторого облака или эллипсоида рассеяния, более вытянутого в одних направлениях и почти плоского в других.

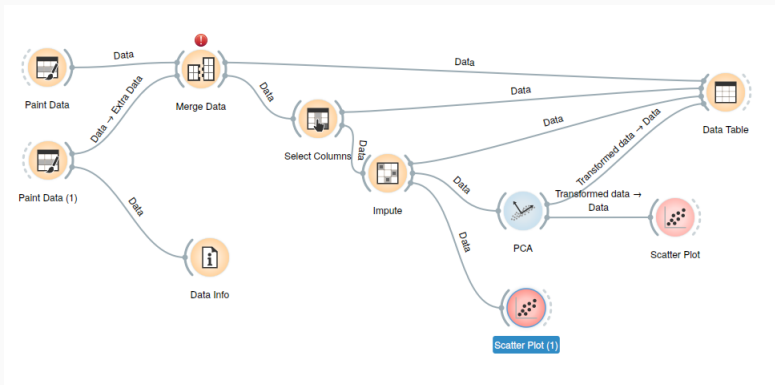
Снижение размерности. Важные моменты

- Если теперь провести новые оси координат соответственно осям такого эллипса рассеяния, то можно говорить о выделении так называемых факторов, более важных и существенных для изучаемого феномена по сравнению с исходными переменными, и оценивать сравнительную значимость этих факторов в смысле их влияния на дисперсию объектов. При этом обычно оказывается, что толщина такого облака рассеяния по некоторым осям настолько мала, что эти оси (малозначимые факторы) можно в дальнейшем вовсе исключить из рассмотрения. Это означает, что нам удалось снизить размерность исходного пространства переменных.

Как работает метод главных компонент

Для выбора первой оси главной компоненты нужно для каждой прямой посчитать сумму квадратов расстояний от всех точек до этой прямой и выбрать среди них ту, для которой эта сумма квадратов будет минимальной. Это похоже на поиск регрессионной прямой: разница состоит в том, что когда мы искали регрессионную прямую, мы считали расстояние «по вертикали», а сейчас считаем обычное расстояние от точки до прямой (вдоль направления, вертикального к прямой). Конечно, всех возможных прямых очень много, но задача отыскания самой лучшей из них оказывается довольно простой: компьютеры с ней справляются очень эффективно даже для пространств высокой размерности.

Снижение размерности.Пример



Методы фильтров выбирают переменные независимо от модели. Они базируются только на общих признаках, таких как корреляция переменной с предсказанием. Методы фильтров подавляют наименее интересные переменные. Другие переменные будут частью классификации или модели регрессии, использованной для классификации или предсказания. Эти методы очень эффективны по времени вычисления и устойчивы к переобучению.

Метод фильтров

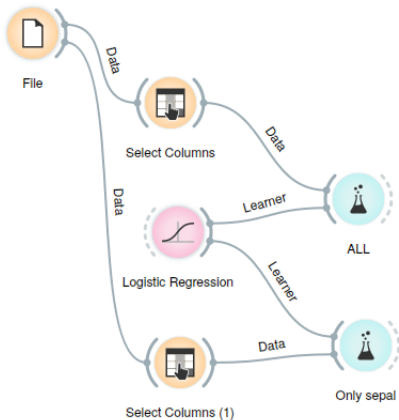


Correlations			
Spearman correlation			
MEDV			
Filter ...			
1	-0.853	LSTAT	MEDV
2	+0.634	MEDV	RM
3	-0.578	INDUS	MEDV
4	-0.563	MEDV	NOX
5	-0.562	MEDV	TAX
6	-0.559	CRIM	MEDV
7	-0.556	MEDV	PTRATIO
8	-0.548	AGE	MEDV
9	+0.446	DIS	MEDV
10	+0.438	MEDV	ZN
11	-0.347	MEDV	RAD

Применение оберток для задания модели вопросов о признаках

Методы обёртывания используют модель априорной оценки результата для оценки поднаборов признаков. Каждый новый поднабор используется для тренировки модели, которая проверяется на контрольной выборке. На этой контрольной выборке считается число ошибок (показатель ошибок модели), которое даёт оценку для данного подмножества. Так как методы обёртывания тренируют модель для каждого поднабора, они вычислительно очень затратны, но дают, как правило, лучший набор признаков для конкретного типа модели.

