学习理论

1、基本概念

2、PAC理论

3、VC维

4、极大似然，最大后验概率，贝叶斯估计

5、模型评估与评价指标

6、模型诊断调参

一、基本概念

机器学习三定义：

1、计算机系统能够利用经验提高自身的性能

2、学习就是一个基于经验数据的函数估计问题

3、提取重要模式、趋势、并理解数据，从数据中学习

引用吾师的话：“三个定义各有侧重点，但都强调经验或数据的重要性”。一般而言，我们常用第二个定义，即基于经验数据的函数估计问题，形式化的给出了机器学习就是一个函数估计问题。但也强调了数据的重要，无数据巧妇难为无米之炊。所以，机器学习中必备三要素，数据、代码、论文。

机器学习常用的分类：

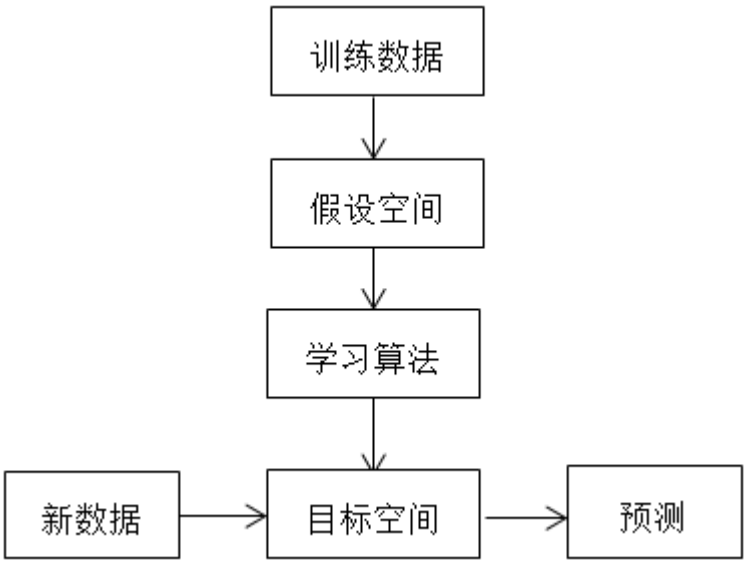
1、监督：样本数据 有标签--分类，回归

2、无监督：样本数据无标签--聚类、异常

3、弱监督：样本数据的标签信息较“弱”--半监督、在线、强化

以上是机器学习最常用的一种基于任务的分类方式，而不同类型下面又有各种算法，有的有千丝万缕的关系，有的却迥然不同，所以很多时候我们不能直接评判那种模型好，只能说某种某些适合某种场景，也不存在一种模型包打天下（深度学习好像是个特例，当然深度学习下面已经衍生了各种算法）。**正是因为我们很难从任务上对比模型，所以笔者就自己所了解的模型从模型的角度进行了对比，分析对比了模型之间的关系，以及适应场景。**

一般而言，无论是监督，无监督，弱监督，都可以形式化如下过程：



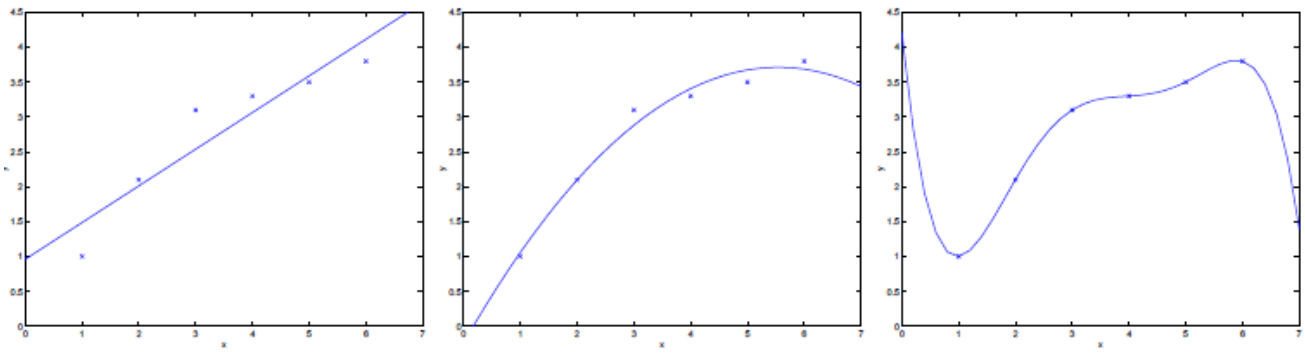
其中，假设空间是我们假设的目标函数空间集（线性，非线性）以及选择目标函数的策略（经验风险，结构风险），通过常用的学习算法（梯度下降，EM算法，坐标下降算法）学习到的最优解即为目标空间，然后当新来数据，依据最优的模型进行预测（分类，回归，聚类）。另外，我们称训练数据为输入空间，一般会做一些特征工程的操作将数据从输入空间映射到特征空间（线性，非线性），然后通过目标函数映射到输出空间。

