朴素贝叶斯与应用

by 寒小阳(hanxiaoyang.ml@gmail.com)

贝叶斯理论简单回顾

在我们有一大堆样本（包含特征和类别）的时候，我们非常容易通过统计得到$p(特征|类别)$.

大家又都很熟悉下述公式：

$$p(x)p(y|x) = p(y)p(x|y)$$

所以做一个小小的变换

$$p(特征)p(类别|特征) = p(类别)p(特征|类别)$$
$$p(类别|特征) = \frac{p(类别)p(特征|类别)}{p(特征)}$$

独立假设

看起来很简单，但实际上，你的特征可能是很多维的

$$p(features|class) = p(f_0, f_1, \ldots, f_n|c)$$

就算是2个维度吧，可以简单写成

$$p(f_0, f_1|c) = p(f_1|c, f_0)p(f_0|c)$$

这时候我们加一个特别牛逼的假设：特征之间是独立的。这样就得到了

$$p(f_0, f_1|c) = p(f_1|c)p(f_0|c)$$

其实也就是：

$$p(f_0, f_1, \ldots, f_n|c) = \Pi_i^n p(f_i|c)$$

贝叶斯分类器

OK，回到机器学习，其实我们就是对每个类别计算一个概率$p(c_i)$，然后再计算所有特征的条件概率$p(f_j|c_i)$，那么分类的时候我们就是依据贝叶斯找一个最可能的类别：

$$p(class_i|f_0, f_1, \ldots, f_n) = \frac{p(class_i)}{p(f_0, f_1, \ldots, f_n)}\Pi_j^n p(f_j|c_i)$$

文本分类问题

下面来看一个文本分类问题，经典的新闻主题分类，用朴素贝叶斯怎么做。