Метод оберток



ALL

Evaluation Results

Method	AUC	CA	F1	Precision
Logistic Regression	0.981	0.927	0.926	0.935

?

Only sepal

Evaluation Results

Method	AUC	CA	F1	Precision
Logistic Regression	0.894	0.773	0.752	0.810

?

Метод вложения для отбора признаков Методы вложения были предложены как попытка комбинации преимуществ двух предыдущих методов. Обучающий алгоритм имеет преимущество собственного процесса выбора переменной и осуществляет выбор признаков и классификацию одновременно. Метод вложения реализуется программно.

Модель алгоритмов машинного обучения.

Задано множество объектов X , множество допустимых ответов Y . Существуют пары $\langle \text{объект}, \text{ответ} \rangle$, которые будем называть прецедентами. Совокупность называется обучающей выборкой (training sample).

Задача обучения по прецедентам заключается в том, чтобы по обучающей выборке восстановить зависимость, то есть построить решающую функцию или алгоритм допускающий эффективную компьютерную реализацию.

В задачах обучения по прецедентам пара <объект,ответ> это не реальные объекты, а лишь доступные данные о них. Данные могут быть неточными, поскольку измерения свойств объектов и ответов обычно выполняются с погрешностями.

При это данные могут быть неполными, поскольку измеряются не все мыслимые признаки, а лишь физически доступные для измерения. В результате одному и тому же описанию x могут соответствовать различные объекты и различные ответы.

Модель алгоритмов, метод обучения.

В таком случае алгоритм или функция описывающие зависимость не является функцией. Устранить эту некорректность позволяет вероятностная постановка задачи. Вместо существования неизвестной целевой зависимости можно считать что существует неизвестное вероятностного распределения на множестве пар объектов и ответов с плотностью $p(x, y)$, из которого случайно и независимо выбираются некоторое количество наблюдений. Такие выборки называются простыми или случайными одинаково распределёнными. Вероятностная постановка задачи считается более общей, так как функциональную зависимость можно представить в виде вероятностного распределения $p(x, y) = p(x)p(y|x)$, положив $p(y|x) = \delta(y - y(x))$, где $\delta(z)$ — дельта-функция.

Основные виды классификаторов машинного обучения.

Методы машинного обучения можно разделить на 3 основные категории: контролируемое, неконтролируемое и подкрепляемое обучение. Контролируемое или обучение с учителем уже обсуждалось выше.

Основные виды классификаторов машинного обучения.

Контролируемое обучение полезно в тех случаях, когда свойство (ярлык) доступно для определенного массива данных (обучающего набора), но на данный момент оно отсутствует и должно быть предсказано для других случаев.

Основные виды классификаторов машинного обучения.

Неконтролируемое обучение используется для обнаружения неявных отношений в данном немаркированном наборе данных.

Формально: Пусть X - множество объектов - описаний некоторых объектов. Необходимо найти множество взаимосвязей этих объектов. Качество выявления взаимосвязей проверяется некоторой метрикой, выбранной исходя из решаемой задачи.

Основные виды классификаторов машинного обучения.

Обучение без учителя используется для решения следующих типов задач: 1. Задача кластеризации. 2. Поиск правил ассоциации. 3. Сокращение размерности данных. 4. Визуализация данных.

Основные виды классификаторов машинного обучения.

Под задачей поиска правил ассоциаций подразумевается выявление в признаковых описаниях объектов (исходных данных) таких наборов и значений признаков, которые особенно часто (неслучайно часто) встречаются в исходных данных. Если же проводить аналогию с первой задачей, то каждое правило в данном случае может быть представлено как кластер.

Основные виды классификаторов машинного обучения.

