Naive Bayes

林俊宇

An Example - Ambiguity of natural languages

The girl saw the boy with a telescope.

Example 2 - Boy or Girl?

- There are 60% boys and 40% girls in one school.
- All the boys are always wearing pants.
- 50% girls are wearing skirts and the others are wearing pants.
- It's easy to count “the probability of he (or her) wearing pants if we picked one student randomly”.

- But if you wander around the campus and meet up with one student wearing pants (unfortunately you were in high myopia and couldn't tell its gender), what is the probablity of this student being a boy (or girl)?

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- It is a basic theorem that has a very important influence on math, finance, gaming theory and biogen.
- We might use this formula in our daily life unconsciously.

Who is Bayes?

Thomas Bayes (1701 – 1761) was an English statistician who is known for formulating a specific case of the theorem that bears his name: Bayes theorem.

Bayes never published what would become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

Background

- In the first decades of the eighteenth century, many problems concerning the probability of certain events, given specified conditions, were solved.
- For example: given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?
- Or the converse: given that one or more balls has been drawn, what can be said about the number of white and black balls in the urn? These are sometimes called "inverse probability" problems.

Example 3 - Spelling Correction

- We might have spelling mistakes when we are typing.
- Assume there is an auto spelling correction machine. How does it work to correct your spelling mistake?

- the -> thew
- why not thaw?

Example 3 - Spelling Correction

$$P(h_i \mid D) = \frac{P(h_i)P(D \mid h_i)}{P(D)}$$

$$P(h_i \mid D) \propto \frac{P(h_i)P(D \mid h_i)}{P(D)}$$

$$P(h_i \mid D) \propto P(h_i)P(D \mid h_i)$$

“the”在词汇表使用频率

想打the却打成thew的可能性

Why Bayes?

- One alternative solution is to choose the word that has the nearest edit distance with “thew”.

- However, “the” and “thaw” are all having 1 edit distance with “thew”.

Why Bayes?

- Actually the letter “e” and “w” are quite close on the keyboard, it is quite likely to type one more “w” accidently.

- The probability of type “thew” instead of “thaw” is higher, as “e” and “a” are much farther.

Maximum Likelihood Estimation (MLE)

- You are selecting h which makes P(D|h) max.

Bayes Theorem

- Selecting h which makes P(h)*P(D|h) max.

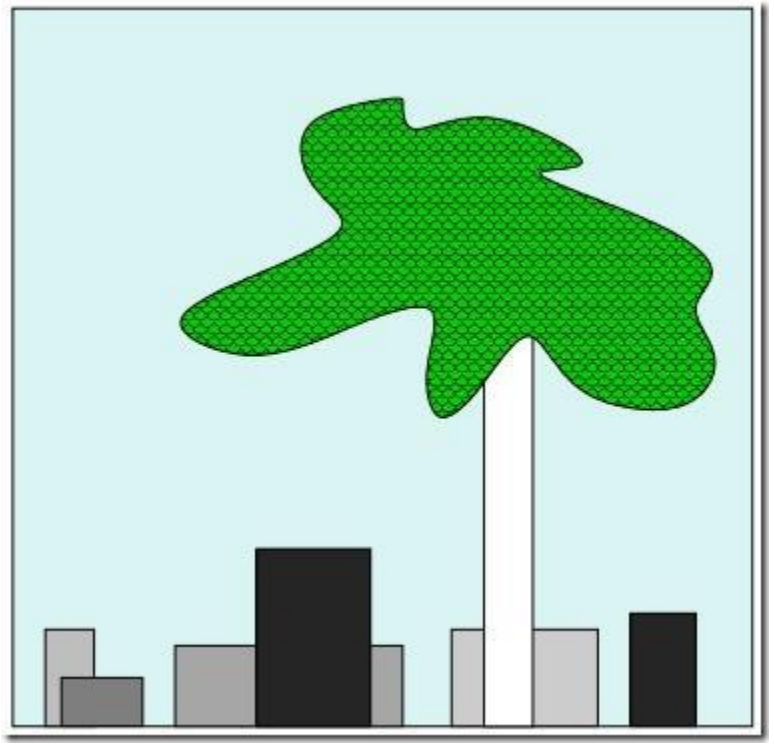
Disadvantages of MLE

- Assume the frequency of “tip” is higher than “top” for this user.
- Input: tlp
- tip or top?
- It is hard to tell by MLE, but easy by Bayes Theorem.

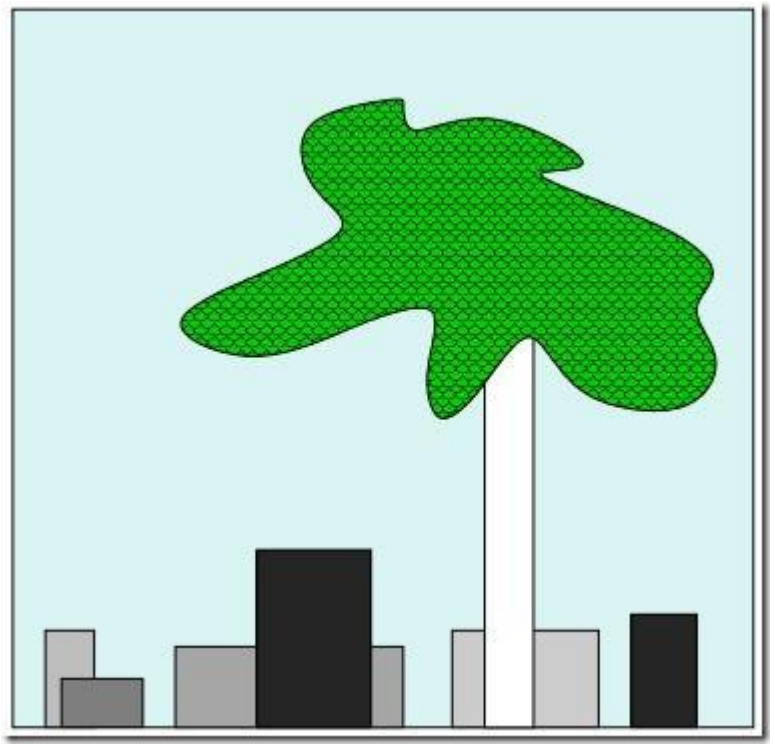
Disadvantages of MLE

- Maximum Likelihood Estimation can't provide all the information for policy decision.