二、PAC理论

概率近似正确（PAC）理论是从概率的角度来衡量模型的正确率，给出了PAC可辨识，样本复杂度界，误差上界。

偏差/方差

偏差和方差是机器学习中很重要的两个概念，在分析模型时对应于欠拟合和过拟合问题。



以回归问题为例，上图中左边为一个线性拟合，可以看出，拟合的程度不够（**欠拟合**），与真实样本的偏差较大，右边的图类似于插值曲线，基本上每个点都拟合的过好（**过拟合**），然而我们的训练集只是样本数真实分布的一个子集，并不代表所有的样本（测试集）都能拟合的很好，一般而言，由于右图模型复杂度较高，往往泛化能力不如简单的模型。而中间的图拟合的程度和模型的复杂度都不错，因此，机器学习中更倾向于中间的模型最优。

经验风险最小

经验风险最小一直以来都是我们构建目标函数的一个准则，以二分类为例，经验风险最小就是使得误判的样本数最少，对于数据集：

$$S = \{x^{(i)}, y^{(i)}\}, 0 \leq i \leq m, y \in \{0, 1\}$$

其中，样本点$(x^{(i)}, y^{(i)})$独立同分布。

假设，我们学习一个模型来进行分类：

$$h_\theta(x) = g(\theta^T x)$$
$$g(z) = I\{z \geq 0\}, g \in \{0, 1\}$$

其中h是一个线性函数，g是一个指示函数，这样我们就有了一个二分类器。

那么，训练误差即为：

$$\hat{\epsilon}(h_\theta) = \hat{\epsilon}_S(h_\theta) = \frac{1}{m}\sum_{i=1}^{m} I\{h_\theta(x^{(i)}) \neq y^{(i)}\}$$

经验风险最小化准则就是最小化训练误差：

$$\hat{\theta} = arg\min_{\theta} \hat{\epsilon}_S(h_\theta)$$

然而，我们发现如上目标函数非凸，一般无法直接优化，而且这样定义目标函数得到最好的模型在真实数据上并不一定测试误差就最小。为了解决优化的问题，采用对数损失，指数损失，Higne损失，线性损失来代替$0-1$损失。

也就是说我们根据最小化经验损失，从h_θ的假设空间H中学习我们的目标空间：

$$\hat{h} = arg\min_{h \in H} \hat{\epsilon}(h)$$

上式，只是我们在训练集上的最小损失，泛化到测试集：

$$\epsilon(h) = P_{x,y\sim D}(h(x) \neq y)$$

也就是说$\hat{\epsilon}$为训练集上的损失，ϵ为泛化到测试集上的损失。当然，我们希望不学习测试数据就能学到测试集上泛化误差最小的h^*是最好的（不切实际）。

Hoeffding不等式，联合上界，一致收敛

假设$\{x^{(1)}, x^{(2)}, .., x^{(i)}, .., x^{(m)}\}$独立同分布，服从伯努利分布：

$$P(x^{(i)} = 1) = \phi, P(x^{(i)} = 0) = 1 - \phi$$

从伯努利分布角度看，当样本趋于无限大时，所有点的均值为ϕ。从样本统计来看，由于它们之间独立同分布，所以有均值为：

$$\hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

Hoeffding不等式的定义为对任意固定值$\gamma > 0$，存在：

$$P(|\hat{\phi} - \phi| > \gamma) < 2exp(-2\gamma^2 m)$$

该引理表示一个随机变量其偏离期望大于γ的概率有上限。可以从高斯分布出发其到均值距离大于γ的概率有上限（切比雪夫不等式）。 注意，上述不等式是针对一维的情况。

联合上界，假设有n个随机变量$\{x_1, x_2, .., x_j, .., x_n\}$，这$n$个随机变量可以相互独立也可以不独立，我们有：

$$P(x_1, x_2, .., x_n) \leq P(x_1) + P(x_2) +, .., + P(x_n)$$

该不等式很容易理解，即所有事件并集发生的概率小于所有事件发生的概率之和，当且仅当n个事件互斥，等号成立。现在，我们将$|\hat{\phi_j}-\phi_j|>\gamma_j$记为事件$x_j$，那么有$P(x_j)\leq 2exp(-2\gamma_j^2m)$，使用联合上界将其推广到$n$维，我们有：

$$\sum_{j=1}^{n}P(|\hat{\phi_j}-\phi_j|>\gamma_j)\leq\sum_{i=1}^{n}2exp(-2\gamma_j^2m)$$