Задача сокращения размерности данных состоит в следующем. Существует большой (значительно большой) объем признаковых описаний объектов. Причем, этот объем обуславливается внушительным количеством измерений признакового пространства. Необходимо представить те же данные в пространстве меньшей размерности, при этом минимизировав потери информации. Группировка по кластерам как раз и будет одним из вариантов решения проблемы.

Основные виды классификаторов машинного обучения.

Задача визуализации данных является по сути частным случаем предыдущей: ее суть – представить исходные данные в отображаемом пространстве, то есть пространстве размерности 2 или 3. Как следует из описанного выше, обучение без учителя в какой-то мере так или иначе сводится к кластеризации. Поэтому для оценки качества обучения данным способом как правило используют метрики качества кластеризации. Причем, при их выборе учитывается, что эти метрики не должны зависеть от исходных

Функция потерь и функционал качества.

Функция потерь (loss function) — это неотрицательная функция $L(a, x)$, характеризующая величину ошибки алгоритма a на объекте x . Если $L(a, x) = 0$, то ответ $a(x)$ называется корректным.

Функция потерь и функционал качества.

$L(a, x) = [a(x) \neq y(x)]$ - индикатор несовпадения с правильным ответом (обычно применяется в задачах классификации);

$L(a, x) = |a(x) - y(x)| > \epsilon$ - индикатор существенного отклонения от правильного ответа, где ϵ - заданный порог точности.

Функция потерь и функционал качества.

Вообще, если L принимает только два значения 0 и 1, причём 1 соответствует ошибке, то функционал Q называется частотой ошибок алгоритма a на выборке X .

$L(a, x) = |a(x) - y(x)|$ – величина отклонения от правильного ответа; функционал Q называется средней ошибкой алгоритма a на выборке X ;

$L(a, x) = (a(x) - y(x))^2$ – квадрат отклонения от правильного ответа; функционал Q называется средней квадратичной ошибкой алгоритма a на выборке X .

$Lw(a, x) = w(x)L(a, x)$ – взвешенная функция потерь, где $w(x)$ – неотрицательная весовая функция, характеризующая степень важности объекта x или величину потери от ошибки на данном объекте; $L(a, x)$ – некоторая функция потерь, например, любая из перечисленных выше.

Принцип минимизации эмпирического риска.

Эмпирическим риском называется средняя ошибка алгоритма на обучающей выборке. Метод минимизации эмпирического риска (empirical risk minimization, ERM) наиболее часто применяется для построения алгоритмов обучения. Он состоит в том, чтобы в рамках заданной модели выбрать алгоритм, имеющий минимальное значение средней ошибки на заданной обучающей выборке.

Принцип минимизации эмпирического риска.

С переобучением метода ERM связано два утверждения, которые на первый взгляд могут показаться парадоксальными.

Утверждение 1. Минимизация эмпирического риска не гарантирует, что вероятность ошибки на тестовых данных будет мала. Легко строится контрпример — абсурдный алгоритм обучения, который минимизирует эмпирический риск до нуля, но при этом абсолютно не способен обучаться.

Принцип минимизации эмпирического риска.

Алгоритм состоит в следующем. Получив обучающую выборку, он запоминает её и строит функцию, которая сравнивает предъявляемый объект с запомненными обучающими объектами. Если предъявляемый объект в точности совпадает с одним из обучающих, то эта функция выдаёт для него запомненный правильный ответ. Иначе выдаётся произвольный ответ (например, случайный или всегда один и тот же). Эмпирический риск алгоритма равен нулю, однако он не восстанавливает зависимость и не обладает никакой способностью к обобщению.

Вывод: для успешного обучения необходимо не только запоминать, но и обобщать.

Принцип минимизации эмпирического риска.

