生成学习模型

本文是观看斯坦福大学机器学习公开课第5课以及阅读配套讲义notes2之后所做笔记。本文主要介绍两种生成学习模型——高斯判别分析和朴素贝叶斯。

by Jimmy

2016年3月2日

在notes1中介绍了线性回归和逻辑回归，这两种模型都是所谓的判别学习模型，今天要介绍一类新模型，即生成学习模型。那么这两类模型有什么不同呢？如果我们直接对$p(y \mid x)$进行建模，或者说尝试找到模型假设$h_\theta(x) : x \mapsto \{0, 1\}$（例如感知机），那么这种模型就称为判别学习模型；如果我们对$p(x \mid y)$进行建模，再利用贝叶斯公式计算出给定x下y的概率，即

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

那么就称为生成学习模型。在生成学习模型中，事实上我们对每一类训练样本都建立了一个模型，这个模型可以计算出测试样本的属于该类的概率。此后，我们只要比较训练样本在所有模型中的概率，取概率最大的类别作为分类结果即可。这个过程可以表示为

$$\begin{aligned}\operatorname*{argmax}_y p(y \mid x) &= \operatorname*{argmax}_y \frac{p(x \mid y)p(y)}{p(x)} \\ &= \operatorname*{argmax}_y p(x \mid y)p(y)\end{aligned}$$

1 高斯判别分析

高斯判别分析（Gaussian discriminant analysis，GDA）是一种生成学习模型，在学习这个模型之前让我们先简单的了解一下多元正态分布（多元高斯分布）。

1.1 多元正态分布

现假设『元』的个数为n，即输入x是一个n维向量，且有$x \sim \mathcal{N}(\mu, \Sigma)$。其中$\mu \in \mathbb{R}^n$表示平均向量（mean vector），$\Sigma \in \mathbb{R}^{n \times n}$表示协方差矩阵（covariance matrix）。多元正态分布的密度函数为

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

其中|Σ|表示协方差矩阵的行列式。

1.2 高斯判别分析模型

对于输入特征x为连续值的分类问题，我们可以利用GDA来进行分类。在此之前，我们做出三个基本假设

$$
\begin{aligned}
y &\sim Bernoulli(\phi) \\
x \mid y = 0 &\sim \mathcal{N}(\mu_0, \Sigma) \\
x \mid y = 1 &\sim \mathcal{N}(\mu_1, \Sigma)
\end{aligned}
$$

到此，我们的模型中出现了4个参数ϕ, Σ, μ_0 & u_1。（注：虽然在假设里出现了两个高斯分布，但通常情况下我们令它们拥有一样的协方差矩阵Σ，这样能够保证两个分布的形状相同，只是中心位置不同。）对所有样本有log-likelihood

$$
\begin{aligned}
l(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p\left(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma\right) \\
&= \log \prod_{i=1}^{m} p\left(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma\right) p\left(y^{(i)}; \phi\right)
\end{aligned}
$$

（注：可以看到，在GDA中likelihood是对联合概率$p(x, y)$进行累乘，而在逻辑回归中likelihood是对条件概率$p(y \mid x)$进行累乘。）最大化l解得参数的最大似然估计为

$$
\begin{aligned}
\phi &= \frac{1}{m} \sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\} \\
\mu_0 &= \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} x^{(i)}}{\sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\}} \\
\mu_1 &= \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\} x^{(i)}}{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\}} \\
\Sigma &= \frac{1}{m} \sum_{i=1}^{m} \left(x^{(i)} - \mu_{y^{(i)}}\right)\left(x^{(i)} - \mu_{y^{(i)}}\right)^T
\end{aligned}
$$

1.3 GDA与逻辑回归

GDA和逻辑回归之间有着有趣的联系，如果将$p(y = 1 \mid x)$视作关于x的函数，那么有

$$
p(y = 1 \mid x; \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1 + \exp(-\theta^T x)}
$$

（注：上式可由贝叶斯公式推得。）其中，θ是关于ϕ, Σ, μ_0 & u_1的函数。上式与逻辑回归的模型假设是一致的，这说明可以从GDA得到逻辑回归。（事实上，如果样本服从指数分布族分布，运用生成学习算法都能够得到逻辑回归，然而此过程不可逆。）GDA是一个比逻辑回归条件假设更强的模型，分类结果也比逻辑回归来得更加准确。所以，如果样本服从高斯分布，用GDA的分类效果更好；如果不能够确定样本服从什么分布（例如服从泊松分布），用逻辑回归的鲁棒性更强。

