

20. 算法-朴素贝叶斯

1. 概念

公式:

在事件B发生的前提下, 事件A发生的概率

在事件A发生的前提下, 事件B发生的概率

事件A发生的概率

事件B发生的概率

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$
$$\frac{2}{4} = \frac{\frac{2}{5} \times \frac{3}{10}}{\frac{4}{10}}$$

2. 应用

过滤垃圾短信
过滤垃圾短信一般的方式是通过黑名单或者建立某种规则, 但是这两种方式都不太灵活, 可以使用朴素贝叶斯的方式计算这个短信是垃圾短信的概率, 我们再进行处理

思路如下:

使用直接统计是不现实的, 因为一条短信当中不可能同时包含 w_1, w_2, w_3, \dots 这些词语

垃圾短信的概率表示

$$P(\text{短信是垃圾短信} | w_1, w_2, \dots, w_n \text{同时出现在该短信中})$$

使用直接统计是不现实的, 因为一条短信当中不可能同时包含 w_1, w_2, w_3, \dots 这些词语

垃圾短信的概率表示

$$P(\text{短信是垃圾短信} | w_1, w_2, \dots, w_n \text{同时出现在该短信中})$$

由于无法直接统计, 但是使用朴素贝叶斯方式可以求出概率

$$P(\text{短信是垃圾短信} | w_1, w_2, \dots, w_n \text{同时出现在该短信中}) = \frac{P(w_1, w_2, \dots, w_n \text{同时出现在垃圾短信中} | \text{短信是垃圾短信}) * P(\text{短信是垃圾短信})}{P(w_1, w_2, \dots, w_n \text{同时出现在非垃圾短信中} | \text{短信是非垃圾短信}) * P(\text{短信是非垃圾短信})}$$

由于事件是独立的, 所以朴素贝叶斯的分子就可以分开来求解

$$P(w_1, w_2, \dots, w_n \text{同时出现在垃圾短信中} | \text{短信是垃圾短信}) = P(w_1 \text{出现在垃圾短信中} | \text{短信是垃圾短信}) * P(w_2 \text{出现在垃圾短信中} | \text{短信是垃圾短信}) * \dots * P(w_n \text{出现在垃圾短信中} | \text{短信是垃圾短信})$$

垃圾短信有 y 个, 其中 x 个包含 w_1 , x/y 就是独立的概率

$$P(w_1 \text{出现在垃圾短信中} | \text{短信是垃圾短信}) = \frac{x}{y}$$

利用同样的方式求出相同的词语同时出现时垃圾短信的概率

$$P(w_1, w_2, \dots, w_n \text{同时出现在垃圾短信中} | \text{短信是垃圾短信}) = \left(\frac{x_1}{y_1}\right) * \left(\frac{x_2}{y_2}\right) * \dots * \left(\frac{x_n}{y_n}\right)$$

这样 $p_1 p_2$ 就是垃圾短信相比非垃圾短信的倍数, 倍数很大才能确定是垃圾短信 如 10 倍

$$P(w_1, w_2, \dots, w_n \text{同时出现在垃圾短信中} | \text{短信是非垃圾短信}) = \left(\frac{x_1}{y_1}\right) * \left(\frac{x_2}{y_2}\right) * \dots * \left(\frac{x_n}{y_n}\right)$$