假设γ_j取统一值γ,那么有：

$$P(|\hat{\phi}-\phi|>\gamma)\leq 2nexp(-2\gamma^2m)$$

上式的意义在于，说明当样本数目m增大时，我们对参数的估计就越逼近真实值。

一致收敛，定义模型的假设空间为:

$$H=\{h_1,h_2,..h_k..,h_N\}$$

首先，我们假设对于所有的h来说，存在训练误差为$\hat{\epsilon}(h_k)=\frac{1}{m}\sum_{i=1}^{m}I(h_k(x^{(i)})\neq y^{(i)})$，$\epsilon$是定义在测试集上的泛化误差，然后我们证明对于任意一个h_k泛化误差ϵ存在上限。

对于h_k，泛化误差$\epsilon(h_k)$是一个以$\hat{\epsilon}(h_k)$为均值服从伯努利分布（分类问题）的随机变量（向量）。由Hoeffding不等式在n维的推广，我们有：

$$P(\forall|\epsilon(h_k)-\hat{\epsilon}(h_k)|>\gamma)\leq 2Nexp(-2\gamma^2m),k=1,2..N$$

也就是说对于假设空间中任意一个模型h_k都满足上式，也就表明不存在一个模型的误差离训练误差的偏差大于一个上限：

$$P(-\exists|\epsilon(h_k)-\hat{\epsilon}(h_k)|>\gamma)\geq 1-2Nexp(-2\gamma^2m),k=1,2..N$$
$$P(|\epsilon(h_k)-\hat{\epsilon}(h_k)|\leq\gamma)\geq 1-2Nexp(-2\gamma^2m),k=1,2..N$$
$$p(|\epsilon(h_k)-\hat{\epsilon}(h_k)|\leq\gamma)\geq 1-\sigma$$

上式表示，任意一个假设空间下的模型h_k的泛化误差都存在上界，这个上界就是定义在偏差上的方差内。由此导出**PAC可辨识**，即从假设空间学习到的模型的误差$\hat{\epsilon}(h_i)$泛化到测试集上的误差$\epsilon(h_i)$的偏差在γ以内的概率大于$1-\sigma$。

样本复杂度界

由$P(|\epsilon(h_k)-\hat{\epsilon}(h_k)|\leq\gamma)\geq 1-2Nexp(-2\gamma^2m)$，为了保证概率大于$1-\sigma$，我们可以分析出至少需要多少样本：

$$1-2Nexp(-2\gamma^2m)\geq 1-\sigma$$
$$\Rightarrow m\geq\frac{1}{2\gamma^2}log\frac{2N}{\sigma}$$
$$\Rightarrow\gamma\geq\sqrt{\frac{1}{2m}log\frac{2N}{\sigma}}$$

由此，我们我们分析出了模型需要达到一定的准确率，需要的样本数目称为样本复杂度。同时我们分析出了在给定m,σ时，模型h_i的泛化误差与m成反比，与n成正比，其中m为样本数目，n表示模型的复杂度。也就是说负杂的模型泛化误差界越大。**注意：由于界的条件很宽，所以得出的界具备参考的价值不大，更多时候是直观的理解，需要样本数的大小与复杂度成正比，与误差范围成反比。**

误差上界

上面我们分析的是同一个模型$\hat{h} = h_k = arg\min_{h \in H} \hat{\epsilon}(h)$的训练误差和泛化误差的关系。但是我们更关心的是训练集上最好模型的泛化误差$\epsilon(\hat{h})$与测试集上最好模型的泛化误差$\epsilon(h^*)$的关系。因为，我们的终极目标是$\epsilon(h^*)$，但是$\epsilon(h^*)$是永远未知的，我们最优模型还是$\epsilon(\hat{h})$，所以我们需要用$\epsilon(h^*)$来定义训练最优模型$\hat{h}$的上界。

$$\hat{h} = arg \min_{h \in H \sim \hat{D}} min(\hat{\epsilon}(h))$$
$$h^* = arg \min_{h \in H \sim D} min(\epsilon(h))$$

其中$\hat{\epsilon}(\hat{h})$表示训练数据集上最好的训练误差，$\epsilon(h^*)$表示测试集上最好的泛化误差。

$$\begin{aligned} \epsilon(\hat{h}) &\leq \hat{\epsilon}(\hat{h}) + \gamma \\ &\leq \hat{\epsilon}(h^*) + \gamma \\ &\leq \epsilon(h^*) + 2\gamma \end{aligned}$$

