斯坦福 ML 公开课笔记 5

这篇笔记针对的是公开课视频的第五个，主要内容包括生成学习算法（generate learning algorithm）、高斯判别分析（Gaussian Discriminant Analysis，GDA）、朴素贝叶斯（Navie Bayes）、拉普拉斯平滑（Laplace Smoothing）。

生成学习算法

之前的视频中讲到的方法都是直接对问题进行求解，比如二类分类问题，不管是感知器算法还是逻辑斯蒂回归算法，都是在解空间中寻找一条直线从而把两种类别的样例分开，对于新的样例只要判断在直线的哪一侧即可；这种直接对问题求解的方法可以成为判别学习方法（discriminative learning algorithm）。而生成学习算法则是对两个类别分别进行建模，用新的样例去匹配两个模型，匹配度较高的作为新样例的类别，比如良性肿瘤与恶性肿瘤的分类，首先对两个类别分别建模，比如分别计算两类肿瘤是否扩散的概率，计算肿瘤大小大于某个值的概率等等；再比如狗与大象的分类，分别对狗与大象建模，比如计算体重大于某个值的概率，鼻子长度大于某个值的概率等等。

形式化的说，判别学习方法是直接对 p(y|x)进行建模或者直接学习输入空间到输出空间的映射关系，其中，x 是某类样例的特征，y 是某类样例的分类标记。而生成学习方法是对 p(x|y)（条件概率）和 p(y)（先验概率）进行建模，然后按照贝叶斯法则求出后验概率 p(y|x)：

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{1}$$

使得后验概率最大的类别 y 即是新样例的预测值：

$$\underset{y}{\operatorname{argmax}}\, p(y|x) = \underset{y}{\operatorname{argmax}} \frac{p(x|y)p(y)}{p(x)} = \underset{y}{\operatorname{argmax}}\, p(x|y)p(y) \tag{2}$$

高斯判别分析

高斯判别分析（GDA）就是一种生成学习算法，不过比较奇怪的是它的名字里居然有判别两个字，可能会让人误以为它是判别学习方法，不过它却是地地道道的生成学习算法。

在 GDA 中，假设 p(x|y)属于多变量正态分布。多变量正态分布是正态分布在多维变量下的扩展，它的参数是一个均值向量（mean vector）μ 和协方差矩阵（covariance matrix）$\sum \in R^{n*n}$, 其中 n 是多维变量的向量长度，$\sum \in R^{n*n}$是对称

正定矩阵。多变量正态分布的概率密度函数为：

$$\mathrm{p}(\mathrm{x};\mu,\Sigma)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)) \tag{3}$$

其中，$|\Sigma|$是行列式的值。

对于服从多变量正态分布的随机变量 x，均值由下面的公式得到：

$$\mathrm{E}[\mathrm{X}]=\int x\mathrm{p}(\mathrm{x};\mu,\Sigma)\,dx=\mu \tag{4}$$

协方差矩阵由协方差函数 Cov 得到：

$$\mathrm{Cov}(\mathrm{X})=\Sigma$$

其中，cov 的计算过程为：

$$\mathrm{Cov}(\mathrm{Z})=\mathrm{E}[(\mathrm{Z}-\mathrm{E}[\mathrm{Z}])(\mathrm{Z}-\mathrm{E}[\mathrm{Z}])^{T}]=\mathrm{E}[\mathrm{Z}Z^{T}]-(\mathrm{E}[\mathrm{Z}])(E[Z])^{T} \tag{5}$$