2 朴素贝叶斯

在GDA中输入特征是连续值，如果输入特征是离散值时我们可以利用朴素贝叶斯来进行分类。下面拿朴素贝叶斯运用最为广泛的垃圾邮件分类场景来说明这个算法。

假设我们拥有一个单词表，里面有5000个单词，那么一封邮件可以被转换为一个5000维的向量（先假设邮件中所出现的单词都在单词表中），类似于$x = [1, 0, 1, 0, \cdots, 0]^T$（$x_i \in \{0, 1\}$，0表示单词表的第i个单词在邮件中不出现，1表示出现）。现给出朴素贝叶斯假设（Naive Bayes assumption）：给定y，输入特征x中的每一项x_i是条件独立的。这显然是个不正确的假设，因为这意味着邮件中每个单词出现与否相互独立，但这个假设可以给我们带来一个简单而且效果不错的模型。根据朴素贝叶斯假设，我们可以得到在给定类别下邮件出现的概率

$$\begin{aligned} & p(x_1, x_2, \cdots, x_{5000} \mid y) \\ & \quad = p(x_1 \mid y) p(x_2 \mid y, x_1) \cdots p(x_{5000} \mid y, x_1, x_2, \cdots, x_{4999}) \\ & \quad = p(x_1 \mid y) p(x_2 \mid y) \cdots p(x_{5000} \mid y) \\ & \quad = \prod_{i=1}^{n} p(x_i \mid y) \end{aligned}$$

朴素贝叶斯模型中存在2n+1个参数，即$\phi_{i|y=1} = p(x_i = 1 \mid y = 1)$，$\phi_{i|y=0} = p(x_i = 1 \mid y = 0)$和$\phi_y = p(y = 1)$。给定训练集$\left\{ \left(x^{(i)}, y^{(i)} \right) ; i = 1, 2, \cdots, m \right\}$，那么联合概率的likelihood为

$$L(\phi_y, \phi_{i|y=0}, \phi_{i|y=1}) = \prod_{i=1}^{m} p\left(x^{(i)}, y^{(i)} \right)$$

最大化likelihood的解得参数的最大似然估计为

$$\begin{aligned} \phi_{j|y=1} &= \frac{\sum_{i=1}^{m} 1\left\{ x_j^{(i)} = 1 \wedge y^{(i)} = 1 \right\}}{\sum_{i=1}^{m} 1\left\{ y^{(i)} = 1 \right\}} \\ \phi_{j|y=0} &= \frac{\sum_{i=1}^{m} 1\left\{ x_j^{(i)} = 1 \wedge y^{(i)} = 0 \right\}}{\sum_{i=1}^{m} 1\left\{ y^{(i)} = 0 \right\}} \\ \phi_y &= \frac{\sum_{i=1}^{m} 1\left\{ y^{(i)} = 1 \right\}}{m} \end{aligned}$$

与GDA一样，参数的似然估计和我们对它们的直观认知是一致的。（注：在GDA和朴素贝叶斯中计算出参数的最大似然估计后，我们可直接用于生成学习模型来进行分类。）

2.1 拉普拉斯平滑（Laplace smoothing）

前面我们假设邮件中所有单词都在单词表中，如果一封待分类邮件中出现了不在单词表中的单词，那么会得到$p(y = 1 \mid x) = 0/0$，这个值的大小无法确定，因而我们无法对该邮件进行分类。拉普拉斯平滑可以解决这个问题。假设$z \in \{1, 2, \cdots, k\}$，给定m个样本，应该有$\theta_j = p(z = j)$最大似然估计

$$\phi_j = \frac{\sum_{i=1}^{m} 1\left\{ z^{(i)} = 1 \right\}}{m}$$

在这种情况下θ_j可能等于0，为了避免这一点，应用拉普拉斯平滑得新的估计量

$$\phi_j = \frac{\sum_{i=1}^{m} 1\left\{z^{(i)} = 1\right\} + 1}{m + k}$$

将这一思想运用到到朴素贝叶斯中，参数应该这样确定

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

2.2 多项式模型（multi-variate Bernoulli event model）

在朴素贝叶斯中，每个单词只有两种取值，如果x_i有多种取值（注：这种场景很常见，例如我们对连续值离散化。），我们可以运用一个更一般更有效的模型——多项式模型。给定邮件$x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_1}^{(i)}\right)$（$n_i$表示第i封邮件中单词数目），那么对所有邮件有联合概率的likelihood等于

$$\begin{aligned} L(\phi, \phi_{i|y=0}, \phi_{i|y=1}) &= \prod_{i=1}^{m} p\left(x^{(i)}, y^{(i)}\right) \\ &= \prod_{i=1}^{m} \left(\prod_{j=1}^{n_i} p\left(x_j^{(i)} \mid y^{(i)}; \phi_{i|y=0}, \phi_{i|y=1}\right) \right) p\left(y^{(i)}; \phi_y\right) \end{aligned}$$

解得参数的最大似然估计（with laplace smoothing）为

$$\phi_y = \frac{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\}}{m}$$
$$\phi_{k|y=1} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{x_j^{(i)} = k \wedge y^{(i)} = 1\right\} + 1}{\sum_{i=1}^{m} 1\left\{y^{(i)} = 1\right\} n_i + |V|}$$
$$\phi_{k|y=0} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} 1\left\{x_j^{(i)} = k \wedge y^{(i)} = 0\right\} + 1}{\sum_{i=1}^{m} 1\left\{y^{(i)} = 0\right\} n_i + |V|}$$

其中，|v|表示单词k取值的类别数（#numbers of buckets）。（注：在$\phi_{k|y=1}$的估计量里，分子是单词k出现在垃圾邮件中的次数，分母是垃圾邮件中单词的总数，因此这个估计量表示在垃圾邮件中单词k出现的概率。）