第一个不等式：对于在训练误差最小的假设类$\hat{h}$，其泛化误差小于训练误差加γ，由一致收敛定理。

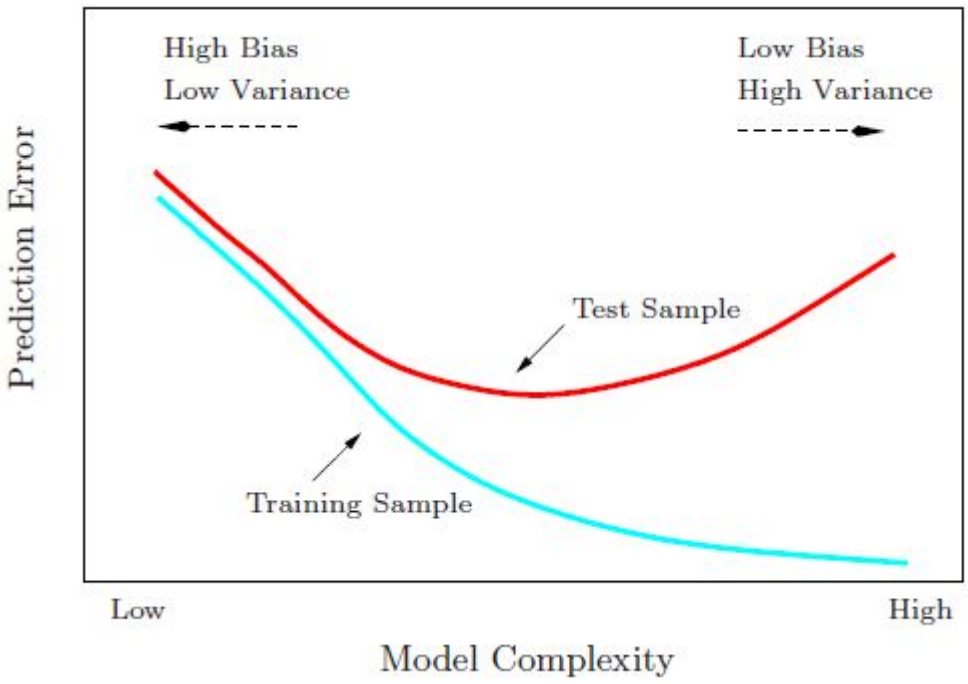
第二个不等式：$\hat{h}$为训练集上误差最小的模型，那么必然有$\hat{\epsilon}(\hat{h}) \leq \hat{\epsilon}(h^*)$。

第三个不等式，对于在测试误差最小的假设类h^*，其训练误差小于泛化误差加γ，由一致收敛定理。

所以我们学习的最好模型的误差$\hat{\epsilon}(\hat{h})$距离我们在测试集上最好模型的误差存在上界：

$$\epsilon(\hat{h}) \leq \epsilon(h^*) + 2\sqrt{\frac{1}{2m} log \frac{2N}{\sigma}}$$

上式表明，我们学习的目标模型的误差服从一个偏差为$\epsilon(h^*)$，方差为$2\sqrt{\frac{1}{2m} log \frac{2N}{\sigma}}$的分布。当我们采用复杂的假设空间来拟合数据时，偏差也许会小，但是使得第二项大。这就指导了我们在选择模型时既要考虑偏差，也要照顾到方差。一般而言，训练误差与测试误差存在如下趋势：

