Note for cs229-note2

iSea @ Jan. 5th, 2015

4 生成型学习算法（Generative Learning algorithms）

4.1 判别模型和生成模型

上一节所说的模型有一个共同点，试图直接学习$p(y|x)$，或者直接从输入映射到输出，这些模型都是**判别（discriminative）模型**。比如对数回归学习$p(y|x;\theta)$中的θ，然后求解条件概率。

现在从另一个方向来分析学习问题，来对$p(x|y)$与$p(y)$建模。比如分类问题中的$p(x|y=1)$表示分类到某类时出现某个feature的概率。然后利用贝叶斯定理

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

更正式的

$$\begin{aligned}\arg\max_y p(y|x) &= \arg\max_y \frac{p(x|y)p(y)}{p(x)} \\ &= \arg\max_y p(x|y)p(y)\end{aligned}$$

来求出$\arg\max_y p(y|x)$。这种方法叫做**生成（generative）模型**。

4.2 高斯判别分析

GDA（Gaussian discriminant analysis，高斯判别分析）假定$p(x|y)$满足多值正态分布，也就是多维下的正态分布。其中mean vector $\mu \in \mathbb{R}^n$、convairance matrix（协方差矩阵）$\Sigma \in \mathbb{R}^{n\times n}$，概率密度函数满足

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

其中$|\Sigma|$是Σ的行列式。可以写作$\mathcal{N}(\mu,\Sigma)$，多值正态分布的性质参考Wiki。

对于分类问题，使用多值正态分布来建模

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

求取极大似然值的log值

$$
\begin{aligned}
l(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma) \\
&= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi) \\
&= \sum_{i=1}^{m}\Big(\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)+\log p(y^{(i)};\phi)\Big) \\
&= \sum_{i=1}^{m}\Big[-\frac{1}{2}\,(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})-\frac{n}{2}\log(2\pi) \\
&\qquad -\frac{1}{2}\log|\Sigma^{-1}|+y^{(i)}\log\phi+(1-y^{(i)})\log(1-\phi)\Big]
\end{aligned}
$$

依次求导，得到

$$
\begin{aligned}
\phi &= \frac{1}{m}\sum_{i=1}^{m}1\{y^{(i)}=1\} \\
\mu_k &= \frac{\sum_{i=1}^{m}1\{y^{(i)}=k\}x^{(i)}}{\sum_{i=1}^{m}1\{y^{(i)}=k\}} \\
\Sigma &= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
\end{aligned}
$$

分析一下式子，ϕ就是在训练集上统计出的$y=1$的样本出现的概率，μ_k则分别为两类样本各自的均值，Σ为整个训练集上的样本方差。求出了这些参数，带入式子就可以求出$p(y)$和$p(x|y)$了，也就估测求出$p(y|x)$了。

高斯判别分析中假定$p(x|y)$满足多值正态分布，可以推出$p(y|x)$是满足logistic函数的性质的。但是反之则不然，只要求$p(x|y)$属于Exponential Family。这也意味着高斯判别分析做了更严格的假定，因此在符合假定的数据集上会更加有效，但是对于不符合高斯分布的数据集，就未必有很好的表现了。相比之下指数回归拥有更好的鲁棒性，应用的更多。

4.3 朴素贝叶斯

类似于高斯判别分析，朴素贝叶斯（Naive Bayes）也做了一个看上去有些苛刻的假定，但是这没有妨碍它成为一个非常简洁有效而流行的分类算法。

GDA中的特征向量是连续的，现在来研究一下离散的情况。**文本分类**指根据文本内容将其分类到某种类别，这里的feature就是文本的内容：单词。每个文本可以被表示为一个长度等于字典长度（设为len）的向量，如果包含第i个单词，将该位置为1。于是$x\in\{0,1\}^{len}$，如果希望知道$p(x|y)$的完整概率分布，取值空间是2^{len}。为了简化问题，需要一些假设。

朴素贝叶斯算法是建立在**朴素贝叶斯假设**上的：每个特征是条件独立的。也就是有

$$
p(x_i|y,x_j,x_k)=p(x_i|y)
$$

那么可以得到

$$
\begin{aligned}
p(x_1,\dots,x_n|y) &= p(x_1|y)p(x_2|y,x_1)\dots p(x_n|y,x1,\dots,x_{n-1}) \\
&= \prod_{i=1}^{n} p(x_i|y)
\end{aligned}
$$

注意假设指特征是条件独立的，而不是完全独立的，和$p(x_i|x_j)=p(x_i)$还是有区别的。

有了这个公式，可以计算出

$$
\begin{aligned}
\arg\max_y p(y=k|x) &= \arg\max_y p(x|y=k)p(y=k) \\
&= \arg\max_y \prod_{i=1}^{n} p(x_i|y=k)p(y=k)
\end{aligned}
$$

得到结果最大的分类值，就可以实现文本分类了。

朴素贝叶斯分类也可以扩展到特征取值是连续的情况，将这些特征值离散化，也就是按照区间范围划分变成离散值，再用上面的假设和方法处理。与GDA相比，朴素贝叶斯往往可以得到更好的结果。

4.4 拉普拉斯平滑

如果training data中不包含某个词，那么

$$
\phi(x_p|y=k) = \frac{\sum_{i=1}^{m}\{x_p^{(i)}=1 \wedge y^{(i)}=k\}}{\sum_{i=1}^{m}\{y^{(i)}=k\}} = 0
$$

这样在分类计算$\prod_{i=1}^{n} p(x_i|y=k)p(y=k)$时就都是0了，无法分类。

为了解决这个情况，假设feature取值空间为V，维度为$|V|$，加入先验概率$1/|V|$来均衡。当training data中包含该feature的样本逐渐变大后，先验影响也会逐渐减小。也就是

$$
\phi(x_p|y=k) = \frac{\sum_{i=1}^{m}\{x_p^{(i)}=1 \wedge y^{(i)}=k\}+1}{\sum_{i=1}^{m}\{y^{(i)}=k\}+|V|}
$$

这就是**Laplace平滑**。