拉普拉斯平滑
(Laplace Smoothing)

Grissom, 2019/07/16

http://sunbo2019.github.io

If categorical variable has a category in test data set and it is absent in the train data set then model will assign zero probability and will be unable to make a prediction.This is known as “Zero Frequency” .To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called **Laplace estimation**.

零概率问题：在计算概率时，如果某个事件在训练集中没有出现过，但在测试集中出现了，会导致整个实例的与**这个事件有关的连乘概率**结果为0。

举例：

在训练集中有三个词：发票、照片、代开

p(发票|正常邮件) = 0.5， p(发票|正常邮件) = 0.5
p(照片|正常邮件) = 0.6， p(照片|垃圾邮件) = 0.4
p(代开|正常邮件) = 0.1， p(代开|垃圾邮件) = 0.9

那么，

① 新邮件A中包含了“发票，照片” 2个词：

p(发票|正常邮件) * p(照片|正常邮件) = 0.5 * 0.6 = 0.3
p(发票|垃圾邮件) * p(照片|垃圾邮件) = 0.5 * 0.4 = 0.2

=> 邮件A被划分为正常邮件。

② 新邮件B中包含了“发票，代开” 2个词：

p(发票|正常邮件) * p(代开|正常邮件) = 0.5 * 0.1 = 0.05
p(发票|垃圾邮件) * p(代开|垃圾邮件) = 0.5 * 0.9 = 0.45

=> 邮件B被划分为垃圾邮件。

③ 新邮件C中包含了“发票，NLP” 2个词：

因为“NLP”在训练集中没有出现过，所以这里的p(NLP|正常邮件) = p(NLP|垃圾邮件) = 0，从而造成：

p(发票|正常邮件) * p(NLP|正常邮件) = 0.5 * 0 = 0
p(发票|垃圾邮件) * p(NLP|垃圾邮件) = 0.5 * 0 = 0

=> 无法对邮件C进行分类

这种情况，就叫着“Zero Frequency”（零概率问题）。

Q：“零概率”？某个词出现在某类文章中的概率就真的为零吗？
A：　并不是这样！它只是在training dataset 中没有被收集到而已。

* 连乘概率为零的问题，可以通过取log将连乘变连加来解决。这是问题的表象而不是核心。
我们之所以在这里纠结“零概率”问题，是因为：**一个事件是未知的，并不意味着它出现的概率为零。**

所以，我们需要找到一种方法在出现“**Zero Frequency**”现象时合理地预估（能解决问题，又符合概率和为1的要求）一个概率出来。于是提出了**“加法平滑”**的处理方案！

- 法国数学家拉普拉斯最早提出用加1的方法来估计未知事件的概率，所以加法平滑也叫做**拉普拉斯平滑**（In statistics, additive smoothing, also called Laplace smoothing）。

- 其本质上是“**劫富济贫**”的思想！
  即：从高频事件上移一部概率到低频事件上。同时保证了概率和=1。当数据集中事件很丰富并且k很小时，从每一个高频事件上移开一个很小的概率（到未知事件上），影响不大，同时又可以避免未知事件的“Zero Frequency”现象。

Laplace Smoothing的数学公式

（1）概率表达： $$P(w_s) = \frac{C(w_s)+1}{N+V}$$

p(w) - 词w出现的概率
C(w) - 词w在数据集中出现的次数 [C for count]
N - 数据集中所有词个数（有重复）
V - 词表长度（无重复）

（2）条件概率表达： $$P_{add-k}(w_i|w_{i-1}) = \frac{c(w_{i-1},w_i)+k}{c(w_{i-1})+kV}$$

例：当k=1时， $$P_{Add-1}(w_i|w_{i-1}) = \frac{c(w_{i-1},w_i)+1}{c(w_{i-1})+V}$$

其实，这就像我们编程时，对可能出错的地方进行try..catch..一样，用“拉普拉斯平滑”来处理某一类概率在训练数据中没有出现，但是在预测的数据集上却出现了的情况。

计算题（1）

Laplace for Single Variable Distribution

- Given: a1, a2, a1, a2, a3, a1, a3, a2
- Asked: Laplace-k smoothed estimate of P(A) with domain of A = {a1, a2, a3} for k=1

$$P(A=a_1)=\frac{3}{8}=\frac{33}{88}$$
$$P(A=a_2)=\frac{3}{8}=\frac{33}{88}$$
$$P(A=a_3)=\frac{2}{8}=\frac{22}{88}$$

K=1 Laplace Smoothing

$$P(A=a_1)=\frac{3+1}{8+3\cdot 1}=\frac{4}{11}=\frac{32}{38}$$
$$P(A=a_2)=\frac{3+1}{8+3\cdot 1}=\frac{4}{11}=\frac{32}{88}$$
$$p(A=a_3)=\frac{2+1}{8+3\cdot 1}=\frac{3}{11}=\frac{24}{88}$$

本质：（1）在原数据集中，增加了一组元素（k=1），即：a1, a2, a1, a2, a3, a1, a3, a2 + **a1, a2, a3**。
（2）稀释了原来高频事件的概率，移一部分概率到低频事件上。**当k很大时，接近均匀分布（uniform distribution）**。

* https://www.youtube.com/watch?v=gCI-ZC7irbY

计算题（2）

Laplace for Conditional Distribution

- Given: (a1, b1), (a2, b2), (a1, b2), (a1, b3), (a2, b2), (a1, b1)
- Asked: Laplace-k estimates of P(B|A), for k=3

<原来的概率>

$$P(B=b_1|A=a_1)=\frac{2}{4}$$

$$P(B=b_2|A=a_1)=\frac{1}{4}$$

$$P(B=b_3|A=a_1)=\frac{1}{4}$$

$$P(B=b_1|A=a_2)=\frac{0}{2}=0$$

$$P(B=b_2|A=a_2)=\frac{2}{2}=1$$

$$P(B=b_3|A=a_2)=\frac{0}{2}=0$$

$\rightarrow$

$$A=\{a_1,a_2\}$$

$$B=\{b_1,b_2,b_3\}$$

k=3，即增加如下一组数据到原集合中:

$$(a_1,b_1),(a_1,b_2),(a_1,b_3)$$

$$(a_1,b_1),(a_1,b_2),(a_1,b_3)$$

$$(a_1,b_1),(a_1,b_2),(a_1,b_3)$$

$$(a_2,b_1),(a_2,b_2),(a_2,b_3)$$

$$(a_2,b_1),(a_2,b_2),(a_2,b_3)$$

$$(a_2,b_1),(a_2,b_2),(a_2,b_3)$$

$\rightarrow$

<拉普拉斯平滑后的概率>

$$P(B=b_1|A=a_1)=\frac{5}{12}$$

$$P(B=b_2|A=a_1)=\frac{4}{12}$$

$$P(B=b_3|A=a_1)=\frac{4}{12}$$

$$P(B=b_1|A=a_2)=\frac{3}{11}$$

$$P(B=b_2|A=a_2)=\frac{5}{11}$$

$$P(B=b_3|A=a_2)=\frac{3}{11}$$