朴素贝叶斯方法

1 模型及其说明

朴素贝叶斯方法

- 贝叶斯公式
  - $$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$
    - 朴素贝叶斯方法就是频率学派在此公式上的一个巧妙的应用，通过最大化后验概率获取某一个确定的分类，在具体使用过程中效果良好
- 1 设输入空间 X ∈ R 为n维向量的集合，输出空间为类标记集合Y= c1, c2,..., 输入为特征向量X ∈ R ,输出为类标记(class label) y ∈Y. X是定义在输入空间x上的随机向量，Y是定义在输出空间D.上的随机变量. P(X, Y)是X和Y的联合概率分布.训练数据集T{(X1, Y1), (X2, Y2), (X3, Y3), ... (Xn, Yn)},在给定的数据集下我们能够求出X和Y的联合概率分布 P(X, Y)
- 2 为了使用贝叶斯公式，我们计算先验概率与似然函数
  - $$P(Y = c_k)(k = 1, 2, \dots, n)$$
  - $$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}|Y = c_k)$$
- 3 所求的联合密度函数就转换为了
  - $$P(X, Y) = \sum_{i=0}^{N} P(X = x|Y = c_k)P(Y = c_k)$$
- 4 在计算条件概率中P(X|Y)是一个非常大的指数级参数，在求取过程中非常复杂。于是，求解过程中给定一个条件独立的假设
  - $$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}|Y = c_k) = \prod_{j=1}^{N} P(X^j = x^j|Y = c_k)$$
- 5 于是模型推导过程如下：
  - 1.1 模型推导

    在已知$X = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$的情况下，需要求得$P(Y = C_k)$的概率,即：$P(X|Y)$. 根据朴素贝叶斯法, $P(X|Y) = \frac{P(X,Y)}{P(Y)}$, 将(1)、(3)、(4)带入得到：

    $$P(Y = C_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{P(Y = c_k)} = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_{j=1}^{n} P(X^k = x^k)P(Y = c_k)} = \frac{P(Y = c_k)\prod_{j=1}^{n} P(X^j = x^j|Y = c_k)}{\sum_k P(Y = c_k)\prod_{j=1}^{n} P(X^j = x^j|Y = c_k)} \quad (5)$$
- 6 但是，在模型中还是过于复杂，对于所有的分类而言，分类器的分母都是相同的，所以模型能够在次简化：
  - 自此(5)所有的值都能通过已知数据获得。于是朴素贝斯分类器(6)可表示为：

    $$y = f(x) = \arg\max_{c_k} \frac{P(Y = c_k)\prod_{j=0}^{n} P(X^j = x^j|Y = c_k)}{\sum_k P(Y = c_k)\prod_{j=1}^{n} P(X^j = x^j|Y = c_k)}$$

    对于所有的分类而言，分类器的分母都是相同的，所以分类器可简化(7)为：

    $$y = f(x) = \arg\max_{c_k} P(Y = c_k)\prod_{j=1}^{n} P(X^j = x^j|Y = c_k)$$
- 7 最大化后验概率
  - 选取0-1损失函数
    - $$L(Y, f(x)) = \begin{cases} 1, Y \neq f(x) \\ 0, Y = f(x) \end{cases}$$
  - 则，经验风险函数可以表示为
    - $$R_{exp}(f) = E[L(Y, f(X))] = E_X \sum_{i=1}^{K} [L(c_k, f(x))]P(c_k|X)$$
  - 在实际过程中我们需要最小化经验风险函数
    - $$\min R_{exp}(f) = \arg\min \sum_{i=1}^{k} L(c_k, y)P(c_k|X) = \arg\min_{c_k} \sum_{i=1}^{k} P(y \neq c_k|X) = \arg\min_{c_k} \sum_{i=1}^{k} 1 - P(y = c_k|X) = \arg\max_{c_k} \sum_{i=1}^{k} P(y = c_k|X)$$
      - 从公式中能够看出最小化经验风险函数，与最大化决策分类函数是一样的，这也就是朴素贝叶斯方法的原理

