补充朴素贝叶斯 (Complement Naive Bayes) 详细介绍

补充朴素贝叶斯 (Complement Naive Bayes)

补充朴素贝叶斯（Complement Naive Bayes, CNB）是朴素贝叶斯分类器的一种变体，由 Rennie 等人于 2003 年提出。它旨在**解决标准多项式朴素贝叶斯在处理不平衡数据集时的一些固有缺陷**，尤其在文本分类任务中表现突出。

核心思想

传统的（多项式）朴素贝叶斯在训练时，会计算每个词在**每个类别中出现的频率**，然后用这些频率来判断新文档属于哪个类别。当数据集高度不平衡时（例如，某些类别包含的文档数量远多于其他类别），多项式朴素贝叶斯往往会**偏向于样本数量多的主导类别**。这是因为在计算词在某个类别下的概率时，主导类别的词频统计会占据更大的权重。

补充朴素贝叶斯的**创新之处**在于，它不是直接建模一个特征属于某个类别的概率，而是建模一个特征属于**该类别的“补集”的概率**（即属于所有其他类别的概率）。

“Model what the class is not, to better understand what it is.”（通过建模类别不属于什么，来更好地理解它是什么。）

具体来说，CNB 为每个类别 C_k 计算一个权重，这个权重是基于数据点不属于 C_k（即属于 C_k 的补集 $\bar{C}_k$）的统计信息。最终，它选择 **“最不可能不属于”** 的类别作为预测结果，换句话说，选择在**补集中得分最低**的类别。

数学原理

对于一个给定的数据点 $x = (x_1, x_2, ..., x_n)$ 和一个类别 C_k，CNB 计算一个“补集”分数 $S(C_k|x)$。这个分数越低，表示数据点 x 越不可能不属于类别 C_k，从而越可能属于 C_k。

其计算方式通常涉及到对数概率：

$$S(C_k|x) = \log P(\bar{C}_k) + \sum_{i=1}^{n} x_i \log P(x_i|\bar{C}_k)$$

其中：

- $P(\bar{C}_k)$ 是类别 C_k 的补集的先验概率，即所有不属于 C_k 的文档的比例。
- $P(x_i|\bar{C}_k)$ 是特征 x_i 在类别 C_k 的补集中出现的概率。这正是 CNB 的关键所在。

$$P(x_i|\bar{C}_k) = \frac{\text{count}(x_i, \bar{C}_k) + \alpha}{\sum_{x \in V}(\text{count}(x, \bar{C}_k) + \alpha)}$$

其中：

- $\text{count}(x_i, \bar{C}_k)$ 是特征 x_i 在所有不属于类别 C_k 的文档中出现的总次数。
- $\sum_{x \in V}(\text{count}(x, \bar{C}_k) + \alpha)$ 是所有特征在类别 C_k 的补集中出现的总次数（包括平滑项）。
- V 是词汇表（所有不同特征的集合）。
- α 是拉普拉斯平滑参数，通常取 $\alpha = 1$。

最终的预测：模型选择使 $S(C_k|x)$ 最小的类别 C_k 作为预测结果。

$$\hat{y} = \arg\min_{C_k} S(C_k|x)$$

通过这种方式，CNB 能够更好地处理不平衡数据，因为它关注的是 **“负面证据”**。如果一个文档包含大量不属于某个类别的词，那么它就不太可能属于那个类别。这种“逆向”的思考方式使其对少数类别的建模更为鲁棒。

Scikit-learn 中的实现

在 Scikit-learn 中，补充朴素贝叶斯通过 `sklearn.naive_bayes.ComplementNB` 类实现。
主要参数：

- `alpha`: 浮点型，默认值为 `1.0`。这是拉普拉斯/利德斯通平滑参数。与多项式朴素贝叶斯中的作用类似，防止零概率问题。
- `fit_prior`: 布尔型，默认值为 `True`。表示是否学习类别先验概率。如果为 `False`，则所有类别都使用均匀先验（即每个类别概率相等）。通常保持为 `True`。
- `class_prior`: 数组类型，默认值为 `None`。如果指定，则为每个类别设置先验概率。如果为 `None` 且 `fit_prior` 为 `True`，则根据训练数据中每个类别的样本比例自动计算先验概率。
- `norm`: 布尔型，默认值为 `False`。表示是否对权重进行二次归一化。这在某些情况下可以进一步提升性能，尤其是在文档长度差异很大的文本分类中。

适用场景

补充朴素贝叶斯最常用于：

- **文本分类**：与多项式朴素贝叶斯类似，但特别适用于词频（TF）或 TF-IDF 特征。
- **不平衡数据集**：这是 CNB 的主要优势。当数据集中某些类别的样本数量远少于其他类别时，CNB 往往比标准的多项式朴素贝叶斯表现更好。

优点

- **对不平衡数据鲁棒**：通过关注类别的补集，有效缓解了少数类别被多数类别“淹没”的问题。
- **在文本分类中表现出色**：通常在文本分类任务中优于多项式朴素贝叶斯，尤其是在不平衡数据集上。
- **计算高效**：与所有朴素贝叶斯算法一样，训练和预测速度快。
- **参数估计更稳定**：经验表明，CNB 的参数估计比 MNB 更稳定。

缺点

- **“朴素”假设**：与所有朴素贝叶斯模型一样，特征条件独立性假设在现实中很少完全成立。
- **主要用于计数型或 TF-IDF 特征**：设计初衷是用于文本数据，因此对其他类型的特征（如连续特征或二元特征）可能不如其各自对应的朴素贝叶斯变体（如高斯朴素贝叶斯或伯努利朴素贝叶斯）适用。

总结

补充朴素贝叶斯是多项式朴素贝叶斯的一个重要改进，特别针对不平衡文本分类任务进行了优化。通过其独特的“补集”建模方式，它能够更公平地评估每个类别的概率，从而在许多现实世界的问题中提供更准确和鲁棒的性能。