规则化和模型选择（Regularization and model selection）

JerryLead
csxulijie@gmail.com
2011 年 3 月 24 日星期四

1 问题

模型选择问题：对于一个学习问题，可以有多种模型选择。比如要拟合一组样本点，可以使用线性回归($y = \theta^T x$)，也可以用多项式回归($y = \theta^T x^{1\sim m}$)。那么使用哪种模型好呢（能够在偏差和方差之间达到平衡最优）？

还有一类参数选择问题：如果我们想使用带权值的回归模型，那么怎么选择权重 w 公式里的参数τ？

形式化定义：假设可选的模型集合是$M = \{M_1, M_2, \dots, M_d\}$，比如我们想分类，那么 SVM、logistic 回归、神经网络等模型都包含在 M 中。

1 交叉验证（Cross validation）

我们的第一个任务就是要从 M 中选择最好的模型。
假设训练集使用 S 来表示
如果我们想使用经验风险最小化来度量模型的好坏，那么我们可以这样来选择模型：

1、 使用 S 来训练每一个M_i，训练出参数后，也就可以得到假设函数h_i。（比如，线性模型中得到θ_i后，也就得到了假设函数$h_\theta(x) = \theta^T x$）
2、 选择错误率最小的假设函数。

遗憾的是这个算法不可行，比如我们需要拟合一些样本点，使用高阶的多项式回归肯定比线性回归错误率要小，偏差小，但是方差却很大，会过度拟合。因此，我们改进算法如下：

1、 从全部的训练数据 S 中随机选择 70%的样例作为训练集S_{train}，剩余的 30%作为测试集S_{CV}。
2、 在S_{train}上训练每一个M_i，得到假设函数h_i。
3、 在S_{CV}上测试每一个h_i，得到相应的经验错误$\hat{\varepsilon}_{S_{\text{CV}}}(h_i)$。
4、 选择具有最小经验错误$\hat{\varepsilon}_{S_{\text{CV}}}(h_i)$的$h_i$作为最佳模型。

这种方法称为 hold-out cross validation 或者称为简单交叉验证。

由于测试集是和训练集中是两个世界的，因此我们可以认为这里的经验错误$\hat{\varepsilon}_{S_{\text{CV}}}(h_i)$接近于泛化错误（generalization error）。这里测试集的比例一般占全部数据的 1/4-1/3。30%是典型值。

还可以对模型作改进，当选出最佳的模型M_i后，再在全部数据 S 上做一次训练，显然训练数据越多，模型参数越准确。

简单交叉验证方法的弱点在于得到的最佳模型是在 70%的训练数据上选出来的，不代表在全部训练数据上是最佳的。还有当训练数据本来就很少时，再分出测试集后，训练数据就太少了。

我们对简单交叉验证方法再做一次改进，如下：

1、 将全部训练集 S 分成 k 个不相交的子集，假设 S 中的训练样例个数为 m，那么每一个子集有 m/k 个训练样例，相应的子集称作$\{S_1, S_2, \dots, S_k\}$。

2、 每次从模型集合 M 中拿出来一个M_i，然后在训练子集中选择出 k-1 个$\{S_1, S_2, S_{j-1}, S_{j+1} \dots, S_k\}$（也就是每次只留下一个$S_j$），使用这 k-1 个子集训练$M_i$后，得到假设函数$h_{ij}$。最后使用剩下的一份$S_j$作测试，得到经验错误$\hat{\varepsilon}_{S_j}(h_{ij})$。

3、 由于我们每次留下一个S_j（j 从 1 到 k），因此会得到 k 个经验错误，那么对于一个M_i，它的经验错误是这 k 个经验错误的平均。

4、 选出平均经验错误率最小的M_i，然后使用全部的 S 再做一次训练，得到最后的h_i。

这个方法称为 k-fold cross validation（k-折叠交叉验证）。说白了，这个方法就是将简单交叉验证的测试集改为 1/k，每个模型训练 k 次，测试 k 次，错误率为 k 次的平均。一般讲 k 取值为 10。这样数据稀疏时基本上也能进行。显然，缺点就是训练和测试次数过多。

