# 6.线性判别分析

## 生成模型和判别模型

1.判别式模型这么做：

我们根据训练数据得到分类函数和分界面，比如说根据SVM模型得到了一个分界面，然后直接计算条件概率 $P(y|x)$ ，我们将最大的 $P(y|x)$ 作为新样本的分类。判别式模型不能反映训练数据本身的特性，能力有限，其只能告诉我们分类的类别。

2.生成式模型这么做

一般我们对每一个类建立一个模型，有多少个类别，我们就建立多少个模型。比方说类别标签有{猫，狗，猪}，那首先根据猫的特征学习出一个猫的模型，再根据狗的特征学习出狗的模型，之后分别计算新测试样本 $x$ 跟三个类别的联合概率 $P(x,y)$ ，然后根据贝叶斯公式：

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

分别计算 $P(y|x)$ ，选择三类中最大的 $P(y|x)$ 作为样本的分类。

我们发现，不管是生成式模型还是判别式模型，它们最终的判断依据都是条件概率 $P(y|x)$ ，但是生成式模型先计算了联合概率 $P(x,y)$ ，再由贝叶斯公式计算得到条件概率。生成式模型可以体现更多数据本身的分布信息，其普适性更广。

最后，两者的联系是：由生成式模型可以得到判别式模型，但由判别式模型不能得到生成式模型。

**做一个总结，判别模型求解的思路是：条件分布------>模型参数后验概率最大------->（似然函数参数先验）最大------->最大似然**

2）现在考虑生成模型。给定输入 $\tilde{x}$ ，生成模型可以给出输入和输出的联合分布 $P(\tilde{x},\tilde{c})$ ，所以生成方法的目标是求出这个联合分布。这里以朴素贝叶斯模型为例，我们要求的目标可以通过：

$$P(\tilde{x},\tilde{c}) = P(\tilde{x}|\tilde{c}) \cdot P(\tilde{c})$$ ----------

这样将求联合分布的问题转化成了求类别先验概率和类别条件概率的问题，朴素贝叶斯方法做了一个较强的假设--------feature的不同维度是独立分布的，简化了类别条件概率的计算，如果去除假设就是贝叶斯网络，这里不再赘述。

以朴素贝叶斯为例，**生成模型的求解思路是：联合分布------->求解类别先验概率和类别条件概率**

## 优缺点比较

生成模型：
优点：
1）生成给出的是联合分布 $P(\tilde{x}, \tilde{c})$，不仅能够由联合分布计算条件分布 $P(\tilde{c}|\tilde{x})$（反之则不行），还可以给出其他信息，比如可以使用 $P(\tilde{x}) = \sum_{i=1}^{k} P(\tilde{x}|\tilde{c}_i) * P(\tilde{c}_i)$ 来计算边缘分布 $P(\tilde{x})$。如果一个输入样本的边缘分布 $P(\tilde{x})$ 很小的话，那么可以认为学习出的这个模型可能不太适合对这个样本进行分类，分类效果可能会不好，这也是所谓的*outlier detection*。
2）生成模型收敛速度比较快，即当样本数量较多时，生成模型能更快地收敛于真实模型。
3）生成模型能够应付存在隐变量的情况，比如混合高斯模型就是含有隐变量的生成方法。

缺点：
1）天下没有免费午餐，联合分布是能提供更多的信息，但也需要更多的样本和更多计算，尤其是为了更准确估计类别条件分布，需要增加样本的数目，而且类别条件概率的许多信息是我们做分类用不到，因而如果我们只需要做分类任务，就浪费了计算资源。
2）另外，实践中多数情况下判别模型效果更好。

判别模型：
优点：
1）与生成模型缺点对应，首先是节省计算资源，另外，需要的样本数量也少于生成模型。
2）准确率往往较生成模型高。
3）由于直接学习 $P(\tilde{c}|\tilde{x})$，而不需要求解类别条件概率，所以允许我们对输入进行抽象（比如降维、构造等），从而能够简化学习问题。

缺点：
1）是没有生成模型的上述优点。

## 常用的模型举例

有两种类型用来决定

$$\vec{w}$$

的线性分类器。第一种模型条件机率

$$P(\vec{x}|\text{class})$$

。这类的算法包括：

- 线性判别分析（LDA） --- 假设为高斯条件密度模型。
- 朴素贝叶斯分类器--- 假设为条件独立性假设模型。

第二种方式则称为判别模型（discriminative models），这种方法是试图去最大化一个训练集（training set）的输出值。在训练的成本函数中有一个额外的项加入，可以容易地表示正则化。例子包含：

- Logit模型---的最大似然估计，其假设观察到的训练集是由一个依赖于分类器的输出的二元模型所产生。
- 感知元（Perceptron） --- 一个试图去修正在训练集中遇到错误的算法。
- 支持向量机--- 一个试图去最大化决策超平面和训练集中的样本间的边界（margin）的算法。