Переобучение появляется именно вследствие минимизации эмпирического риска. Пусть задано конечное множество из D алгоритмов, которые допускают ошибки независимо и с одинаковой вероятностью. Число ошибок любого из этих алгоритмов на заданной обучающей выборке подчиняется одному и тому же биномиальному распределению. Минимум эмпирического риска — это случайная величина, равная минимуму из D независимых одинаково распределённых биномиальных случайных величин. Её ожидаемое значение уменьшается с ростом D . Соответственно, с ростом D увеличивается переобученность — разность вероятности ошибки и частоты ошибок на обучении.

Принцип минимизации эмпирического риска.

В данном модельном примере легко построить доверительный интервал переобученности, так как функция распределения минимума известна. Однако в реальной ситуации алгоритмы имеют различные вероятности ошибок, не являются независимыми, а множество алгоритмов, из которого выбирается лучший, может быть бесконечным. По этим причинам вывод количественных оценок переобученности является сложной задачей, которой занимается теория вычислительного обучения. До сих пор остаётся открытой проблема сильной завышенности верхних оценок вероятности переобучения.

Утверждение 3. Переобучение связано с избыточной сложностью используемой модели. Всегда существует оптимальное значение сложности модели, при котором переобучение минимально.

Обобщающая способность (generalization ability, generalization performance). Говорят, что алгоритм обучения обладает способностью к обобщению, если вероятность ошибки на тестовой выборке достаточно мала или хотя бы предсказуема, то есть не сильно отличается от ошибки на обучающей выборке. Обобщающая способность тесно связана с понятиями переобучения и недообучения.

Переобучение, переподгонка (overtraining, overfitting) – нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложных моделей.

Недообучение – нежелательное явление, возникающее при решении задач обучения по прецедентам, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей.

Скользящий контроль или кросс-проверка или кросс-валидация (cross-validation, CV) — процедура эмпирического оценивания обобщающей способности алгоритмов, обучаемых по прецедентам.

Фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке, затем оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках.

Если выборка независима, то средняя ошибка скользящего контроля даёт несмещённую оценку вероятности ошибки. Это выгодно отличает её от средней ошибки на обучающей выборке, которая может оказаться смещённой (оптимистически заниженной) оценкой вероятности ошибки, что связано с явлением переобучения.

Скольльзящий контроль является стандартной методикой тестирования и сравнения алгоритмов классификации, регрессии и прогнозирования.

Логистическая регрессия представляет собой мощный статистический способ прогнозирования вероятности возникновения некоторого события с одной или несколькими независимыми переменными. Логистическая регрессия определяет степень зависимости между категориальной зависимой и одной или несколькими независимыми переменными путем использования логистической функции, являющейся аккумулятивным логистическим распределением.

Линейная регрессионная модель не всегда способна качественно предсказывать значения зависимой переменной. Выбирая для построения модели линейное уравнение, мы естественным образом не накладываем никаких ограничений на значения зависимой переменной. А такие ограничения могут быть существенными.

Например, при проектировании оптимальной длины шахты лифта в новом здании необходимо учесть, что эта длина не может превышать высоту здания вообще.

Линейная регрессионная модель может дать результаты, несовместимые с реальностью. С целью решения данных проблем полезно изменить вид уравнения регрессии и подстроить его для решения конкретной задачи.

Вообще, логит регрессионная модель предназначена для решения задач предсказания значения непрерывной зависимой переменной, при условии, что эта зависимая переменная может принимать значения на интервале от 0 до 1.

Сигмоид — это гладкая монотонная возрастающая нелинейная функция, имеющая форму буквы «S», которая часто применяется для «приближения» значений некоторой величины к нулю и единице

$$y = a_0 + a_1x_1 + a_2x_2 \dots + a_nx_n$$

$$f(y) = \frac{1}{1 + e^{-y}}$$

Данный алгоритм может быть использован для:

- оценки компартмента предприятия с использованием косвенных данных связи с другими предприятиями в прошлом объемы по типам продукции, регион и т.п.;
- измерении показателей успешности тех или иных мер;
- предсказании доходов с определенного продукта;

