伯努利朴素贝叶斯 (Bernoulli Naive Bayes) 详细介绍

伯努利朴素贝叶斯 (Bernoulli Naive Bayes)

伯努利朴素贝叶斯是朴素贝叶斯分类器的一种变体，它特别适用于**二元（Boolean）或布尔型特征**的分类问题。与多项式朴素贝叶斯关注特征的计数或频率不同，伯努利朴素贝叶斯关注特征**是否存在**（即特征值为 0 或 1）。

核心思想

伯努利朴素贝叶斯，像所有朴素贝叶斯模型一样，基于朴素贝叶斯假设：给定类别的情况下，特征之间是条件独立的。它假设每个特征的概率分布服从**伯努利分布**。

这意味着对于每个特征，我们只关心它是否出现，而不关心它出现的次数。

例如：在文本分类中，如果使用伯努利朴素贝叶斯，一个文档的特征不是某个单词出现的次数，而是这个单词是否出现在文档中（1 表示出现，0 表示未出现）。

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，我们想要计算 $P(C_k|x)$。根据贝叶斯定理，我们有：

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

其中各项的含义与之前朴素贝叶斯模型相同。我们仍然需要最大化 $P(x|C_k)P(C_k)$。

朴素贝叶斯假设：特征之间是条件独立的。

$$P(x|C_k) = P(x_1, x_2, ..., x_n|C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

其中 x_i 是数据点 x 的第 i 个特征。

伯努利分布假设：对于每个特征 x_i 在给定类别 C_k 的情况下，其概率分布服从伯努利分布。这意味着我们估计的是特征 x_i 在类别 C_k 中出现（值为 1）的概率 $P(x_i = 1|C_k)$ 和不出现（值为 0）的概率 $P(x_i = 0|C_k)$。

$$P(x_i|C_k) = p_{ik}^{x_i}(1 - p_{ik})^{1-x_i}$$

其中：

- $p_{ik} = P(x_i = 1|C_k)$ 是特征 x_i 在类别 C_k 中出现的概率。
- x_i 只能取值 0 或 1。

我们可以通过最大似然估计来计算 p_{ik}：

$$p_{ik} = \frac{\text{Number of samples in } C_k \text{ where } x_i = 1 + \alpha}{\text{Number of samples in } C_k + \alpha \times 2}$$

其中 α 是拉普拉斯平滑参数（通常取 $\alpha = 1$），用于处理训练集中未出现的特征组合，防止概率为零。这里的分母是 Number of samples in $C_k + \alpha \times$ number of possible outcomes for x_i，由于 x_i 只有 0 或 1 两种可能，所以是 $\alpha \times 2$。

先验概率 $P(C_k)$ 的估计：

$$P(C_k) = \frac{\text{Number of samples in } C_k}{\text{Total number of samples}}$$

最终，模型会选择使 $P(x|C_k)P(C_k)$ 最大的类别 C_k 作为预测结果。

Scikit-learn 中的实现

在 Scikit-learn 中，伯努利朴素贝叶斯通过 `sklearn.naive_bayes.BernoulliNB` 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。
- `binarize`: 浮点型或 `None`，默认值为 `0.0`。
  - 如果为浮点数，表示一个阈值。输入数据中所有大于此阈值的特征值将被二值化为 1，小于或等于此阈值的将被二值化为 0。
  - 如果为 `None`，则假定输入数据已经是二元的（即特征值只有 0 和 1）。
  - **重要提示**：在实际应用中，通常建议在将数据传递给 `BernoulliNB` 之前，手动对数据进行二值化（例如使用 `sklearn.preprocessing.Binarizer`），以确保特征正确地表示存在性。

适用场景

伯努利朴素贝叶斯最常用于：

- **文本分类**：当特征表示单词是否在文档中出现时（例如，使用二值化的词袋模型）。
- **具有二元特征**的任何分类问题。

优点

- **简单且高效**：训练和预测速度快。
- **适用于二元特征**：对于那些只关心特征是否存在而非频率的应用，它是一个合适的选择。
- **对稀疏数据友好**：特别是在文本数据中，如果使用二值化特征，可以有效处理高维稀疏数据。

缺点

- **“朴素”假设**：特征条件独立性假设在现实中很少完全成立。
- **仅适用于二元特征**：要求输入特征是二元的，对于非二元特征需要进行预处理（如二值化）。
- **对计数信息不敏感**：它丢弃了特征出现的次数信息，这在某些情况下可能导致信息损失。例如，一个单词出现 100 次和出现 1 次被同等对待，这可能不适用于所有文本分类任务。

与其他朴素贝叶斯变体的区别

- `MultinomialNB` **(多项式朴素贝叶斯)**：适用于**计数型离散特征**，如文本分类中的词频。假设特征服从多项式分布。
- `GaussianNB` **(高斯朴素贝叶斯)**：适用于**连续特征**，假设特征服从高斯（正态）分布。
- `ComplementNB` **(补充朴素贝叶斯)**：也是为文本分类设计的变体，尤其适用于**不平衡数据集**。

总之，伯努利朴素贝叶斯是一个简单而有效的分类器，特别适用于处理二元特征数据，其核心在于关注特征的“有无”而非“多少”。