朴素贝叶斯算法
Naïve Bayes Algorithms

部分课件来源：余超老师课程助教团队

目录

贝叶斯概率的基础思想

- **概率：** 某人对一个命题信任的程度，这个概率不像频率概率范畴是描述某个随机事件发生的可能性的一个未知的定值，而是指一个未知的可以变化的值。
- **先验概率：** 不知道其他信息情况下，根据以往经验和分析得到的概率。
- **后验概率：** 在得到随机事件的结果之后对先验概率进行修正之后的概率。
- **全概率公式：** 将复杂事件的概率求解转换为简单概率求解的方法。
- **贝叶斯公式/定理：** 贝叶斯概率与贝叶斯统计的基础，基本上是用来计算后验概率的公式。

相关公式

- **变量相互独立：**

$$\mathbf{P}(\mathrm{X}|\mathrm{Y}) = \mathbf{P}(\mathrm{X}) \text{ or } \mathbf{P}(\mathrm{Y}|\mathrm{X}) = \mathbf{P}(\mathrm{Y}) \text{ or } \mathbf{P}(\mathrm{X},\mathrm{Y}) = \mathbf{P}(\mathrm{X})\mathbf{P}(\mathrm{Y})$$

- **条件概率：**

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

- **贝叶斯规则：**

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

$$\mathbf{P}(\mathrm{X},\mathrm{Y}) = \mathbf{P}(\mathrm{X}|\mathrm{Y})\mathbf{P}(\mathrm{Y}) = \mathbf{P}(\mathrm{Y}|\mathrm{X})\mathbf{P}(\mathrm{X})$$

- **链式法则：**

$$\mathbf{P}(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_\mathrm{N}) = \mathbf{P}(\mathrm{X}_1|\mathrm{X}_2,\ldots,\mathrm{X}_\mathrm{N}) * \mathbf{P}(\mathrm{X}_2|\mathrm{X}_3,\ldots,\mathrm{X}_\mathrm{N}) * \ldots * \mathbf{P}(\mathrm{X}_{\mathrm{N}-1}|\mathrm{X}_\mathrm{N}) * \mathbf{P}(\mathrm{X}_\mathrm{N})$$

1.1 朴素贝叶斯法的学习与分类

考虑一个分类问题，我们希望根据动物的某些特征来区分猫（y = 1）和狗（y = 0）。

- **判别模型**
  - 找到将猫和狗分开的决策边界或分类原则。
  - 为了分类一只新动物，判别模型会检查它落在决策边界的哪一边，并直接做出决定。
- **生成模型**
  - 分别建立猫和狗的外观模型。
  - 要对新动物进行分类，将其与猫/狗模型进行匹配，并查看它看起来更像哪个模型。

1.1 朴素贝叶斯法的学习与分类

- **判别模型**
  - 直接估计后验概率 $p(y|x)$ 。
  - 常用方法：k-近邻，etc.
- **生成模型**
  - 估计先验概率 $p(y)$ 和条件概率 $p(x|y)$。
  - 根据贝叶斯定理计算后验概率 $p(y|x)$ 。
  - 常用方法：Naïve Bayes，etc.

1.1 朴素贝叶斯法的学习与分类

朴素贝叶斯法根据**贝叶斯定理**来估计每个类别的**后验概率**。

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_i p(x|y_i)p(y_i)} \propto p(x|y)p(y)$$

朴素贝叶斯法的目标是找到

$$y = \arg\max_y p(y|x) = \arg\max_y \frac{p(x,\ y)}{p(x)} = \arg\max_y p(x|y)p(y)$$

1.1 朴素贝叶斯法的学习与分类

朴素贝叶斯法对条件概率分布作了**条件独立性假设**。这是一个较强的假设，具体地，该假设是：

$$p(x_1, x_2, ..., x_n|y) = \prod_{k=1}^{n} p(x_k|y)$$

条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的。

1.1 朴素贝叶斯法的学习与分类

给定一个包含 M 个文本的数据集，其中每个有 K 维特征向量 $X = (x_1, ..., x_K)$ 和一个情感标签 e_i，为了预测测试文本，需要估计：

$$\begin{aligned}\arg\max_{e_i} p(e_i|X) &= \arg\max_{e_i} \frac{P(X|e_i)p(e_i)}{p(X)} \\ &= \arg\max_{e_i} p(X|e_i)p(e_i) \\ &= \arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)\end{aligned}$$

1.2 两种概率模型

$$\arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

- 伯努利模型：

$$\mathrm{p}(x_k|e_i) = \frac{n_{e_i}(x_k)}{N_{e_i}} \qquad p(e_i) = \frac{N_{e_i}}{N}$$