极端情况下，k 可以取值为 m，意味着每次留一个样例做测试，这个称为 leave-one-out cross validation。

如果我们发明了一种新的学习模型或者算法，那么可以使用交叉验证来对模型进行评价。比如在 NLP 中，我们将训练集中分出一部分训练，一部分做测试。

2 特征选择（Feature selection）

特征选择严格来说也是模型选择中的一种。这里不去辨析他们的关系，重点说明问题。假设我们想对维度为 n 的样本点进行回归，然而，n 可能大多以至于远远大于训练样例数 m。但是我们感觉很多特征对于结果是无用的，想剔除 n 中的无用特征。n 个特征就有2^n种去除情况（每个特征去或者保留），如果我们枚举这些情况，然后利用交叉验证逐一考察在该情况下模型的错误率，太不现实。因此需要一些启发式搜索方法。

第一种，前向搜索：

1、 初始化特征集 F 为空。

2、 扫描 i 从 1 到 n，

如果第 i 个特征不在 F 中，那么将特征 i 和 F 放在一起作为F_i（即$F_i = F \cup \{i\}$）

在只使用F_i中特征的情况下，利用交叉验证来得到F_i的错误率。

3、 从上步中得到的 n 个F_i中选出错误率最小的F_i，更新 F 为F_i。

如果 F 中的特征数达到了 n 或者预设定的阈值（如果有的话），那么输出整个搜索过程中最好的 F，没达到转到 2

前向搜索属于 wrapper model feature selection。Wrapper 这里指不断地使用不同的特征集来测试学习算法。前向搜索说白了就是每次增量地从剩余未选中的特征选出一个加入特征集中，待达到阈值或者 n 时，从所有的 F 中选出错误率最小的。

既然有增量加，那么也会有增量减，后者称为后向搜索。先将 F 设置为{1,2,..,n}，然后每次删除一个特征，并评价，直到达到阈值或者为空，然后选择最佳的 F。

这两种算法都可以工作，但是计算复杂度比较大。时间复杂度为$O(n+(n-1)+(n-2)+\cdots+1)=O(n^2)$。

第二种，过滤特征选择（Filter feature selection）：

过滤特征选择方法的想法是针对每一个特征x_i，i 从 1 到 n，计算x_i相对于类别标签y的信息量S(i)，得到 n 个结果，然后将 n 个S(i)按照从大到小排名，输出前 k 个特征。显然，这样复杂度大大降低，为 O(n)。

那么关键问题就是使用什么样的方法来度量S(i)，我们的目标是选取与 y 关联最密切的一些x_i。而 y 和x_i都是有概率分布的。因此我们想到使用互信息来度量S(i)，对于x_i是离散值的情况更适用，不是离散值，将其转变为离散值，方法在第一篇《回归认识》中已经提到。

互信息（Mutual information）公式：

$$\mathrm{MI}(x_i,y)=\sum_{x_i\in\{0,1\}}\sum_{y\in\{0,1\}}p(x_i,y)\log\frac{p(x_i,y)}{p(x_i)p(y)}.$$

当x_i是 0/1 离散值的时候，这个公式如上。很容易推广到x_i是多个离散值的情况。

这里的$p(x_i,y)$，$p(x_i)$和$p(y)$都是从训练集上得到的。

若问这个 MI 公式如何得来，请看它的 KL 距离（Kullback-Leibler）表述：

$$\mathrm{MI}(x_i,y)=\mathrm{KL}\left(p(x_i,y)||p(x_i)p(y)\right)$$

也就是说，MI 衡量的是x_i和 y 的独立性。如果它俩独立（$p(x_i,y)=p(x_i)p(y)$），那么 KL 距离值为 0，也就是说x_i和 y 不相关了，可以去除x_i。相反，如果两者密切相关，那么 MI 值会很大。在对 MI 进行排名后，最后剩余的问题就是如何选择 k 值（前 k 个x_i）。我们继续使用交叉验证的方法，将 k 从 1 扫描到 n，取最大的 F。不过这次复杂度是线性的了。比如，在使用朴素贝叶斯分类文本的时候，词表长度 n 很大。使用 filter 特征选择方法，能够增加分类器的精度。

