贝叶斯

有监督的分类算法

数学知识

先验概率

事件发生前的预判概率。可以是基于历史数据的统计，可以由背景常识得出，也可以是人的主观观点给出。一般都是单独事件概率，如P(x),P(y) 。

条件概率

一个事件发生后另一个事件发生的概率。一般的形式为P(x|y)表示y发生的条件下x发生的概率。

后验概率

事件发生后求的反向条件概率；或者说，基于先验概率求得的反向条件概率。概率形式与条件概率相同。

贝叶斯公式

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

在机器学习的视角下，我们把XX理解成**“具有某特征”**，把YY理解成**“类别标签”**(一般机器学习为题中都是`X=>特征`，`Y=>结果`对吧)。在最简单的二分类问题(`是`与`否`判定)下，我们将YY理解成**“属于某类”**的标签。于是贝叶斯公式就变形成了下面的样子:

$$P(\text{”属于某类”}|\text{”具有某些特征”})=\frac{P(\text{”具有某些特征”}|\text{”属于某类”})P(\text{”属于某类”})}{P(\text{”具有某些特征”})}$$

在我们实际使用的场景，一般是计算具有某些特征，属于某些类的概率。例如：

$$P(\text{”属于某类1”}|\text{”具有某些特征”})=\frac{P(\text{”具有某些特征”}|\text{”属于某类1”})P(\text{”属于某类1”})}{P(\text{”具有某些特征”})}$$

$$P(\text{”属于某类2”}|\text{”具有某些特征”})=\frac{P(\text{”具有某些特征”}|\text{”属于某类2”})P(\text{”属于某类2”})}{P(\text{”具有某些特征”})}$$

举例（垃圾邮件识别）

识别"办理正规发票，增值税发票"类垃圾邮件。

思路

其实是邮件分类问题，首先计算出垃圾邮件的概率和非垃圾邮件的概率，如果是垃圾邮件的概率较大，就可以认为是垃圾邮件。

公式

$$P(\text{”垃圾邮件”}\mid\text{”办理正规发票，增值税发票”})=\frac{P(\text{”办理正规发票，增值税发票”}\mid\text{”垃圾邮件”})P(\text{”垃圾邮件”})}{P(\text{”办理正规发票，增值税发票”})}$$

$$P(\text{”正常邮件”}\mid\text{”办理正规发票，增值税发票”})=\frac{P(\text{”办理正规发票，增值税发票”}\mid\text{”正常邮件”})P(\text{”正常邮件”})}{P(\text{”办理正规发票，增值税发票”})}$$

因为中文的可能性太多，覆盖所有的句子很难做

条件独立假设

词语与词语之间是没有关系的，那么

$$P(\text{”办理正规发票，增值税发票”}\mid\text{”垃圾邮件”})=P(\text{”办理”}\mid\text{”垃圾邮件”})*P(\text{”正规”}\mid\text{”垃圾邮件”})*P(\text{”发票”}\mid\text{”垃圾邮件”})*P(\text{”增值税”}\mid\text{”垃圾邮件”})*P(\text{”发票”}\mid\text{”垃圾邮件”})$$

分类

多项式模型

伯努利模型

高斯模型