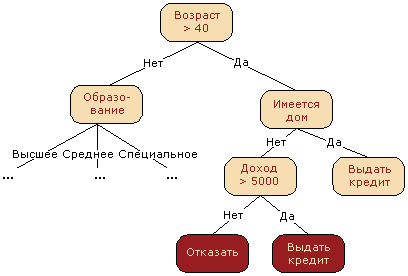
**Деревья принятия решений (ДПР, decision tree) или решающие деревья**

**1.**

В модели деревьев принятия решений используется древовидная структура данных, которая предоставляет несколько возможных путей принятия решения и исход для каждого пути.

Зачастую дерево решений служит обобщением опыта экспертов, средством передачи знаний будущим сотрудникам или моделью бизнес-процесса компании.



Модели на основе деревьев принятия решений обладают многими преимуществами:

* Они понятны и легко интерпретируемы, а процесс, на основе которого они достигают прогноза, абсолютно прозрачен
* Модели на основе деревьев принятия решений легко справляются с сочетанием численных (количество ног) и категорийных (вкусный/не вкусный) признаков и даже могут классифицировать данные, для которых атрибуты отсутствуют

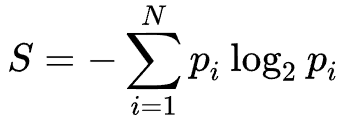
Вместе нахождение "оптимального" дерева принятия решений для набора обучающих данных - задача вычислительно очень сложная. Очень легко (и это очень плохо) построить деревья принятия решений, которые избыточно подогнаны под обучающую выборку и которые недостаточно хорошо обобщают не встречавшиеся ранее данные. Мы обратимся к способам решения этих проблем.

Обычно деревья принятия решений подразделяются на **классификационные** деревья (которые генерируют категорийные значения) и **регрессионные** деревья (которые генерируют численные или непрерывные значения).

Для того чтобы построить дерево принятия решений, необходимо решить, какие вопросы задавать и в каком порядке. На каждом уровне дерева имеются некоторые возможности, которые были исключены и которые еще остались. Получив информацию о том, что у животного не более пяти ног, исключают возможность, что это кузнечик, но оставлен вариант с уткой. Каждый возможный вопрос разделяет оставшиеся возможные варианты в соответствии с полученным ответом. В идеале хотелось бы подобрать вопросы, ответы на которые дают максимум информации о предмете предсказания ("да/нет").

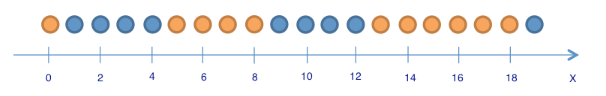
Идея о "количестве информации" объясняется термином энтропия. Это слово часто приводят для обозначения беспорядка, хаоса. Здесь оно используется для обозначения неопределенности или непредсказуемости, связанной с данными.

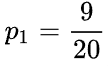
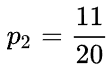
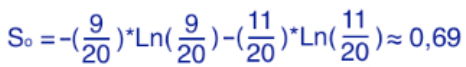
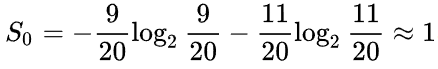
Пусть имеется множество S данных, каждый маркированный элемент которого принадлежит одному из классов конечного множества классов CI' . . . ,Сп. Если все точки данных принадлежат только одному классу, то неопределенность фактически отсутствует, вследствие чего требуется низкая энтропия. Если же точки данных распределены по классам равномерно, то существует большая неопределенность, и значит, требуется высокая энтропия.



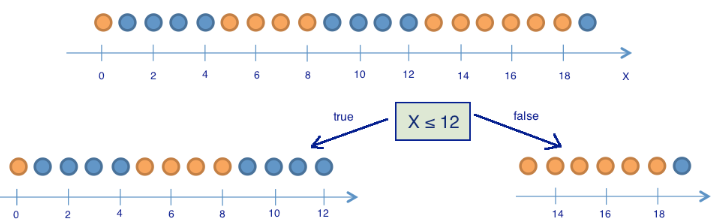
где pi – вероятности нахождения системы в i-ом состоянии.

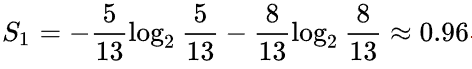
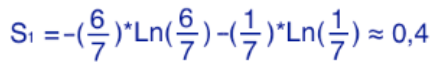
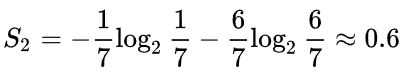
Каждый шаг дерева принятия решений предполагает постановку вопроса, ответ на который расщепляет данные на один или более подмножеств. Например, вопрос "у него больше пяти лап?" разделяет животных на тех, у которых их больше пяти (пауки) и остальных (ехидна).