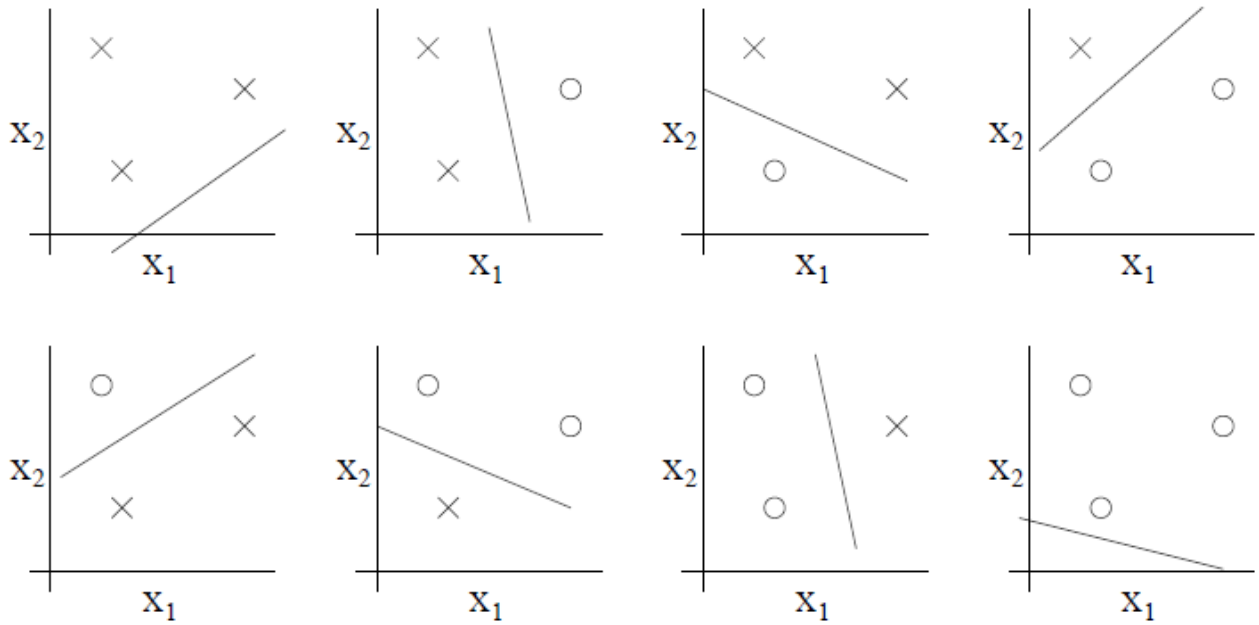


三、VC维

在PAC理论中，我们用假设空间的取值N来描述模型的复杂度，然而很多时候假设空间的取值是无限的，比如线性模型中模型属于连续空间，我们无法用取值来衡量模型的复杂度，VC维的主要价值在于用VC维（维度）衡量模型的复杂度，同时给出了误差上界（个人见解）。

VC维：给定一个样本集$S = \{x^1, x^2, \dots, x^m\}$，我们称假设空间$H$可以打散$S$，当且仅当对于样本集$S$的任何一种标签（与样本的分布无关）都能被$H$线性可分。一般来说： 等于假设类的参数个数

比如下图：一个二维的假设空间其最大能打散（线性可分）的样本集数为3，VC维为3。



对于假设空间H，其$VC(H) = d$，那么至少以$1-\sigma$的概率，对于假设空间下h，我们有：

$$P\left(|\epsilon(h_k) - \hat{\epsilon}(h_k)| \leq \sqrt{\frac{d}{m}log\frac{m}{d} + \frac{1}{m}log\frac{1}{\sigma}}\right) \geq 1 - \sigma$$

也就是说至少$1-\sigma$的概率有，泛化误差与训练误差的方差满足下式：

$$\epsilon(\hat{h}) \leq \hat{\epsilon}(h^*)| + \sqrt{\frac{d}{m}log\frac{m}{d} + \frac{1}{m}log\frac{1}{\sigma}}$$

其中$\hat{h} = arg\min_{h\in H\sim\hat{D}} min(\hat{\epsilon}(h))$,$h^* = arg\min_{h\in H\sim D} min(\epsilon(h))$。

上式可以看作是PAC理论误差分析的VC版，而VC维中定义的假设空间的复杂度是定义在VC维上的，一般而言，VC维与模型的参数有关，特征的维度。

四、极大似然，最大后验概率，贝叶斯估计

假设空间是目标函数所在的一个空间集，我们在选择假设空间时往往需要偏差/方差。在假设空间确定之后，我们需要定义一个合适的策略从假设空间中选择最优的目标函数（模型），上面提到过**经验风险最小化**准则：

$$\min_{h\in H}\frac{1}{m}\sum_{i=1}^{m}L(h_\theta(x^{(i)}),y^{(i)})$$

其中$L(h_\theta(x^{(i)}),y^{(i)})$表示我们定义的损失函数，比如最直接的$0-1$损失，对数损失，指数损失，Hinge损失，线性损失。

另外，考虑到偏差/方差权衡问题，我们都知道复杂的模型泛化误差较大，因此为了防止过拟合（说白了就是我们不能完全信任数据集），我需要对模型的复杂度进行限制，引入正则项。提出了一种**结构风险最小化**准则：

$$\min_{h\in H}\left(\frac{1}{m}\sum_{i=1}^{m}L(h_\theta(x^{(i)}),y^{(i)})+\lambda J(\theta)\right)$$

其中$J(\theta)$用于衡量模型的复杂度，在VC维中，我们知道模型的复杂度与参数有关，所以$J(\theta)$是一个关于参数的函数。比如常用的范数。

以上两条准则都是我们的经验准则，以及我们的损失函数，正则项的定义皆出于我们的经验。下面，我们从概率的角度来分析经验风险最小和结构风险最小。

由机器学习第二定义，我们知道机器学习是一个根据经验数据的函数参数估计问题，转化为概率表示即为$P(\theta|X)$，其中θ为假设空间中的参数，由贝叶斯公式有：

$$P(\theta|X)=\frac{P(X|\theta)P(\theta)}{P(X)}$$

其中$P(\theta|X)$为后验概率，$P(X|\theta)$是似然函数，$P(\theta)$为先验概率，$P(X)$为证据，其中监督问题$X=(X,Y)$。

最大似然估计

最大似然估计可以看作是统计学派的观点，即参数分布由数据集分布确定，即在特定数据集下，参数是固定的，用数据说话，所以，参数估计问题就是最大化似然函数：

$$\max_{\theta}\sum_{i=1}^{m}p(x_i|\theta)$$

对应的，二分类问题中，我们的似然函数为：

$$\begin{aligned}&p\left(y=1|x;\theta\right)=h_\theta(x)\\&p\left(y=0|x;\theta\right)=1-h_\theta(x)\\&p\left(y|x;\theta\right)=(h_\theta(x))^{y}.(1-h_\theta(x))^{(1-y)}\end{aligned}$$