接下来，视频展示了几组二元正态分布的概率密度的图形，包括均值为 0，协方差矩阵为单位矩阵的图形，还有改变均值和协方差矩阵时的图形。

介绍完多变量正态分布，就正式进入 GDA 模型的介绍。GDA 模型针对的是输入特征为连续值时的分类问题。这个模型的基本假设是目标值 y 服从伯努利分布，条件概率 p(x|y)服从正态分布。于是，它们的概率密度为：

$$\mathrm{p}(\mathrm{y})=\varphi^{y}(1-\varphi)^{1-y} \tag{6}$$

$$\mathrm{p}(\mathrm{x}|\mathrm{y}=0)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_0)^{T}\Sigma^{-1}(x-\mu_0)) \tag{7}$$

$$\mathrm{p}(\mathrm{x}|\mathrm{y}=1)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu_1)^{T}\Sigma^{-1}(x-\mu_1)) \tag{8}$$

于是，数据集的极大似然函数的对数如下所示：

$$\begin{aligned}\mathrm{L}(\varphi,\mu_0,\mu_1,\Sigma)&=\log\prod_{i=1}^{m}p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma)\\&=\log\prod_{i=1}^{m}p(x^{(i)}|y^{(i)};\phi,\mu_0,\mu_1,\Sigma)p(y^{(i)};\varphi)\end{aligned} \tag{9}$$

对极大似然函数对数最大化，我们就的得到了 GDA 模型的各参数的极大似然估计，即得到了如何使用 GDA 算法的方法。各参数的极大似然估计如下：

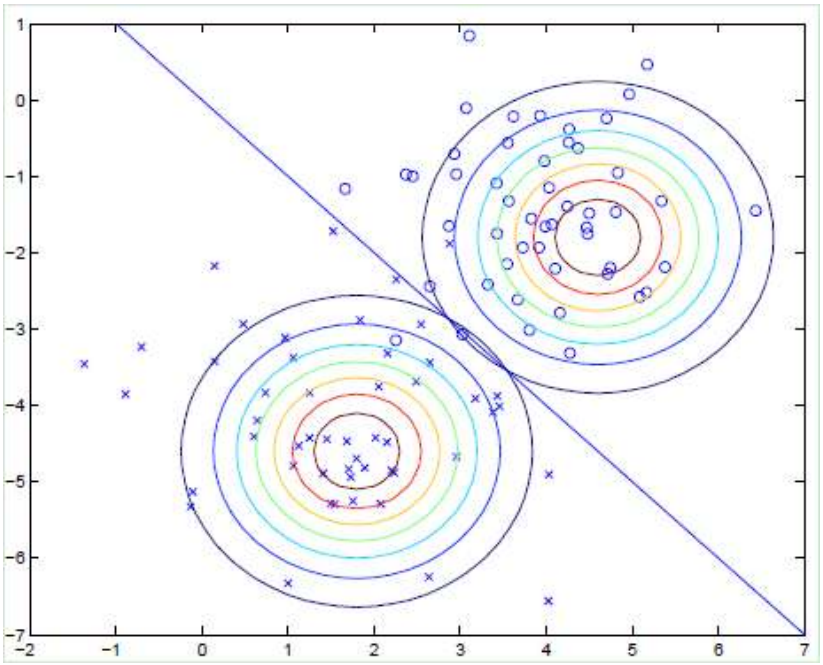
$$\varphi=\frac{1}{m}\sum_{i=1}^{m}I\{y^{(i)}=1\} \tag{10}$$

$$\mu_0=\frac{\sum_{i=1}^{m}I\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{m}I\{y^{(i)}=0\}} \tag{11}$$

$$\mu_1=\frac{\sum_{i=1}^{m}I\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{m}I\{y^{(i)}=1\}} \tag{12}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \quad (13)$$

一个二维的 GDA 模型例子如下图所示：



注意到两个二维高斯分布分别对两类数据进行拟合；它们使用相同的协方差矩阵；但却有不同的均值；在直线所示的部分，p(y=1|x)=p(y=0|x)=0.5。

GDA 模型与 logistic 模型的联系

在公式 2 中，我们使用 p(x|y)p(y)作为 p(y|x)的拟合。归一化后，可以得到：

$$\mathrm{p}(\mathrm{y}=1|\mathrm{x}) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \quad (14)$$

