判别模型与生成模型

王裕

一、定义

判别模型：学习得到条件概率分布 P(y|x)，即在特征 x 出现的情况下标记 y 出现的概率。

生成模型：学习得到联合概率分布 P(x,y)，即特征 x 和标记 y 共同出现的概率，然后求条件概率分布。能够学习到数据生成的机制。

理解判别模型与生成模型之间的区别是非常重要的。当分析适用于你的数据集的各种分类器时，你会遇到各种各样的分类器，包括 SVM（支持向量机）、朴素贝叶斯、逻辑斯蒂回归，等等，但你必须了解哪些分类器将最适合你的特定数据集以及你希望从数据分析中得到什么。

二、两个模型基本的区别

判别模型是学得一个分类面（即学得一个模型），该分类面可用来区分不同的数据分别属于哪一类；而生成模型是学得各个类别各自的特征（即可看成学得多个模型），可用这些特征数据和要进行分类的数据进行比较，看新数据和学得的模型中哪个最相近，进而确定新数据属于哪一类。

举个例子：若分类目标是对图像中的大象和狗进行分类。判别方法学得一个模型，这个模型可能是判断图中动物鼻子的长度是否大于某一阈值，若大于则判断为大象，否则判断为狗；生成学习则分别构建一个大象的特征模型与狗的特征模型，来了一个新图像后，分别用大象模型与狗模型与其进行比较，若新图像与狗相似度更高则判断为狗，否则判断为大象。

若已知某分类任务的生成模型，是可以求得该任务的判别模型，反之则不行。这和概率论中的全概率密度函数以及边沿概率密度函数是一致的（即已知全概率密度可求得边沿概率密度，但已知边沿概率密度不能求得全概率密度，$p(y|x)=\frac{p(x,y)}{p(x)}$）。

三、两个模型的具体特点

1、判别模型

判别模型的特点：

1、SVM 和决策树是判别模型，因为它们学习类之间的明确边界。SVM 是一种最大边缘分类器，即在给定核函数的情况下，学习一个使两类样本间距离最大的决策边界。样本与已知的决策边界之间的距离可以使 SVM 成为一个“软”分类器。决策树通过最大化信息增益(或其他标准)的方式递归地划分空间来学习决策边界。

2、判别模型一般不用于离群点检测，而生成模型通常用于离群点检测。当然，最好应基于特定的应用程序进行评估。

3、判别模型不提供数据集的特征和类之间关系的清晰表示。它们不是使用资源对每个类进行完整的建模，而是侧重于对类之间的边界进行丰富的建模。因此，给定相同的容量(例如，计算机程序中的位)，与生成模型相比，判别模型可能生成更复杂的边界表示形式。

4、判别算法允许对点进行分类，而不需要提供实际生成点的模型。所以这些可以是：概率算法试图学习 P(y|x)（例如，逻辑斯蒂回归）或试图直接从点到类学习映射的非概率算法(例如，感知器和 SVM 只是一个提供分离的超平面，但它们没有生成新点的模型)。

另一种思考方法是生成算法对模型做一些结构假设，而判别算法做的假设较少。例如，朴素贝叶斯假设特征具有条件独立性，而逻辑斯蒂回归(与朴素贝叶斯相对应的判别模型)则没有。

总的来说，判别模型比生成模型更强大，因为它对于较大的数据集比对于较小的数据集更有效。也就是说，它们可能会过度拟合较小的数据集。

2、生成模型

典型的几个生成模型有：朴素贝叶斯法、马尔科夫模型、高斯混合模型。这种方法一般建立在统计学和贝叶斯理论的基础之上。

生成模型的特点：

生成算法提供一个模型模拟数据是如何生成的，而判别算法只是提供分类（不一定以概率的方式）。例如，我们来比较高斯混合模型和 k 均值聚类模型。在高斯混合模型中，我们有一个很好的生成点的概率模型(选择一个具有一定概率的组件，然后通过从组件的高斯分布中采样来生成一个点)，但是对于 k 均值

聚类模型，却不是这样的。

1、生成算法对模型做出某种结构假设（例如，朴素贝叶斯假设特征具有条件独立性）。生成模型通常被解释为概率图形模型，它提供了数据集中的独立关系的丰富表示。

2、当你试图强行使一个分类器成为生成分类器时（比如说，逻辑斯蒂回归可以用 P(y|x)和 P(x)来代替 P(x,y)，因此可以成为一个生成分类器），你并没有使用完整的生成模型来做出分类决策。

注意，你可以使判别分类器生成，因为你正在向逻辑斯蒂回归中添加一些尚未存在的东西。也就是说，当你执行一个朴素贝叶斯分类时，你直接计算 P(y|x)∝P(x|y)P(y)(右边的项 P(x|y)和 P(y)是生成新文档的条件)；但是当你在逻辑斯蒂回归中计算 P(y|x)时，你不是在计算这两个东西，你只是把逻辑函数应用到点积上。

1、生成模型在较小的数据集上的应用通常优于判别模型，因为它们的生成假设在模型中设置了一些防止过度拟合的结构。例如，让我们考虑朴素贝叶斯与逻辑斯蒂回归。当然，朴素贝叶斯的假设很少得到满足，因此随着数据集的增大，逻辑斯蒂回归往往会优于朴素贝叶斯(因为它可以捕获朴素贝叶斯无法捕获的相关关系）。

2、根据不同的应用情形，生成模型可能有许多优势。假设你正在处理非平稳分布，其中在线测试数据可能是由训练数据不同的潜在分布生成的。与 SVM 中的决策边界相比，检测分布变化并相应地更新生成模型通常更为简单，尤其是在在线更新需要无监督的情况下。

参考文献：

https://www.cnblogs.com/xiaoshayu520ly/p/9079435.html

https://deveshbatra.github.io/Generative-vs-Discriminative-models

https://blog.csdn.net/u010358304/article/details/79748153