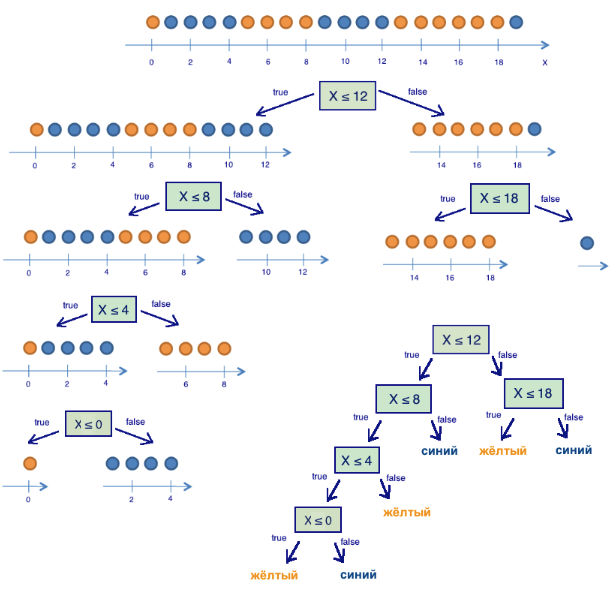
Здесь 9 синих шариков и 11 желтых. Если мы наудачу вытащили шарик, то он с вероятностью  будет синим и с вероятностью  – желтым. Значит, энтропия состояния .

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



В левой группе оказалось 13 шаров, из которых 8 синих и 5 желтых. Энтропия этой группы равна  . В правой группе оказалось 7 шаров, из которых 1 синий и 6 желтых. Энтропия правой группы равна . Как видим, энтропия уменьшилась в обеих группах по сравнению с начальным состоянием, хоть в левой и не сильно. Поскольку энтропия – по сути степень хаоса (или неопределенности) в системе, уменьшение энтропии называют приростом информации.

Получается, разделив шарики на две группы по признаку "координата меньше либо равна 12", мы уже получили более упорядоченную систему, чем в начале. Продолжим деление шариков на группы до тех пор, пока в каждой группе шарики не будут одного цвета.



Для правой группы потребовалось всего одно дополнительное разбиение по признаку "координата меньше либо равна 18", для левой – еще три. Очевидно, энтропия группы с шариками одного цвета равна 0 (), что соответствует представлению, что группа шариков одного цвета – упорядоченная.

В итоге мы построили дерево решений, предсказывающее цвет шарика по его координате. Отметим, что такое дерево решений может плохо работать для новых объектов (определения цвета новых шариков), поскольку оно идеально подстроилось под обучающую выборку (изначальные 20 шариков). Для классификации новых шариков лучше подойдет дерево с меньшим числом "вопросов", или разделений, пусть даже оно и не идеально разбивает по цветам обучающую выборку.

Алгоритм обучения: *нужно находить правила (предикаты), на основе которых разбивать тренировочный набор данных, таким образом, чтобы уменьшалось среднее значение энтропии*. Процесс деления множества данных на части, приводящий к уменьшению энтропии, можно рассматривать как *производство информации*.

**Случайные леса**

Поскольку деревья принятия решений могут приспосабливаться к обучающим данным, неудивительно, что они имеют тенденцию к переобучению. Один из способов, который помогает избежать этого - статистический метод случайных лесов (random forests). Данный метод подразумевает создание нескольких ДПР, которые затем принимают коллективное решение на основе голосования о том, каким образом классифицировать входящие значения.

Для построения случайного леса к данным применяются бутстрап-методы. Вместо того чтобы обучать деревья на всех элементах inputs обучающей выборки, каждое дерево обучают на результате, получаемом после вызова bootstrap\_sample(inputs). Поскольку теперь каждое дерево строится на разных данных, деревья будут друг от друга отличаться. (Дополнительное преимущество заключается в том, что для проверки каждого дерева совершенно справедливо можно использовать невыборочные данные, а значит, в качестве обучающей выборки можно смело воспользоваться всеми данными целиком, если разумно подойти к измерению производительности.) Этот метод известен как бутстрап-агрегирование или бэггинг.

Чтобы не заморачиваться над критерием остановки при построении дерева, можно поступить следующим образом: *выбирать случайные подмножества из обучающей выборки данных, и для каждого подмножества строить своё дерево принятия решений (в принципе, даже не важно какой критерий остановки будет использоваться)*:



Можно заметить, что единичное дерево принятия решений описывает область, которая полностью содержит красные точки, в то время как ансамбль деревьев описывает фигуру, которая более близка к окружности.

Модели на основе случайных лесов являются одними из самых популярных и универсальных.

**Вместо заключения**

Деревья принятия являются неплохой альтернативой, в тех случаях когда надоедает подстраивать абстрактные веса и коэффициенты в других алгоритмах классификации, либо, когда приходится обрабатывать данные со смешанными (категориальными и числовыми) атрибутами.