实际上，它可以被表示成逻辑斯蒂分布的形式：

$$\mathrm{p}(\mathrm{y}=1|\mathrm{x};\varphi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+\exp(-\theta^T x)} \quad (15)$$

其中，θ是参数$\varphi,\mu_0,\mu_1,\Sigma$某种形式的函数。GDA 的后验分布可以表示成逻辑斯底分布形式的合理性我没有证明，有兴趣的可以自己证明。

实际上，可以证明，不仅仅当先验概率分布服从多变量正态分布时可以推导出逻辑斯蒂回归模型，当先验分布属于指数分布族中的任何一个分布（比如泊松分布）时都可以推导出逻辑斯蒂回归模型；而反之则不成立，逻辑斯蒂回归模型的先验概率分布不一定必须得是指数分布族中的成员；因而也说明了逻辑斯蒂回归模型在建模上的鲁棒性。

由此，我们得到了推导逻辑斯蒂回归模型的两种方法。第一种是前面的视频

里讲到的通过指数分布族来推导；第二种则是刚才提到的通过生成学习假设先验概率分布的方式进行推导。

那么如何选择 GDA 与逻辑斯蒂回归模型呢？由上面的分析可以知道，GDA 与逻辑斯蒂回归是泛化与特化的关系，GDA 比逻辑斯蒂回归有更多的前置假设。当数据服从或大致服从正态分布时，使用 GDA 会达到更好的效果，因为 GDA 利用了更多的信息构建模型。但是当数据不服从正态分布时，那么逻辑斯蒂回归更有效，因为它做出更少的假设，构建的模型更加强壮，更加具有鲁棒性。生成学习还有另外的一个好处，就是可以使用比判别学习模型更少的数据构建出强壮的模型。

朴素贝叶斯

GDA 针对的是特征向量 x 为连续值时的问题。而朴素贝叶斯（Navie Bayes，NB）则针对的是特征向量 x 为离散值时的问题。

NB 算法的标准应用也是最常见的应用就是文本分类问题，例如邮件是否为垃圾邮件的分类。

同其他分类算法一样，NB 算法也需要有相应的标准好的数据集。对于文本分类问题来说，使用向量空间模型（vector space model，VSM）来表示文本。何为 VSM？首先，我们需要有一个词典，词典的来源可以是现有的词典，也可以是从数据中统计出来的词典，对于每个文本，我们用长度等于词典大小的向量表示，如果文本包含某个词，该词在词典中的索引为 index，则表示文本的向量的 index 处设为 1，否则为 0。

如果按直接对 p(x|y)进行建模，那么会遇到参数过多的问题，我们假设词典里拥有 50000 个词语，即向量长度为 50000，向量中每个分量的取值为{0,1}，那么可能有 2^{50000} 个可能的结果，对其建模则需要 2^{50000}-1 个参数。因而，NB 模型做了另外的假设，成为朴素贝叶斯假设，又朴素贝叶斯假设推导出的分类器成为朴素贝叶斯分类器。

朴素贝叶斯假设即是在给定分类 y 后，假设特征向量中的各个分量是相互独立的。如下式所示：

$$\begin{aligned} \mathrm{p}(x_1, x_2, \dots, x_{50000}|\mathrm{y}) &= \mathrm{p}(x_1|\mathrm{y})\mathrm{p}(x_2|\mathrm{y}, x_1) \dots \mathrm{p}(x_{50000}|x_1, x_2, \dots, x_{49999}) \\ &= \mathrm{p}(x_1|\mathrm{y})\mathrm{p}(x_2|\mathrm{y}) \dots \mathrm{p}(x_{50000}|\mathrm{y}) \end{aligned} \tag{16}$$

