PART#1.DECISION TREES

by googlenerds

What for -- why to grow tree?

- Trees are much more interpretable than anything else
- The only practical way to handle real (messy) data
- No secret that Yandex Search ranks pages using boosted decision trees (GBDT)

- Wanna try to capture how people take decisions
- Have: objects with various attributes
- Problem:

- how to reveal objects' alikeness?

- how create a rule which MAX the probability the most alike objects are in 1 class?

formalization:

Задача восстановления зависимости $y\colon X \to Y$, $|Y| < \infty$
по точкам *обучающей выборки* (x_i, y_i), $i = 1, \dots, \ell$.

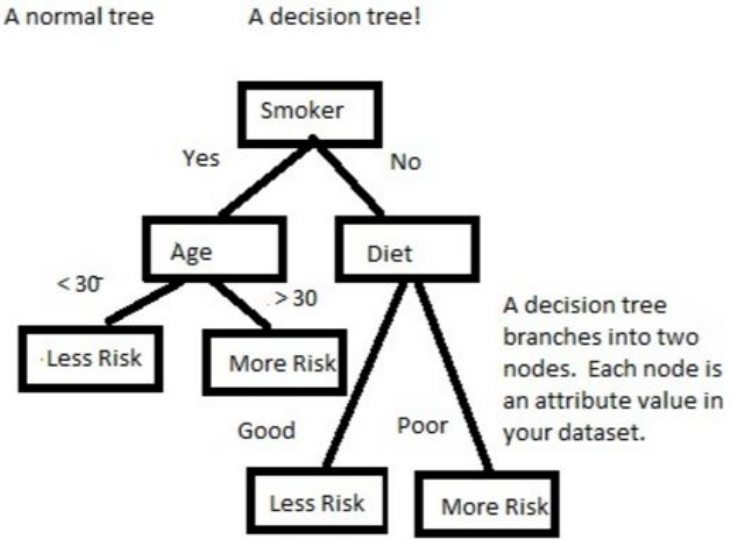
Дано: векторы $x_i = (x_i^1, \dots, x_i^n)$ — объекты обучающей выборки,
$y_i = y(x_i)$ — классификации, ответы учителя, $i = 1, \dots, \ell$:

$$\begin{pmatrix} x_1^1 & \dots & x_1^n \\ \dots & \dots & \dots \\ x_\ell^1 & \dots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Найти: функцию $a(x)$, способную классифицировать объекты
произвольной *тестовой выборки* $\tilde{x}_i = (\tilde{x}_i^1, \dots, \tilde{x}_i^n)$, $i = 1, \dots, k$:

$$\begin{pmatrix} \tilde{x}_1^1 & \dots & \tilde{x}_1^n \\ \dots & \dots & \dots \\ \tilde{x}_k^1 & \dots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \dots \\ a(\tilde{x}_k) \end{pmatrix}$$

Decision tree - what's it?



- Много-итерационная интуитивно-естественная процедура упорядочения элементов хаоса в групки по схожести признаков объектов

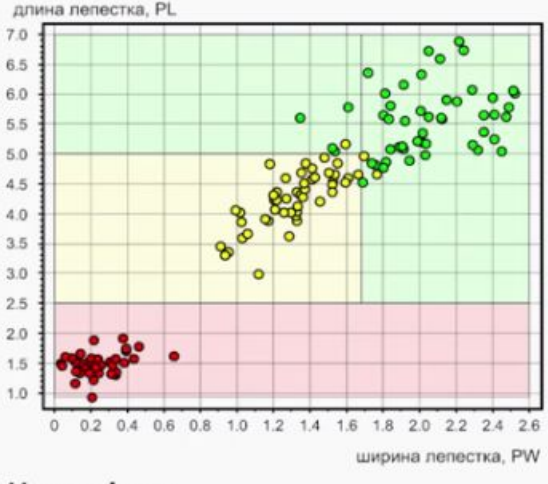
!Процедура поиска предикатов повторяется рекурсивно для каждого подмножества

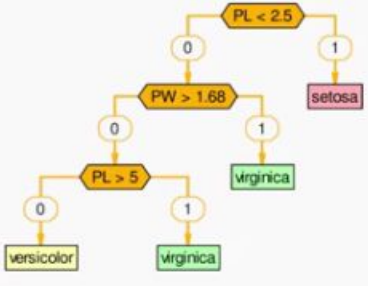
- Поиск ограничивающих объёмов в гиперпространстве: Если рассматривать классифицируемые объекты как точки в многомерном пространстве, то можно увидеть, что предикаты(правила), разделяющие множество данных на подмножества, являются гиперплоскостями, а процедура обучения классификатора является поиском ограничивающих объёмов (в общем, как и для любого другого вида

#1 Real tree example

Пример решающего дерева

Задача Фишера о классификации цветков ириса на 3 класса,
в выборке по 50 объектов каждого класса, 4 признака.