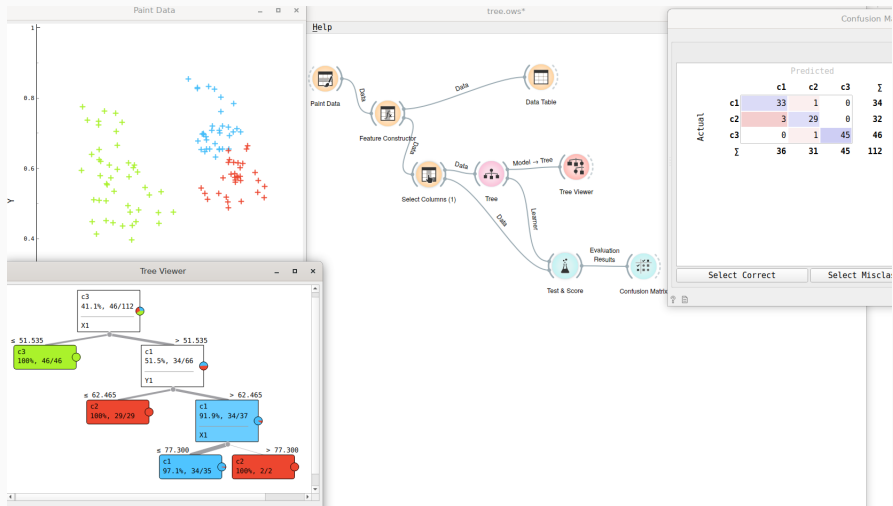
Дерево принятия решений — средство поддержки принятия решений, которое использует древовидный граф или модель принятия решений, а также возможные последствия их работы, включая вероятность наступления события, затраты ресурсов и полезность.

Деревья принятия решений и случайные леса

Один из самых распространенных алгоритмов машинного обучения. Используется в статистике и анализе данных для прогнозных моделей. Структура представляет собой "листья" и "ветки". На "ветках" дерева решения записаны атрибуты, от которых зависит целевая функция, в "листьях" записаны значения целевой функции, а в остальных узлах – атрибуты, по которым различаются случаи.

Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких входных переменных. Осуществляется перебор столбцов.

Дерево принятия решения. Пример



Дерево принятия решения. Пояснения

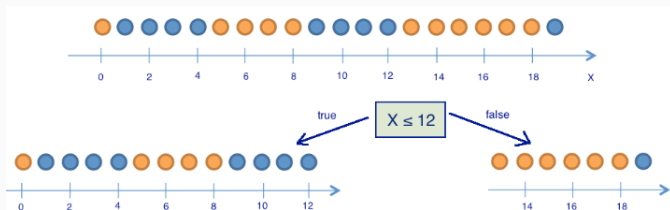


Если мы наудачу вытащили шарик, то он с вероятностью $p_1 = \frac{9}{20}$ будет синим и с вероятностью $p_2 = \frac{11}{20}$ – желтым. Значит, энтропия состояния $S_0 = -\frac{9}{20}\log_2 \frac{9}{20} - \frac{11}{20}\log_2 \frac{11}{20} \approx 1$

Само это значение пока ни о чем нам не говорит. Как изменится энтропия, если разбить шарики на две группы – с координатой меньше либо равной 12 и больше 12?

<https://www.pvsm.ru/python/249240>

Дерево принятия решения. Пояснения



В левой группе оказалось 13 шаров, из которых 8 синих и 5 желтых. Энтропия этой группы равна $S_1 = -\frac{5}{13}\log_2 \frac{5}{13} - \frac{8}{13}\log_2 \frac{8}{13} \approx 0.96$. В правой группе оказалось 7 шаров, из которых 1 синий и 6 желтых. Энтропия правой группы равна

$$S_2 = -\frac{1}{7}\log_2 \frac{1}{7} - \frac{6}{7}\log_2 \frac{6}{7} \approx 0.6$$

<https://www.pvsm.ru/python/249240>

Метод опорных векторов (SVM)

Метод опорных векторов (SVM) — это набор алгоритмов, использующихся для задач классификации и регрессионного анализа. Учитывая, что в N -мерном пространстве каждый объект принадлежит одному из двух классов, SVM генерирует $(N-1)$ -мерную гиперплоскость с целью разделения этих точек на 2 группы. Это как если бы вы на бумаге изобразили точки двух разных типов, которые можно линейно разделить. Помимо того, что метод выполняет сепарацию объектов, SVM подбирает гиперплоскость так, чтобы та характеризовалась максимальным удалением от ближайшего элемента каждой из групп.