其中，$n_{e_i}(x_k)$表示类别为e_i的文本中出现x_k的文本数量；N_{e_i}表示类别为e_i的文本数量；N表示全部文本的数量。

在伯努利模型中，当某一文本出现单词x_k，那么$n_{e_i}(x_k)$则$+1$。

1.2 两种概率模型

$$\arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

- 多项式模型：

$$p(x_k|e_i) = \frac{nW_{e_i}(x_k)}{nW_{e_i}} \qquad p(e_i) = \frac{N_{e_i}}{N}$$

其中，$nW_{e_i}(x_k)$表示类别为e_i的文本中出现x_k的总次数；nW_{e_i}表示类别为e_i的文本中单词的总数；N_{e_i}表示类别为e_i的文本数量；N表示全部文本的数量。

因此，我们可以得知，伯努利模型中，我们关注的是单词x_k是否有在文本中出现；而多项式模型中，则关注的是单词x_k在文本中出现的次数。

1.2 两种概率模型

ID	text	class label
1	good,thanks	joy
2	No impressive, thanks	sad
3	Impressive good	joy
4	No, thanks	?

ID	goods	thanks	no	impressive	class label
1	1	1	0	0	joy
2	0	1	1	1	sad
3	1	0	0	1	joy
4	0	1	1	0	?

Bernoulli Model（伯努利模型）：

$$P_{(thanks|joy)} = 1/2$$

Multinomial Model（多项式模型）：

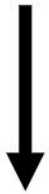
$$P_{(thanks|joy)} = 1/4$$

思考题：这两个模型分别有什么优缺点

1.2 两种概率模型

$$\arg\max_{e_i} \prod_{k=1}^{K} p(x_k|e_i)p(e_i)$$

ID	text	class label
1	good,thanks	joy
2	No impressive, thanks	sad
3	Impressive good	joy
4	No, thanks	?



ID	goods	thanks	no	impressive	class label
1	1	1	0	0	joy
2	0	1	1	1	sad
3	1	0	0	1	joy
4	0	1	1	0	?

Target function:

$$p(joy|d_4) = p(joy) \cdot p(d_4|joy)$$
$$p(sad|d_4) = p(sad) \cdot p(d_4|sad)$$

Example:

$$\begin{aligned} p(joy|d_4) &= p(d_4|joy) \cdot p(joy) \\ &= p(\text{“}thanks\text{”}, \text{“}no\text{”}|joy) \cdot p(joy) \\ &= p(\text{“}thansk\text{”}|joy) \cdot p(\text{“}no\text{”}|joy) \cdot p(joy) \\ &= \frac{1}{4} \times 0 \times \frac{2}{3} = 0 \end{aligned}$$

1.3 拉普拉斯平滑

思考：在前面的文本分类算法中，如果测试文本中的单词没有在训练文本中出现会造成什么结果？

会影响到后验概率的计算结果，使分类产生偏差。解决这一问题的方法是采用**拉普拉斯平滑** (Laplacian smoothing)：

$$Bernoulli: p(x_k|e_i) = \frac{n_{e_i}(x_k) + 1}{N_{e_i} + 2}$$

$$Multinomial: p(x_k|e_i) = \frac{nw_{e_i}(x_k) + 1}{nw_{e_i} + V_{e_i}}$$

这里V_{e_i}值得是类别为e_i的文本中不重复的单词数量。

知识补充

训练集 验证集 测试集的区别

数据类型	有无标签	作用
训练集(training set)	有	用来训练模型或确定模型参数的，如k-NN中权值的确定等。 相当于平时练习。
验证集(validation set)	有	用来确定网络结构或者控制模型复杂程度的参数，修正模型。 相当于模拟考试。
测试集(test set)	无	用于检验最终选择最优的模型的性能如何。 相当于期末考试。

实验任务

- 第11周：朴素贝叶斯
- 第12周：*k*-近邻算法

- 实验任务：文本情感分析
  - 在朴素贝叶斯分类、*k*-NN分类与*k*-NN回归中，至少完成一项；
    - 用朴素贝叶斯法完成文本信息情感分类训练，要求使用拉普拉斯平滑技巧。
  - 鼓励尝试多种算法及算法中的不同策略/参数；
  - 完成一份实验报告，注意实验报告要求。

实验提交

- **作业名称：实验5**
- **截止时间：5月11日 23:00**
- 本次实验提交样例：压缩包20******_wangxiaoming.zip，

  内含：
  - 20******_wangxiaoming.pdf
  - /code：文件夹，内含所有实验代码并附上readme
  - /result：文件夹，内含实验结果（根据完成情况，至少包含一个）
    - 20******_wangxiaoming_NB_classification.csv
    - 20******_wangxiaoming_KNN_classification.csv
    - 20******_wangxiaoming_KNN_regression.csv