На графике:
в осях двух самых информативных признаков (из 4)
два класса разделились без ошибок, на третьем 3 ошибки.

#2 TOUCH THE TREE

Есть хаос. Как себе его представить? Пусть есть 10 шариков. И их можно по-всякому переставлять

Есть много вариантов перестановок этих шариков, а точнее:

N!/(Ngreen!*Nred!*Nyellow!) - multinomial coeff W.

Обзавём каждую перестановку от 0 до W-1

Строка из log2(W) бит однозначно кодирует каждую из перестановок - это инфа

Поскольку перестановка состоит из N шариков, то средн$\frac{\log_2(W)}{N}$ичество бит, приходящихся на один элемент перестановки можно ть как: Ещё это называется **комбинаторная энтропия**

Чем более однородно множество (преобладают шарики какого-то одного цвета) — тем меньше его комбинаторная энтропия, и наоборот — чем больше различных элементов в множестве, тем выше его энтропия

$$S = -\sum p_i \cdot \log_2 p_i$$

Это энтропия Шеннона, предел комбинаторной энтропии. Эту удобнее использовать в вычислениях. Выведено **тут**

ВЫВОД
нужно находить правила (предикаты) по которым бы уменьшалось среднее значение энтропии.

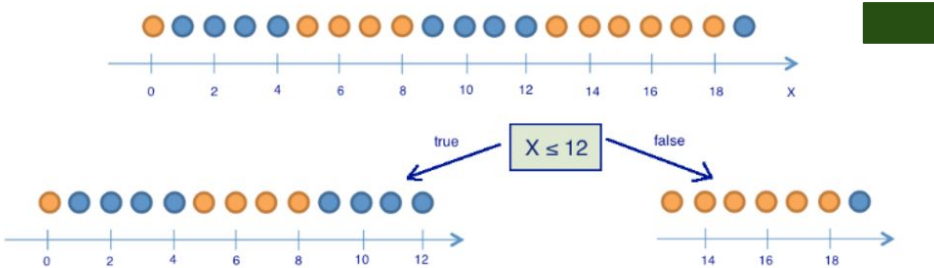
Процесс деления множества данных на части, приводящий к уменьшению энтропии, можно рассматривать как *производство информации*.

#2 TOUCH THE TREE

$$S = -\sum p_i \cdot \log_2 p_i$$

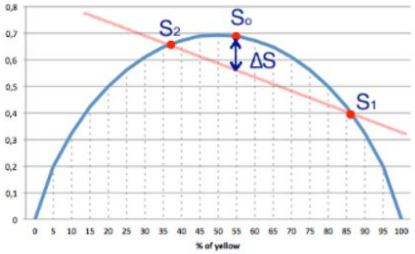
Для примера, рассмотрим множество двухцветных шариков, в котором цвет шарика зависит только от координаты *x*:
(из практических соображений, при расчётах удобно использовать энтропию Шеннона)