其中我们用一个线性函数来表示在数据x_i,θ的条件下$y=1$的概率为$h_\theta(x)$。

最大似然估计有：

$$\begin{aligned}\max L(\theta)&=p\left(y|x;\theta\right)\\&=\prod_{i=1}^{m}p\left(y^{(i)}|x^{(i)};\theta\right)\\&=\prod_{i=1}^{m}\left(h_\theta(x^{(i)})\right)^{y^{(i)}}.\left(1-h_\theta(x^{(i)})\right)^{\left(1-y^{(i)}\right)}\\&\Leftrightarrow\max logL(\theta)\\&\Leftrightarrow\min -logL(\theta)=\sum_{i=1}^{m}-\left(y^{(i)}log(h_\theta(x^{(i)}))+(1-y^{(i)})log(1-h_\theta(x^{(i)}))\right)\end{aligned}$$

由此，我们就导出了Logistic回归。

最大后验概率估计

贝叶斯学派认为，我们已知的数据只是在服从某一分布的参数的某种情况下的一组有限数据，也就是说，在数据确定时，参数同样是随机变量服从一个先验分布。

$$
\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{P(X)} \\
&\approx P(X|\theta)P(\theta)
\end{aligned}
$$

对应的，回归问题中，假设$P(X|\theta)$服从高斯分布，$P(\theta)$服从高斯分布（伯努利分布），那么后验概率：

$$
\begin{aligned}
p\left(y^{(i)}|x^{(i)};\theta\right) &= \frac{1}{\sqrt{2\pi}}exp\left(-\frac{\left(y^{(i)}-\theta^T x^{(i)}\right)^2}{2\sigma_1^2}\right) \\
p(\theta) &= \frac{1}{\sqrt{2\pi}}exp\left(-\frac{(w-0)^2}{2\sigma_2^2}\right)
\end{aligned}
$$

也就是说，似然函数的均值为$h_\theta = \theta^T x$，而先验分布的均值为0（为了模型尽可能的简单）。

最大后验概率有：

$$
\begin{aligned}
\max R(\theta) &= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)p(\theta) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}}exp\left(-\frac{\left(y^{(i)}-\theta^T x^{(i)}\right)^2}{2\sigma_1^2}\right)\frac{1}{\sqrt{2\pi}}exp\left(-\frac{(w-0)^2}{2\sigma_2^2}\right) \\
\max logR(\theta) &= \sum_{i=1}^{m} \log\frac{1}{\sqrt{2\pi}}exp\left(-\frac{\left(y^{(i)}-\theta^T x^{(i)}\right)^2}{2\sigma_1^2}\right)\frac{1}{\sqrt{2\pi}}exp\left(-\frac{(w-0)^2}{2\sigma_2^2}\right) \\
&= L(\theta)+J(\theta) \\
&= mlog\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma_1^2}\cdot\sum_{i=1}^{m}\left(y^{(i)}-\theta^T x^{(i)}\right)^2 + mlog\frac{1}{\sqrt{2\pi}} - \frac{1}{2\sigma_2^2}\cdot\sum_{i=1}^{m}(\theta)^2 \\
&\Leftrightarrow \min\frac{1}{2\sigma^2}\cdot\sum_{i=1}^{m}\left(y^{(i)}-\theta^T x^{(i)}\right)^2 + \lambda\sum_{i=1}^{m}(\theta)^2 = L(\theta)+J(\theta)
\end{aligned}
$$

由此，我们得出线性回归的岭回归,对应的先验分布服从伯努利分布可以导出Lasso回归。

贝叶斯估计

贝叶斯估计是建立在完整的贝叶斯公式上的，不存在最大后验概率的近似。

$$
\begin{aligned}
P(\theta|X) &= \frac{P(X|\theta)P(\theta)}{P(X)} \\
&= \frac{P(X|\theta)P(\theta)}{\int_\theta p(X|\theta)p(\theta)d_\theta}
\end{aligned}
$$

其中，分母表示是在所有θ可能的取值下对应X的分布的积分。在最大后验概率中，我们认为分母的意义就在于归一化，对后验概率的分布无影响。

可以看出，贝叶斯估计是“先验知识”+“样本信息”=“后验概率” 。由此，我们发现机器学习中损失函数加正则项的解释完全可以从概率角度进行推导。

五、模型评估与评价指标

一、模型评估

在不同的假设空间下，依据各自的准则选择出最优模型后（学习），往往需要对这些模型进行评估。一般而言，把训练数据划分为训练集-验证集-测试集。

训练集：用来训练不同模型，获得模型及其训练误差；

验证集：与训练集相对独立，获取训练模型在该集上的预测误差，用来做模型选择；

测试集：与训练集和验证集独立，获得一般误差和其他模型评价指标，用来评价已选择出的模型。

常用的验证方法有:

