part#2.Nearest neighbours

by googlenerds

What for -- why to Use?

What are non-parametric methods and why care?
Your favourite music streamer recommendations runs something non-parametric

The problem (example)

We have tracked how many times each user has visited a website

We know some of the users a little more than others

Sometimes we know their gender and how old are they (e.g. from their Facebook profile)

Classification: what is the gender of arbitrary user ?

Regression: what is the age of arbitrary user ?

formalization:

- There is an unknown dependency $f : \mathcal{X} \to \mathcal{Y}$
- We have a few noisy observations of f, called the training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i = f(x_i) + \text{noise})\}_i \subset \mathcal{X} \times \mathcal{Y}$
- The problem is called classification if $\mathcal{Y} \subset \mathbb{N}$
- The problem is called regression if $\mathcal{Y} \subseteq \mathbb{R}$

Nearest neighbour - what's it?

- Let's just assign $f(x)$ to the value $y \in \mathcal{Y}$ of the nearest known example x_{nearest} from $\mathcal{D}_{\text{train}}$ determined by a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$
- This procedure is known as the nearest neighbor algorithm (NN)
- No parameters, no model, we just remember the training set
- Major drawback: any outlier gets some followers

The distance function

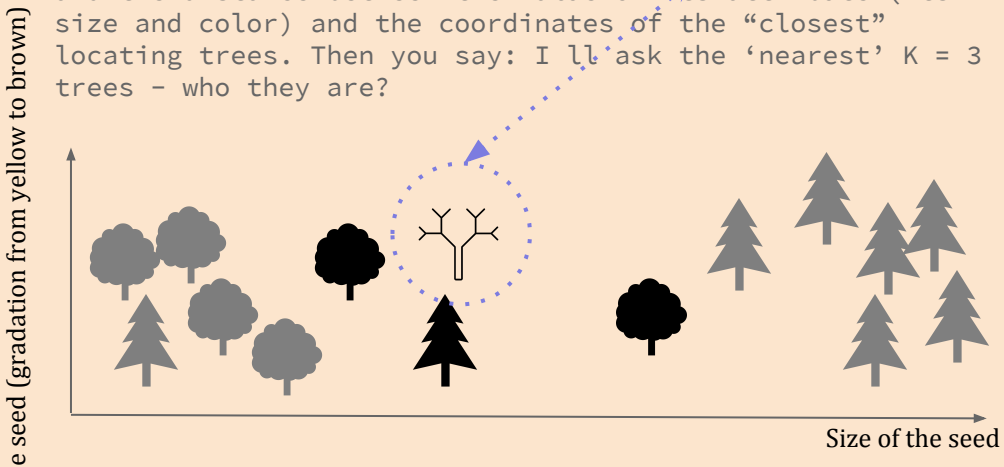
- The function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is called a distance function iff:
  - it is non-negative: $\forall_{x,x'} \;\; d(x, x') \geq 0$
  - it's zero implies identity: $d(x, x') = 0 \implies x = x'$
  - it is symmetrical: $d(x, x') = d(x', x)$
  - it satisfies the triangle inequality: $\forall_z d(x, x') \leq d(x, z) + d(z, x')$
- Let's use euclidean distance defined in $\mathcal{X} = \mathbb{R}^D$:

$$d(x, x') = \sqrt{\sum_i (x_i - x'_i)^2}$$

Now for 5-years old:

With the algorithm you have the sample with the attributes and with the answers for classification/regression problem. The color/size of the seed and the info about plant type which grew from it:

Now look at your object you re **not sure about** (newborn tree) and the distance between the value of its attribute (its size and color) and the coordinates of the "closest" locating trees. Then you say: I ll ask the 'nearest' K = 3 trees - who they are?



You see that your trees voted for your newborn to become a bush, not a fir-tree (2 vs 1). The majority opinion in the case would be the answer for the classification problem, the average - the solution of the regression one.

KNN Parameters

You can vary the K meaning:

Large K make the prediction smooth and and

#1 Real tree example

#2 TOUCH THE TREE

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.