3 贝叶斯统计和规则化（Bayesian statistics and regularization）

题目有点绕，说白了就是要找更好的估计方法来减少过度拟合情况的发生。

回顾一下，线性回归中使用的估计方法是最小二乘法，logistic 回归是条件概率的最大似然估计，朴素贝叶斯是联合概率的最大似然估计，SVM 是二次规划。

以前我们使用的估计方法是最大似然估计（比如在 logistic 回归中使用的）：

$$\theta_{\mathrm{ML}}=\arg\max_{\theta}\prod_{i=1}^{m}p(y^{(i)}|x^{(i)};\theta).$$

注意这里的最大似然估计与维基百科中的表述

http://zh.wikipedia.org/wiki/%E6%9C%80%E5%A4%A7%E5%90%8E%E9%AA%8C%E6%A6%82%E7%8E%87

有些出入，是因为维基百科只是将样本（观察数据）记为 X，然后求 P(X)的最大概率。然而，对于我们这里的样本而言，分为特征 x 和类标签 y。我们需要具体计算 P(X)。在判别模型（如 logistic 回归）中，我们看待 P(X)=P(x,y)=P(y|x)P(x)，而 P(x)与θ独立无关，

因此最后的 argmax P(X)由 argmaxP(y|x)决定，也就是上式θ_{ML}。严格来讲θ_{ML}并不等于样本 X 的概率，只是 P(X)决定于θ_{ML}，θ_{ML}最大化时 P(X)也最大化。在生成模型，如朴素贝叶斯中，我们看待 P(X)=P(y)P(x|y)，也就是在某个类标签 y 下出现特征 x 的概率与先验概率之积。而 P(x|y)在 x 各个分量是条件独立情况下可以以概率相乘方式计算出，这里根本没有参数θ。因此最大似然估计直接估计 P(x,y)即可，变成了联合分布概率。

在该上式中，我们视参数θ为未知的常数向量。我们的任务就是估计出未知的θ。

从大范围上说，最大似然估计看待θ的视角称为频率学派（frequentist statistics），认为θ不是随机变量，只是一个未知的常量，因此我们没有把$p(y^{(i)}|x^{(i)};\theta)$写成$p(y^{(i)}|x^{(i)},\theta)$。

另一种视角称为贝叶斯学派（Bayesian），他们看待θ为随机变量，值未知。既然θ为随机变量，那么θ不同的值就有了不同的概率p(θ)（称为先验概率），代表我们对特定的θ的相信度。我们将训练集表示成$S=\{(x^{(i)},y^{(i)})\}$，i 从 1 到 m。我们首先需要求出θ的后验概率：

$$\begin{aligned} p(\theta|S) &= \frac{p(S|\theta)p(\theta)}{p(S)} \\ &= \frac{\left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)\right)p(\theta)}{\int_\theta \left(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)p(\theta)\right)d\theta} \end{aligned} \tag{1}$$

这个公式的推导其实比较蹊跷。第一步无可厚非，第二步中先看分子，分子中p(S|θ)最完整的表达方式是$(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(x^{(i)})$。由于在分母中也会出现$p(x^{(i)})$，所以$p(x^{(i)})$会被约掉。当然作者压根就没有考虑$p(x^{(i)})$，因为他看待 P(S)的观点就是 x->y，而不是(x,y)。再来看分母，分母写成这种形式后，意思是对所有的θ可能值做积分。括号里面的意思是$\prod_{i=1}^{m} p(y^{(i)}|x^{(i)})$，然后将其展开成分母的模样，从宏观上理解，就是在求每个样例的概率时，先以一定的概率确定θ，然后在$x^{(i)}$和θ的作用下再确定$y^{(i)}$的概率。而如果让我推导这个公式，我可能会这样写分母$p(S)=\int_\theta (p(S|\theta)p(\theta))d\theta$，这样推导出的结果是$p(S)=\int_\theta (\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(\theta)d\theta$。我不知道自己的想法对不对，分歧在于如何看待$\theta$，作者是为每个样例都重新选定$\theta$，而我是对总体样本选择一个$\theta$。

后记：我看了 Andrew NG 的教学视频，发现视频上的结果和讲义上的不一致，应该讲义上由于笔误写错了，正确的分母是$p(S)=\int_\theta (\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(\theta)d\theta$