2 具体应用

- $P(y)$ 的估计：使用频率学派的极大似然估计——训练集中类别的相对频率：

  $$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1, 2, \cdots, K$$

- $P(x_i|y)$ 的估计：形成了不同的朴素贝叶斯模型；

1 GaussianNB 高斯朴素贝叶斯方法

- GaussianNB 高斯朴素贝叶斯

  对于特征分量 x_i 是连续值情形，假设它服从高斯分布。

  概率密度函数：

  $$P(x_i|y_k) = \frac{1}{\sqrt{2\pi\sigma^2_{i;y_k}}} exp(-\frac{(x_i - \mu_{i;y_k})^2}{2\sigma^2_{i;y_k}})$$
- 它的 2 个模型参数，可以使用频率学派的最大似然方法从训练集中进行估计：
  - 均值：$\mu_{i;y_k} = \frac{\sum x_{i;y_k}}{N}$；
  - 方差：$\sigma^2_{i;y_k} = \frac{\sum(x_{i;y_k} - \mu_{i;y_k})^2}{N}$；
- 它们分别表示某个类别 yk 中第 i 个特征 xi 的均值与方差。因为朴素贝叶斯假设这些特征是条件独立的，所以可以把每一个特征分量的分布当作一维高斯分布。

通过已知数据训练得到均值与方法，在预测时只要使用均值与方差就能求出对应的条件概率然后就能通过模型求出其分类了

2 Bernoulli Naive Bayes 伯努利朴素贝叶斯

- 伯努利朴素贝叶斯：
  - $P(x_i|y)$ 的计算公式如下：

    $$P(x_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$
  - 与高斯贝叶斯方法一样先计算条件概率，然后计算分类概率，求最大值就是其分类

3 Multinomial Naive Bayes 多项式朴素贝叶斯

- 该方法是应用于文本分类（Text Classification）的 2 个经典朴素贝叶斯变体之一。- 它的特征数据服从多项式分布，使用单词计数向量（Word Count Vector）来表示输入数据；- 分布参数由参数向量表示：

  $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$

  其中，n 表示特征个数，例如，在文本分类中，n 表示单词容量。θ_{yi} 表示特征 i 出现在属于类别 y 的样本 x 中的概率 $P(x_i|y)$。

  由于特征条件分布的独立性假设，参数 θ_y 的每个分量都可以单独使用最大似然方法进行估算，常用情况下，使用平滑版本：

  $$P(x_i|y) = \hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

  其中，$N_{yi} = \sum_{x \in T} x_i$ 是训练集 T 中，特征 i 出现在类别 y 中的个数，$N_y = \sum_{i=1}^{n} N_{yi}$ 是类别 y 中所有特征的个数总和，n 表示特征 x_i 的取值个数。

  上式中，$\alpha \geq 0$，当 $\alpha = 0$ 时，它等价于最大似然估计。$\alpha > 0$ 可以防止待估计的概率值出现 0 的情况，避免对后续计算产生影响。当 $\alpha = 1$ 时，该平滑公式被称为拉普拉斯平滑（Laplace Smoothing），当 $\alpha < 1$ 时，被称为 Lidstone 平滑。
- 在上述基础上我给出具体的推导过程：
  - 4.1 模型推导 ¶

    假设x是一个离散的变量，取值可能是$1, 2, 3 \dots k$。那么关于x的概率分布的集合可以使用一个参数为向量P的分布表示，其中$P(X = k) = P_K$.另一种写法如下：

    $$p(x|P) = \prod p_k^{\sigma(x=k)} \quad (1)$$

    这里的$\sigma(x = k)$是一个指示函数，当$x = k$时，它的结果为1，否则为0。那么N个独立同分布的样本$X = \{x_1, x_2, \dots, x_n\}$的联合概率分布就是：

    $$p(X|p) = \prod p_k^{c_k}$$
    $$c_k = \sum \sigma(x_i = k)$$

    其实c_k的意思就是观测值是k的变量的个数。构造分类器：

    $$f(x) = \arg\max_p p(X|p) = \prod p_x^{c_k} \quad (2)$$
    $$s.t. \sum_{k=1}^{K} p_x = 1$$

    接下来，我们使用拉格朗日乘数法，求得其解析解， 构造拉格朗日函数：

    $$f(p, \lambda) = \sum c_k log\, p_x + \lambda\left(\sum p_x - 1\right)$$
  - 对p_k与λ分别求偏导得：

    $$\frac{\partial f(p, \lambda)}{\partial p} = \sum \frac{c_k}{p_x} - \sum \lambda = 0$$
    $$\frac{\partial f(p, \lambda)}{\partial \lambda} = \sum p_x - 1 = 0$$

