实验 2 实验报告

朴素贝叶斯

朴素贝叶斯（Naive Bayes） 既可以是一种**算法**——朴素贝叶斯算法，也可以是一种**模型**——朴素贝叶斯分类模型（分类器）。主要用来做分类任务。其理论基础是基于贝叶斯定理和条件独立性假设的一种分类方法。

朴素贝叶斯算法

朴素贝叶斯算法可以直接利用贝叶斯定理来实现也就是将实例分到后验概率最大的类中。这等价于期望风险最小化。这就是朴素贝叶斯法所采用的原理。

简单的朴素贝叶斯分类器

【预测过程】：

- 有一个朴素贝叶斯分类模型（器），它能够区分出 k 个类 (c_1, c_2, …, c_k)$(c1, c2, \cdots, ck)$，用来分类的特征有 n 个 (F_1, F_2, …, F_n)$(F1, F2, \cdots, Fn)$。
- 现在有个样本 s，我们要用 NB 分类器对它做预测，则需要先提取出这个样本的所有特征值 F1 到 Fn，将其带入到下式中进行 k 次运算:

  $$P(C=c_j)\prod_{i=1}^{n}P(F_i=f_i|C=c_j)$$ $P(C=cj) i=1\prod nP(Fi=fi \mid C=cj)$

- 然后比较这 k 次的结果，选出使得运算结果达到最大值的那个 c_j(j=1, 2, …, k)$cj(j=1, 2, \cdots, k)$ —— 这个 cj 对应的类别就是预测值。

【示例】：假设我们当前有一个模型。

- 总共只有两个类别：c1 和 c2；
- 有三个 Feature：F1，F2 和 F3。
  - F1 有两种可能性取值：f11 和 f12；
  - F2 有三种可能性取值：f21，f22，f23；
  - F3 有两种可能性取值：f31，f32。

通过训练过程以获得如下的值：

```
P(C=c1)
P(C=c2)
P(F1=f11|C=c1)
```

```
P(F1=f12|C=c1)
P(F2=f21|C=c1)
P(F2=f22|C=c1)
P(F2=f23|C=c1)
P(F3=f31|C=c1)
P(F3=f32|C=c1)
P(F1=f11|C=c2)
P(F1=f12|C=c2)
P(F2=f21|C=c2)
P(F2=f22|C=c2)
P(F2=f23|C=c2)
P(F3=f31|C=c2)
P(F3=f32|C=c2)
```

将这些概率值都算出来后，即可用来做预测。

比如我们有一个需要预测的样本 X，它的特征值分别是 f11，f22，f31，那么：

- 样本 X 被分为 c1 的概率是：$P(C=c_1|x) = P(C=c_1| F_1=f_{11}, F_2=f_{22}, F_3=f_{31}) \propto P(C=c_1)P(F_1=f_{11} | C=c_1)P(F_2=f_{22} | C=c_1)P(F_3=f_{31} | C=c_1)$ $P(C=c1|x)=P(C=c1|F1=f11,F2=f22,F3=f31)\propto P(C=c1)P(F1=f11|C=c1)P(F2=f22|C=c1)P(F3=f31|C=c1)$
- 样本 X 被分为 c2 的概率是：$P(C=c_2|x) = P(C=c_2| F_1=f_{11}, F_2=f_{22}, F_3=f_{31}) \propto P(C=c_2)P(F_1=f_{11} | C=c_2)P(F_2=f_{22} | C=c_2)P(F_3=f_{31} | C=c_2)$ $P(C=c2|x)=P(C=c2|F1=f11,F2=f22,F3=f31)\propto P(C=c2)P(F1=f11|C=c2)P(F2=f22|C=c2)P(F3=f31|C=c2)$

两者都算出来以后，只需要对比 P(C=c1|x) 和 P(C=c2|x) 谁更大，那么这个样本的预测值就是对应类别。

【总结】：在训练样本的基础上做一系列概率运算，然后用这些算出来的概率按朴素贝叶斯公式“拼装”成分类模型——这就成了朴素贝叶斯分类器。

存在的问题

朴素贝叶斯分类器这个模型的训练过程都不需要先从模型函数推导目标函数，再优化目标函数求 Cost 最小的解吗？朴素贝叶斯公式就是朴素贝叶斯分类器的训练算法？？

之所以这样简单，是因为我们**简单地将频率当成了概率**。但在现实应用中，这种方法往往不可行，因为这种方法实际上默认了“未被观测到”的就是“出现概率为 0”的。这样做显然是不合理的。