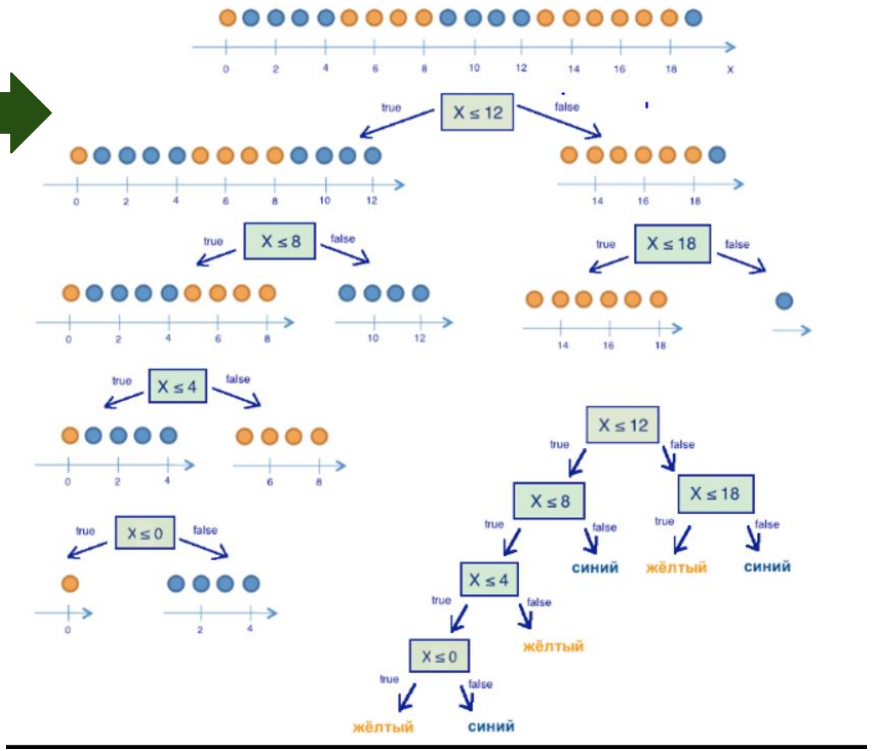
$$S_0 = -\left(\frac{9}{20}\right)*\mathrm{Ln}\left(\frac{9}{20}\right) - \left(\frac{11}{20}\right)*\mathrm{Ln}\left(\frac{11}{20}\right) \approx 0{,}69$$



$$S_2 = -\left(\frac{8}{13}\right)*\mathrm{Ln}\left(\frac{8}{13}\right) - \left(\frac{5}{13}\right)*\mathrm{Ln}\left(\frac{5}{13}\right) \approx 0{,}66$$

$$S_1 = -\left(\frac{6}{7}\right)*\mathrm{Ln}\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right)*\mathrm{Ln}\left(\frac{1}{7}\right) \approx 0{,}4$$





How to grow a tree hypothetically?

КАК НАЙТИ ПРЕДИКАТ?

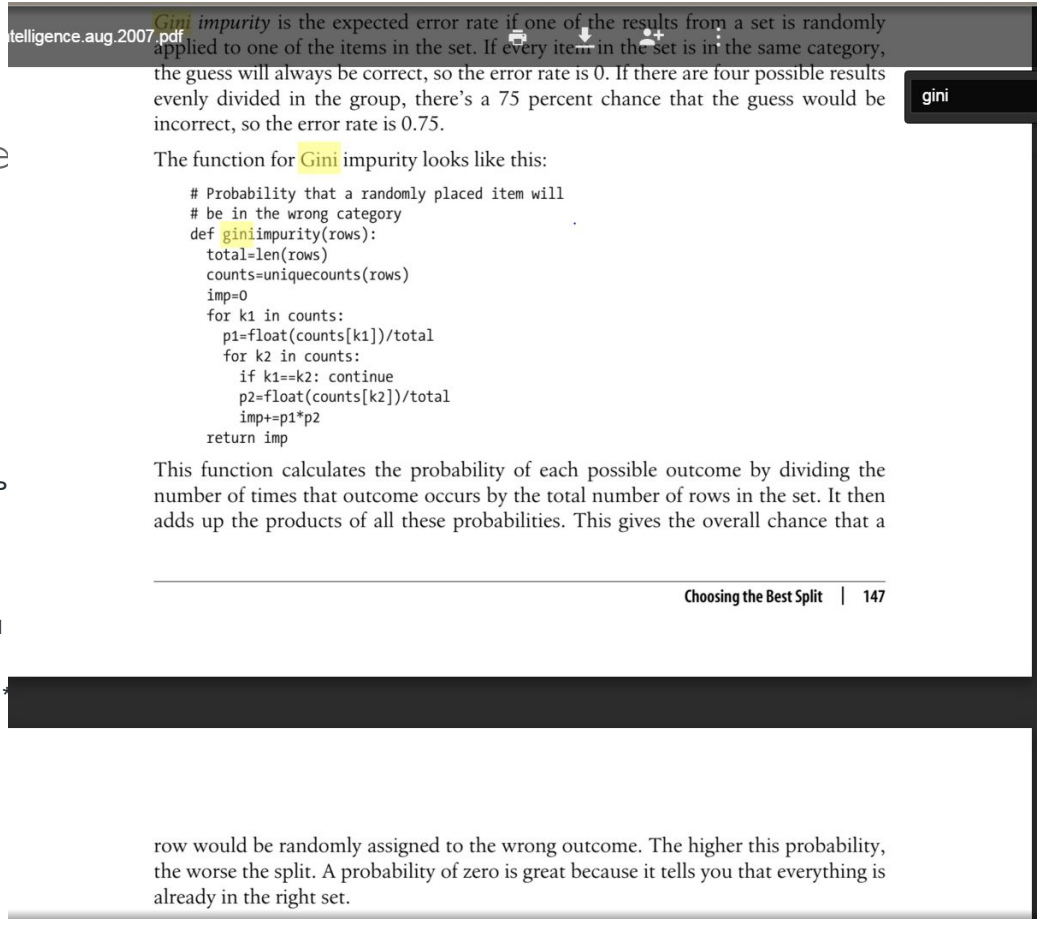
Разбив исходный набор данных на две части по некому предикату, можно рассчитать энтропию каждого подмножества, после чего рассчитать среднее значение энтропии — если оно окажется меньшим чем энтропия исходного множества, значит предикат содержит некую обобщающую информацию о данных.

```
s0 = вычисляем энтропию исходного множества

Если s0 == 0 значит:
   Все объекты исходного набора, принадлежат к одному классу
   Сохраняем этот класс в качестве листа дерева

Если s0 != 0 значит:
   Перебираем все элементы исходного множества:
      Для каждого элемента перебираем все его атрибуты:
         На основе каждого атрибута генерируем предикат, который разбивает исходное множество на два подмножества
         Рассчитываем лин комб энтропии
         Вычисляем ΔS
   Нас интересует предикат, с наибольшим значением ΔS
   Найденный предикат является частью дерева принятия решений, сохраняем его

   Разбиваем исходное множество на подмножества, согласно предикату
   Повторяем данную процедуру рекурсивно для каждого подмножества
```