朴素贝叶斯假设在文本分类问题上的解释是文本中出现某词语时不会影响其他词语在文本中的概率。

以 VSM 与 NB 假设为基础，我们就得到了 NB 方法的参数：

$$\varphi_y = p(y = 1) \tag{17}$$

$$\varphi_{j|y=1} = \mathrm{p}(x_j = 1|\mathrm{y} = 1) \tag{18}$$

$$\varphi_{j|y=0} = \mathrm{p}(x_j = 1|\mathrm{y} = 0) \tag{19}$$

于是，我们就得到了 NB 方法的极大似然估计的对数函数：

$$\begin{aligned}\mathrm{L}(\varphi_y, \varphi_{j|y=1}, \varphi_{j|y=0}) &= \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) = \prod_{i=1}^{m} p(x^{(i)}|y^{(i)})p(y^{(i)}) \\ &= \prod_{i=1}^{m}\left(\prod_{\mathrm{j}=1}^{n} p(x_j^{(i)}|y^{(i)})\right)p(y^{(i)})\end{aligned} \tag{20}$$

其中，n 为词典的大小。最大化该函数，我们得到参数的极大似然估计：

$$\varphi_{j|y=1} = \frac{\sum_{i=1}^{m} I\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} I\{y^{(i)} = 1\}} \tag{21}$$

$$\varphi_{j|y=0} = \frac{\sum_{i=1}^{m} I\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} I\{y^{(i)} = 0\}} \tag{22}$$

$$\varphi_y = \frac{\sum_{i=1}^{m} I\{y^{(i)} = 1\}}{m} \tag{23}$$

对于新样本，按照如下公式计算其概率值：

$$\begin{aligned}\mathrm{p}(\mathrm{y} = 1|\mathrm{x}) &= \frac{p(x|y=1)p(y=1)}{p(x)} = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \\ &= \frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right)p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right)p(y=1) + \left(\prod_{i=1}^{n} p(x_i|y=0)\right)p(y=0)}\end{aligned} \tag{24}$$

以上就是最基本的 NB 方法。注意到特征向量的每个分量都只能取值{0,1}，我们可以将其扩展为{0,1,2,…,k}，而概率分布由伯努利分布变为多项式分布。对于一些连续的变量，我们可以将其离散化使其可以用 NB 方法解决，离散化的方法为将连续变量按值分段。

拉普拉斯平滑

拉普拉斯平滑（Laplace Smoothing）又被称为加 1 平滑，是比较常用的平滑方法。平滑方法的存在时为了解决零概率问题。

所谓的零概率问题，就是在计算新实例的概率时，如果某个分量在训练集中从没出现过，会导致整个实例的概率计算结果为 0。针对文本分类问题就是当一个词语在训练集中没有出现过，那么该词语的概率为 0，使用连乘法计算文本出现的概率时，整个文本出现的概率也为 0。这显然是不合理的，因为不能因为一个事件没有没观测到就判断该事件的概率为 0。

对于一个随机变量 z，它的取值范围是{1,2,3…,k}，对于 m 次试验后的观测结果$\{z^{(1)}, z^{(2)}, z^{(3)}, \ldots, z^{(m)}\}$，极大似然估计按照下式计算：

$$\varphi_j = \frac{\sum_{i=1}^{m} I\{z^{(i)} = j\}}{m} \tag{25}$$

使用 Laplace 平滑后，计算公式变为：

$$\varphi_j = \frac{\sum_{i=1}^{m} I\{z^{(i)} = j\} + 1}{m + \mathrm{k}} \tag{26}$$

即在分母上加上取值范围的大小，在分子加 1。

回到 NB 算法，我们可以修正各分量的计算公式：

$$\varphi_{j|y=1} = \frac{\sum_{i=1}^{m} I\left\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\right\} + 1}{\sum_{i=1}^{m} I\{y^{(i)} = 1\} + 2} \tag{27}$$

$$\varphi_{j|y=0} = \frac{\sum_{i=1}^{m} I\left\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\right\} + 1}{\sum_{i=1}^{m} I\{y^{(i)} = 0\} + 2} \tag{28}$$