    可以解的：

    $$\lambda = \frac{c_x}{p_x} \quad (3)$$
    $$p_x = \frac{c_x}{\sum c_x}$$

    将(3)带入到(1)就能求得条件概率，于是模型求得方法与高斯朴素贝叶斯方法一样进行。 但是在(3)中，我们能够看出，如果存在一个$c_k = 0$，则求出的条件概率始终为0，显然这个样子是不符合实际情况的，于是我们加以变形：

    $$p_x = \frac{c_x + \alpha}{\sum c_x + n\alpha} \quad (4)$$

    上式中，$\alpha \geq 0$，当 $\alpha = 0$时，它等价于最大似然估计。 $\alpha > 0$ 可以防止待估计的概率值出现0 的情况，避免对后续计算产生影响。当$\alpha = 1$时，该平滑公式被称为拉普拉斯平滑（Laplace Smoothing），当$\alpha < 1$ 时，被称为 Lidstone 平滑。

3 实验验证

1 GaussianNB 高斯朴素贝叶斯方法

- Iris部分数验证
- Iris全数验证

部分验证中，准确率相同，均为100%，在全数据验证中，准确率为97.37%，且实验结果与sklearn的训练效果相同，这表明模型没有问题

```
Gaussian Naive Bayse traing done !
labels probability {0.0: 0.4714285714285714, 1.0: 0.5285714285714286}
The parameters of features' conditional probability: {0.0: [(5.042424242424243, 0.34818168631482066), (3.472727272727273, 0.4084506774665272)], 1.0: [(5.978378378378378, 0.5266596085863139), (2.775675675675676, 0.3208143400156524)]}
Accuracy = 100.00
```

```
Gaussian Naive Bayse traing done !
labels probability {0.0: 0.35714285714285715, 1.0: 0.29464285714285715, 2.0: 0.3482142857142857}
The parameters of features' conditional probability: {0.0: [(5.005, 0.35633551605193664), (3.4375, 0.3685698712591684), (1.4525000000000001, 0.17462459735100322), (0.25, 0.10954451150103323)], 1.0: [(5.954545454545455, 0.5635874856617119), (2.796969696969697, 0.35117538438346996), (4.257575757575758, 0.5093253428476247), (1.3333333333333333, 0.218350103450325 1)], 2.0: [(6.646153846153846, 0.6007884366082042), (2.9871794871794872, 0.31393245000950093), (5.620512820512819, 0.5459 31096985826), (2.051282051282051, 0.2827264645914981)]}
Number of mislabeled points out of a total 38 points : 1
Accuracy = 97.37%
```

2 Bernoulli Naive Bayes 伯努利朴素贝叶斯

***注意，该方法处理特征 xi 的分布服从伯努利分布的情形。**

3 Multinomial Naive Bayes 多项式朴素贝叶斯

4 新闻分类器

本实验在给定的数据集下实现了新闻分类，具体步骤如下:

1. 读取文件，使用分词器`jieba`分词， 得到全数据、训练数据、测试数据三部分
   * 全数据用于特征向量的选取

2. 生成断句文本，在创建特征向量时屏蔽掉一些不需要的文本
   * 如果不讲这些词语屏蔽的话就会引入过多的噪声，这类词语出现频率高，而且对于分类的效果并不好，有' 啊 '、'的'、' 嗯 '等词语，这类词语出现频率高，而且对于分类的效果并不好，而且由于文本的处理，会引入空格、回车等字符，也会影响分类

3. 将训练数据、测试数据向量化，使之能够使用`MultinomialNB`分类， 分类器搭建完成

在测试中并不是特征值选取的越多越高，在最后选取了3500个特征值时能够达到100%的准确率

4 对比与回顾

- 感知机
- KNN
- 朴素贝叶斯

	感知机	KNN	朴素贝叶斯
样本要求	线性可分	无	样本的各个属性间相互独立
模型类别	判别模型	判别模型	生成模型
有明显的训练模型	有	无	有
算法	梯度下降法	KD树搜索	概率计算，没有明确的计算过程
模型评价	效果一般般，不能通过训练次数确定模型的好坏，模型不唯一，模型有待改进。	训练效果较好，KD树的构造耗时长，在数据变化较大的情况下不适用，不能良好的确定K值的大小，只能在已知数据集的情况下估算K,但是这样的K值泛化能力不够强	训练效果好，在不同的情况下使用不同的概率计算公式，得到的效果往往较好。