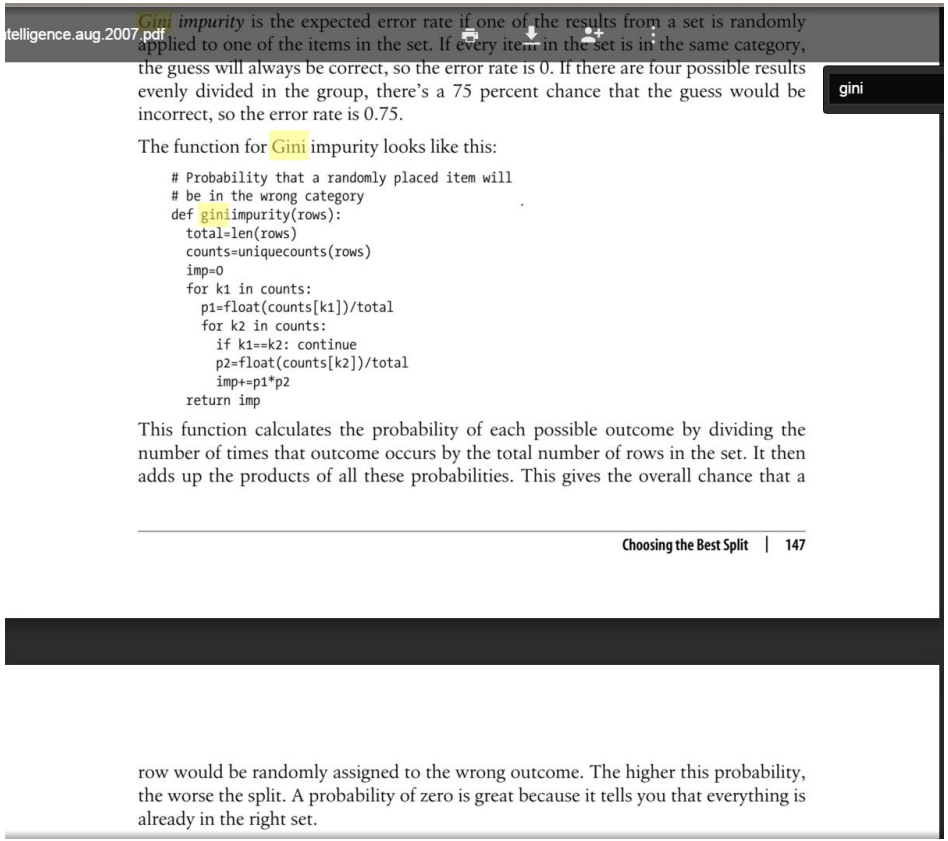
```
s0 = вычисляем энтропию исходного множества

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 creteria

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

Поиск наилучшего атрибута

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.



telligence.aug.2007.pdf

gini

Gini impurity is the expected error rate if one of the results from a set is randomly applied to one of the items in the set. If every item in the set is in the same category, the guess will always be correct, so the error rate is 0. If there are four possible results evenly divided in the group, there's a 75 percent chance that the guess would be incorrect, so the error rate is 0.75.

The function for Gini impurity looks like this:

```
# Probability that a randomly placed item will
# be in the wrong category
def giniimpurity(rows):
  total=len(rows)
  counts=uniquecounts(rows)
  imp=0
  for k1 in counts:
    p1=float(counts[k1])/total
    for k2 in counts:
      if k1==k2: continue
      p2=float(counts[k2])/total
      imp+=p1*p2
  return imp
```

This function calculates the probability of each possible outcome by dividing the number of times that outcome occurs by the total number of rows in the set. It then adds up the products of all these probabilities. This gives the overall chance that a

Choosing the Best Split | 147

row would be randomly assigned to the wrong outcome. The higher this probability, the worse the split. A probability of zero is great because it tells you that everything is already in the right set.

Criteria donskogo

Reverse to Donskogo

Cross validation

Andrew ng vs yandex

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)

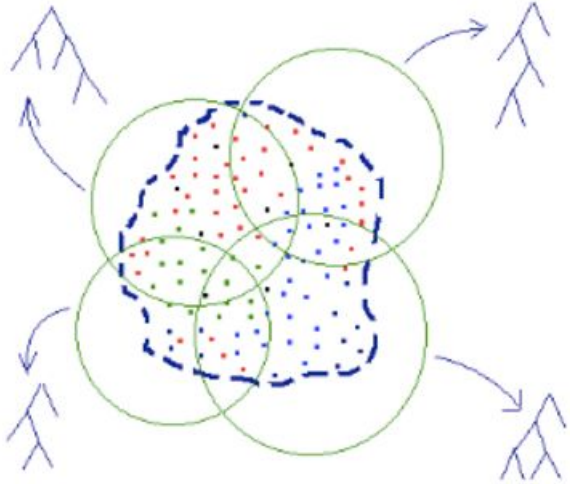
STOP WHEN: restrictions on # of nodes OR

Min error → 0

Random forest

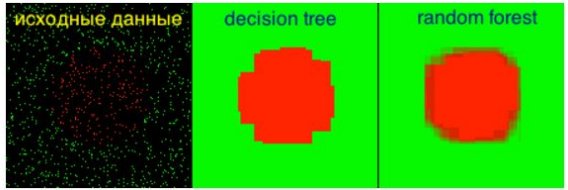
Можно сделать ансамбль деревьев, которые будут голосовать за каждый объект на тему его принадлежности к классу: берём несколько рандомных выборок, строим свои деревья - и потом суммируя голоса определяем принадлежность объекта к классу

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