交叉验证法（hold-out cross validation）：

1、随机的分割训练样本S为S_{train} 和S_{cy}，一般70和30，前者为训练集，后者为验证/测试集集。

2、为每一类模型M_i在训练集S_{train}上学习，每个假设类得到一个目标函数h_i。

3、对这些目标函数在验证样本上进行验证，得出泛化误差

K-fold cross validation：

1随机的把样本S分为k份，得到了训练子集$S_1, S_2, .., S_k$ 。

2对于每一个模型M_i，从1到k选择留下一份S_j作为验证集，其余的作为训练集。进行k次训练得到k个训练误差$\hat{\epsilon_{ij}}, j = 1..k$，，然后在测试集上进行测试得到泛化误差$\epsilon_{ij}, j = 1, .., k$ ，求均值作为该模型的验证误差。

3、选出模型M_i中验证误差最小的

一般情况下 较常用，也称十重交叉验证。特别的在样本数非常少时，当$k = m$时，称为留一法leave-one-out cross validation。

二、评价指标

平均平方根误差（RMSE）,平均平方误差（MSE）,平均绝对值误差（MAE）

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|h_\theta(x^{(i)} - y^{(i)}\right|$$

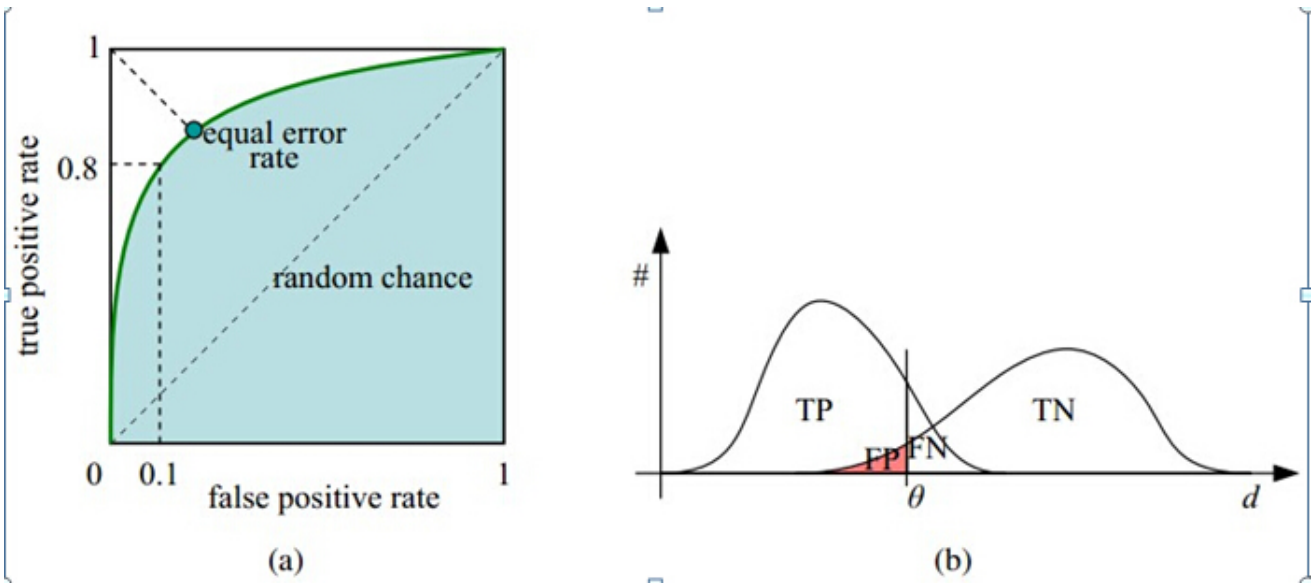
精度（Accuracy ）召回率（Recall）精确率（Precision） F1

真实值/预测值	1	0	
1	TP（真正例）	FN（假反例）	Recall=TP/(TP+FN)实际为1
0	FP（假正例）	TN（真反例）	
	Precision=TP/(TP+FP)预测为1		Accuracy=(TP+TN)/(TP+FN+FP+TN)

$$F1 = \frac{2PR}{P+R} = \frac{2*TP}{2*TP+FN+FP}$$

受试者工作特征曲线 (ROC：receiver operating characteristic curve)与曲线下方面积(AUC:area under curve)

