朴素贝叶斯分类器

分类问题综述

目标：实现一个分类器f，当输入一个样本$x_i = (x^{(1)}, x^{(2)}, \ldots\ldots x^{(n)})$时，可以给出分类$y_i = f(x_i)$，我们的目的就是为了构造这个分类器$f$。

贝叶斯公式

贝叶斯公式：

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

另一种表示方法：

$$P(\text{类别}|\text{特征}) = \frac{P(\text{特征}|\text{类别})P(\text{类别})}{P(\text{特征})} \tag{2}$$

针对该公式分析：

- $P(Y)$为先验概率，先验概率分布：$P(Y = c_k), k = 1, 2, \ldots K$（分类器输出一共有$K$类，其中取到第$c_k$类的概率）。

  以手写数字识别为例，输出有10类，则$k = 0, 1, 2 \ldots 9$，其中

$$P(Y = c_0) = \frac{\text{训练集标签为}0\text{的图片数量}}{\text{全部图片数量}} \tag{3}$$

- $P(X|Y)$为条件概率，条件概率分布：$P(X = x|Y = c_k) = \{X^{(1)} = x^{(1)}, X^{(2)} = x^{(2)}, \ldots, X^{(n)} = x^{(n)}|Y = c_k\}$(在第k类数据中，取得特征$x$的概率)。

  以手写数字识别为例，$P(X = x|Y = c_0) = \frac{P(X=x,Y=c_0)}{P(Y=c_0)}$表示，在标签为0的数据集中，符合特征$X^i = x^i$的数量，即

$$P(X = x|Y = c_0) = \frac{\text{在训练集}0\text{这一类当中，特征为}x\text{的数量}}{\text{训练集标签为}0\text{的图片数量}} \tag{4}$$

- $P(X)$为证据，$P(X) = \sum_k P(X = x|Y = c_k)P(Y = c_k)$

等式右边的三项均可由数据集计算得到，因此可以求得目标$P(Y|X)$，对应的意义就是给一个样本X，它是类别Y的概率：

$$P(Y = c_k|X = x) = \frac{P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)} \tag{5}$$

我们的目标是使得等式左边这个值尽可能的大，即：给定一个样本x，可以肯确定的判断他是第y_i类，又因为分母是定值，所有得到如下结论：

$$\hat{y} = argmax_{c_k} P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k) \tag{6}$$

即(6)式是我们用于分类的分类器，使用方法为：给一个测试样本x，将$Y = c_k$分别带入不同的c，当$\hat{y}$取得最大时的c_k就是样本x的类别。