Example 4 - How many boxes are behind the tree?



Example 4 - How many boxes are behind the tree?



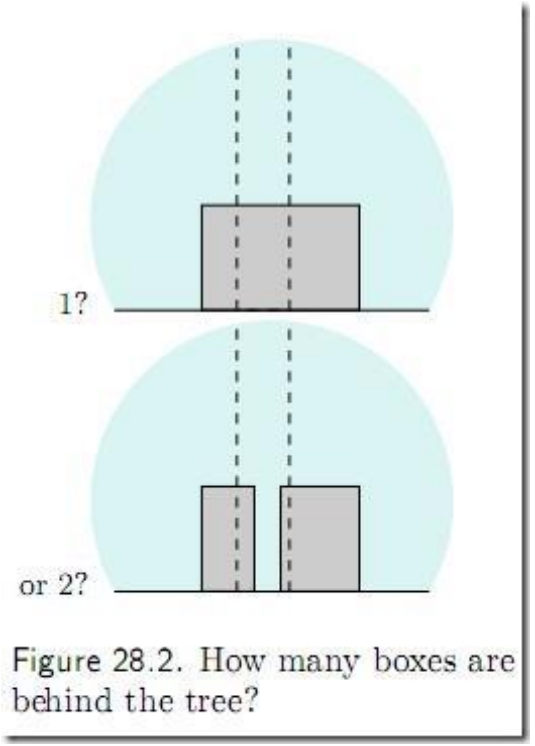


Figure 28.2. How many boxes are behind the tree?

Occam's Razor

- The simplest solution tends to be the correct one.

Example 5 - 中文分词

- 给定一个句子，如：南京市长江大桥
- 如何对这个句子进行最靠谱的分词？
  1. 南京市/长江/大桥
  2. 南京/市长/江大桥

Example 5 - 中文分词

- 令X为句子，Y为词串，即求使得P(Y|X)最大的Y。

根据贝叶斯公式可得：

$$P(Y \mid X) \propto P(Y)P(X \mid Y)$$

- 事实上P(X|Y)恒等于1，则最终变成了最大化P(Y)，即寻找一种分词使得词串的概率最大化。

Example 5 - 中文分词

- 如何计算一个词串W1, W2, W3, ...
- 根据联合概率的公式展开：

  $P(W1, W2, W3) = P(W1) * P(W2|W1) * P(W3|W2, W1)$

- 在这里作一个假设，假设句子中一个词的出现概率只依赖于它前面的有限的 k 个词
- (如果只依赖于前面的一个词，就是2元语言模型2-gram)

  $P(W3|W2, W1) = P(W3|W2)$

Example 5 - 中文分词

- P(W1, W2, W3) = P(W1) * P(W2|W1) * P(W3|W2)
- P(“江大桥”|“市长”)趋近于0
- P(“大桥”|”长江”)远大于0
- 所以“南京市/长江/大桥”的分词方式胜出

朴素贝叶斯 - 垃圾邮件分类

- 给定一封邮件，判定它是否属于垃圾邮件。
- D ={这封邮件} (D 由 N 个单词组成)
- h+ ={垃圾邮件}
- h- ={正常邮件}
- 问题可以形式化地描述为求：

  $P(h+|D) = P(h+) * P(D|h+) / P(D)$

  $P(h-|D) = P(h-) * P(D|h-) / P(D)$

朴素贝叶斯 - 垃圾邮件分类

- P(h+|D) = P(h+) * P(D|h+) / P(D)
- P(h-|D) = P(h-) * P(D|h-) / P(D)
- 其中 P(h+) 和 P(h-) 这两个先验概率都是很容易求出来的，只需要计算一个邮件库里面垃圾邮件和正常邮件的比例就行了。

朴素贝叶斯 - 垃圾邮件分类

- P(h+|D) = P(h+) * P(D|h+) / P(D)
- P(h-|D) = P(h-) * P(D|h-) / P(D)
- 然而 P(D|h+) 却不容易求，因为 D 里面含有 N 个单词 d1, d2, d3, .. ，所以P(D|h+) = P(d1,d2,..,dn|h+)

朴素贝叶斯 - 垃圾邮件分类

- 根据联合概率分布公式：
  P(d1,d2,..,dn|h+)=P(d1|h+) * P(d2|d1, h+) * P(d3|d2,d1, h+) * ...
- 这里作一个更激进的假设，
  P(d1,d2,..,dn|h+)=P(d1|h+) * P(d2|h+) * P(d3|h+) *...
- 这个就是所谓的条件独立假设，也正是朴素贝叶斯方法的朴素之处。而计算 P(d1|h+) * P(d2|h+) * P(d3|h+) * .. 就太简单了，只要统计 di 这个单词在垃圾邮件中出现的频率即可。

Thanks For Your Listening!