$p(y^{(i)}|x^{(i)},\theta)$在不同的模型下计算方式不同。比如在贝叶斯 logistic 回归中，

$$p(y^{(i)}|x^{(i)},\theta)=h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta(x^{(i)}))^{(1-y^{(i)})},$$

其中$h_\theta(x^{(i)})=1/(1+\exp(-\theta^T x^{(i)}))$，p 的表现形式也就是伯努利分布了。

在θ是随机变量的情况下，如果新来一个样例特征为 x，那么为了预测 y。我们可以使用下面的公式：

$$p(y|x,S)=\int_\theta p(y|x,\theta)p(\theta|S)d\theta$$

p(θ|S)由前面的公式得到。假若我们要求期望值的话，那么套用求期望的公式即可：

$$E[y|x,S]=\int_y yp(y|x,S)dy$$

大多数时候我们只需求得p(y|x, S)中最大的 y 即可（在 y 是离散值的情况下）。

这次求解p(y|x, S)与之前的方式不同，以前是先求θ，然后直接预测，这次是对所有可能的θ作积分。

再总结一下两者的区别，最大似然估计没有将θ视作 y 的估计参数，认为θ是一个常数，只是未知其值而已，比如我们经常使用常数 c 作为 y=2x+c 的后缀一样。但是$p(y^{(i)}|x^{(i)};\theta)$的计算公式中含有未知数θ。所以再对极大似然估计求导后，可以求出θ。

而贝叶斯估计将θ视为随机变量，θ的值满足一定的分布，不是固定值，我们无法通过计算获得其值，只能在预测时计算积分。

然而在上述贝叶斯估计方法中，虽然公式合理优美，但后验概率p(θ|S)很难计算，看其公式知道计算分母时需要在所有的θ上作积分，然而对于一个高维的θ来说，枚举其所有的可能性太难了。

为了解决这个问题，我们需要改变思路。看p(θ|S)公式中的分母，分母其实就是 P(S)，而我们就是要让 P(S)在各种参数的影响下能够最大（这里只有参数θ）。因此我们只需求出随机变量θ中最可能的取值，这样求出θ后，可将θ视为固定值，那么预测时就不用积分了，而是直接像最大似然估计中求出θ后一样进行预测，这样就变成了点估计。这种方法称为最大后验概率估计（Maximum a posteriori）方法

Θ估计公式为

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta)p(\theta).$$

θ_{MAP}与θ_{ML}一样表示的是P(S)，意义是在从随机变量分布中以一定概率p(θ)选定好θ后，在给定样本特征$x^{(i)}$上$y^{(i)}$出现的概率积。

但是如果让我推导这个公式的时候，我会这么做，考虑后验概率p(θ|S)，我们的目标是求出最有可能的θ。而对于θ的所有值来说，分母是一样的，只有分子是不同的。因此argmax $p(\theta|S) = arg\max_{\theta}(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(x^{(i)})$。也就是$\theta_{MAP}$的推导式。但这个公式与上面的有些不同，同样还是看待每个样本一个θ，还是总体样本一个θ的问题。

与最大似然估计对比发现，MAP 只是将θ移进了条件概率中，并且多了一项p(θ)。一般情况下我们认为$\theta \sim N(0, \tau^2 I)$，实际上，贝叶斯最大后验概率估计相对于最大似然估计来说更容易克服过度拟合问题。我想原因是这样的，过度拟合一般是极大化$p(y^{(i)}|x^{(i)};\theta)$造成的。而在此公式中多了一个参数θ，整个公式由两项组成，极大化$p(y^{(i)}|x^{(i)},\theta)$时，不代表此时p(θ)也能最大化。相反，θ是多值高斯分布，极大化$p(y^{(i)}|x^{(i)},\theta)$时，p(θ)概率反而可能比较小。因此，要达到最大化θ_{MAP}需要在两者之间达到平衡，也就靠近了偏差和方差线的交叉点。这个跟机器翻译里的噪声信道模型比较类似，由两个概率决定比有一个概率决定更靠谱。作者声称利用贝叶斯 logistic 回归（使用θ_{MAP}的 logistic 回归）应用于文本分类时，即使特征个数 n 远远大于样例个数 m，也很有效。