Gini criteria

Количество объектов, которые лежат в 1 и том же классе, когда оъекты одного класса более кучкуются

перед словом Джини могут оказываться другие слова, каждый раз это значит что-то своё. критерий, думаю, означает, что речь идёт о разбиении в решающем дереве или случайных лесах. в каждой вершинке там нужно при построении дерева как-то выпустить две ветки вниз для детей. это можно сделать по-разному. обычно перебираются все возможные варианты и для каждого подсчитывается некоторая функция полезности. критерий Джини --- одна из возможных функций. \Sum_{k} p_{k} * (1-p_{k}), где p_k --- вероятность попадаения в k-ый класс при случайном выборе из всех объектов данной вершины.



Criteria donskogo

Reverse to Donskogo

Cross validation

Andrew ng vs yandex

Vocabulary

CART

Часто сходится на локальном решении (к примеру, на первом шаге была выбрана гиперплоскость, которая максимально делит пространство на этом шаге, но при этом это не приведёт к оптимальному решению)

Caution! Тревога! Дерево переобучилось!

Step1. Build all possible hyperplanes, which divide your surface on 2 parts

Step2. MIN the entropy: choose that case where in 1 part there s a max #of elements of 1 class -- you got 2 leaves

Step3. Take the worst (most chaotic) leaf and → Step1: instead of leaf you got a node with 2 leaves……

Тестовые выборки + кросс-валидация → проводимм обратный анализ(pruning)

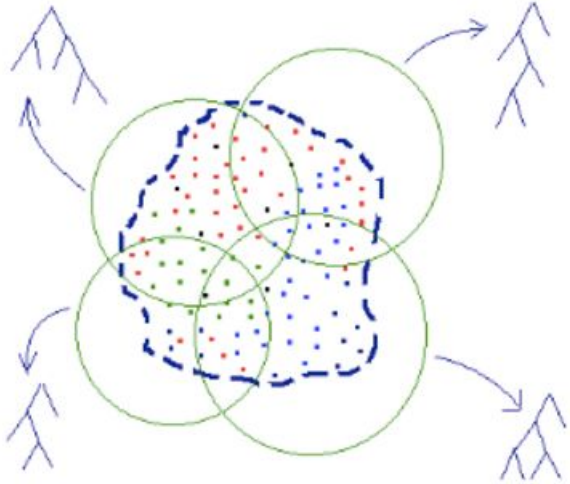
STOP WHEN: restrictions on # of nodes OR