SVM. Пример



Метод ансамблей основан на обучающих алгоритмах, которые формируют множество классификаторов, а затем сегментируют новые точки данных, отталкиваясь от голосования или усреднения. Оригинальный метод ансамблей — не что иное, как Байесовское усреднение, но более поздние алгоритмы включают исправления ошибок выходного кодирования, бэггинг (bagging) и бустинг (boosting).

Бустинг направлен на превращение слабых моделей в сильные путем построения ансамбля классификаторов. Бэггинг также агрегирует усовершенствованные классификаторы, но используется при этом параллельное обучение базовых классификаторов. Бэггинг — улучшающее объединение, а бустинг — улучшающее пересечение.

Метод случайного леса (Random Forest) представляет собой дальнейшее улучшение бэггинга деревьев решений, которое заключается в устранении корреляции между деревьями. Как и в случае с бэггингом, мы строим несколько сотен деревьев решений по обучающим бутстреп-выборкам. Однако на каждой итерации построения дерева случайным образом выбирается m из p подлежащих рассмотрению предикторов и разбиение разрешается выполнять только по одной из m этих переменных.

Смысл этой процедуры, оказавшейся весьма эффективной для повышения качества получаемых решений, заключается в том, что с вероятностью $(p-m)/p$ блокируется какой-нибудь потенциально доминирующий предиктор, стремящийся войти в каждое дерево. Если доминирование таких предикторов разрешить, то все деревья в итоге будут очень похожи друг на друга, а получаемые на их основе предсказания будут сильно коррелировать и снижение дисперсии будет не столь очевидным. Благодаря блокированию доминантов, другие предикторы получают свой шанс, и вариация деревьев возрастает.

Выбор малого значения m при построении случайного леса обычно будет полезным при наличии большого числа коррелирующих предикторов. Естественно, если случайный лес строится с использованием $m=p$, то вся процедура сводится к простому бэггингу.

Часто результат работы алгоритма на фиксированной тестовой выборке визуализируют с помощью ROC-кривой (ROC = receiver operating characteristic, иногда говорят «кривая ошибок»), а качество оценивают как площадь под этой кривой – AUC (AUC = area under the curve). [https : //dyakonov.org/](https://dyakonov.org/) и далее

Табл. 1: Таблица для построения ROC-кривой

Оценка логистической регрессии	класс
0.6	1
0.5	1
0.4	1
0.2	0
0.2	1
0.1	0

Чтобы нарисовать ROC-кривую нужно координатную плоскость, разбить на m равных частей горизонтальными линиями и на n – вертикальными, где m – число 1 среди правильных меток теста (в нашем примере $m=3$), n – число нулей ($n=4$). В результате квадрат разбивается сеткой на $m \times n$ блоков.

Теперь будем просматривать строки таблицы сверху вниз и прорисовывать на сетке линии, переходя их одного узла в другой. Стартуем из точки $(0, 0)$. Если значение метки класса в просматриваемой строке 1, то делаем шаг вверх; если 0, то делаем шаг вправо. Ясно, что в итоге мы попадём в точку $(1, 1)$, т.к. сделаем в сумме m шагов вверх и n шагов вправо.

Если у нескольких объектов значения оценок равны, то мы делаем шаг в точку, которая на a блоков выше и b блоков правее(по диагонали), в нашем случае 4 и 5 строки имеют один и тот же результат логистической регрессии, поэтому будет сдвиг по диагонали В частности, если все объекты имеют одинаковую метку, то мы сразу шагаем из точки $(0, 0)$ в точку $(1, 1)$.