Min error → 0

Random forest

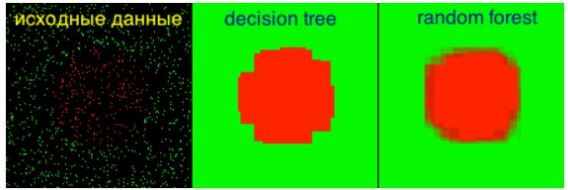
Можно сделать ансамбль деревьев, которые будут голосовать за каждый объект на тему его принадлежности к классу: берём несколько рандомных выборок, строим свои деревья - и потом суммируя голоса определяем принадлежность объекта к классу

Так что пограничная область будет довольно адекватной:



производится “bagging” — выборка случайных двух третей наблюдений для обучения, а оставшаяся треть **oob (out-of-bag) data** используется для оценки результата. Такую операцию проделывают сотни или тысячи раз. Результирующая модель будет будет результатом “голосования” набора полученных при моделировании деревьев

- \+ Высокое качество результата, особенно для данных с большим количеством переменных и малым количеством наблюдений.
- \+ Возможность распараллелить
- \+ Не требуется тестовая выборка
- \- Каждое из деревьев огромное, в результате модель получается огромная
- \- Долгое построение модели, для достижения хороших результатов.
- \- Сложная интерпретация модели (Сотни или тысячи больших деревьев сложны для интерпретации)



Stochastic gradient boosting

Step1. Grow “weak” tree with fixed amount of nodes.

Step2. Calculate the ▲ between:

the weak-tree prediction * learn rate (weakness coeff)

and reality

Yi+1=Yi-Yi*learnrate

continue

Learn ID3 (induction of decision tree)

Жадный алгоритм:

Решающие деревья ID3: достоинс

Достоинства:

- Интерпретируемость и простота классификации.
- Гибкость: можно варьировать множество $\mathscr{B}$.
- Допустимы разнотипные данные и данные с пропусками.
- Трудоёмкость линейна по длине выборки $O(|\mathscr{B}|h\ell)$.
- Не бывает отказов от классификации.

Недостатки:

- Жадный ID3 переусложняет структуру дерева, и, как следствие, сильно переобучается.
- Фрагментация выборки: чем дальше v от корня, тем меньше статистическая надёжность выбора β_v, c_v.
- Высокая чувствительность к шуму, к составу выборки, к критерию информативности.

Advantages and dis-

- \+ Деревья удобны, когда требуется не просто классифицировать данные, но ещё и объяснить ПОЧЕМУ тот или иной объект отнесён к такому то классу
- \+ когда надоедает подстраивать абстрактные веса и коэффициенты в других алгоритмах классификации, либо, когда приходится обрабатывать данные со смешанными (категориальными и числовыми) атрибутами.
-
- \- Переобучение. Одним из возможных критериев остановки может быть небольшое значение ΔS. Но при таком подходе, всё же, невозможно дать универсальный совет: при каких значениях ΔS следует прекращать построение дерева.
-
-

Tree can solve

Classification task

Regression task

Help neural network

???????KWESCHANS?????????

Q: Do we have trees with multiple branches from one node (more that 2 branches from one node?)

A: Classification or regression trees do not have to be binary, but most are. (ternary tree then)

Q:Как понять как останавливаться?
A:
1)Наперед заданная энтропия (?)
Доля объектов принадлежащих 1 классу >= const
2) норм количество узлов - больше не надо
3) с каждым новым листом мы не имеем прироста инфы > fix
4) не имеем прироста инфы (нет способа найти такой предикат)

Как понять outliers (Din-din effect):не нужно ли нам исключить вручную какой то объект в опр момент, чтобы потом быстрее и лучше классифицировать ?

Q: Как при непрерывном значении признака объекта найти способ, по которому определить пороговое значение для предиката (опр множество предикатов, из которых выбирать приводящий к минимиз энтропии) (10^6 диаметров шаров - такое дискр, что почти непр)

A: Среднеарифметическое гэпов между значениями атрибутов х обучающей выборки

Q трудоемкость ID3 линейна по числу признаков и по длине выборки - whyy?

Sources

https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie/lecture/d9Rvs/alghoritm-postroieniia-rieshaiushchiegho-dierieva

https://habrahabr.ru/post/171759/

https://habrahabr.ru/post/116385/

http://cda.psych.uiuc.edu/multivariate_fall_2012/systat_cart_manual.pdf

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#workings

http://xn--90abr5b.xn--p1ai/exams/%D1%81%D0%B0%D0%BE%D0%B4/35.html