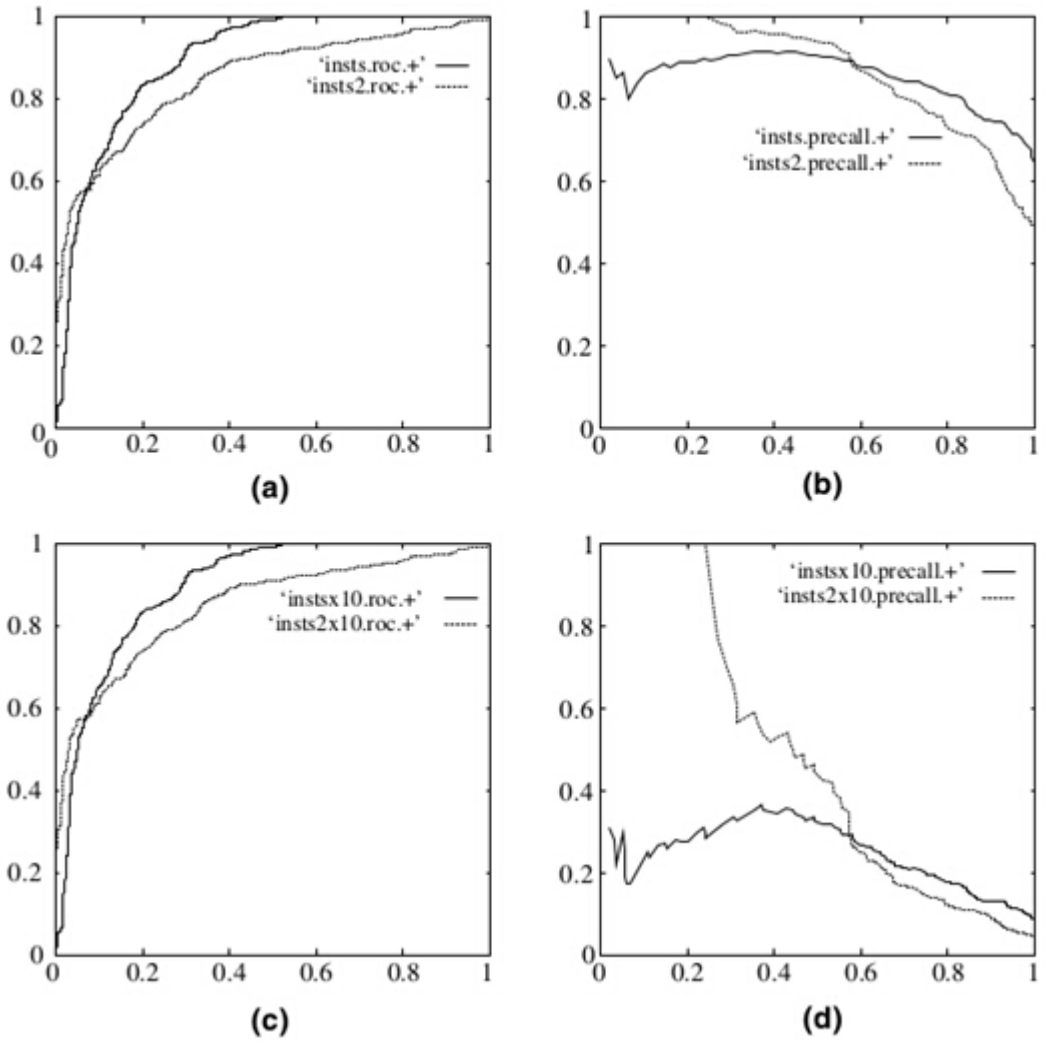
我们都知道在分类问题上确定最终的分类是通过一个指示函数（阈值函数），设置一个阈值进行分类。不同的阈值决定不同的召回率和精确率，因此，ROC曲线是在多组阈值下描述召回率和精确率的曲线，如下图：



曲线的横坐标表示误分类为正类占所有负例的比例，纵坐标表示正确分类为正类占所有正例的比例（这样的好处，由于单独比上各类的总样本数，不会因为数据不均衡问题导致在不同测试集上曲线大变样）。曲线的含义是，当我们不断的调整阈值识别更多的正例时，不可避免的引入了负例误判为正类。

AUC值则表示曲线下方的面积，面积越大则表示在调整阈值是引入负例的概率较小，也就是说曲线面积越大，分类性能越好。

下图是ROC曲线和Presision-Recall曲线的对比：



在上图中，a)和c)为Roc曲线，b)和d)为Precision-Recall曲线。a)和b)展示的是分类其在原始测试集(正负样本分布平衡)的结果，c)和d)是将测试集中负样本的数量增加到原来的10倍后，分类器的结果，可以明显的看出，ROC曲线基本保持原貌，而Precision-Recall曲线变化较大。

DCG(Discounted Cumulative gain)与NDCG(Normalize DCG):两个指标是信息检索下常用的指标，按检索排序计算得分，每个检索排序对应真实排序有一个得分r，将模型前K个检索结果的得分累加。

$$DCG@K = \sum_{i=1}^{K} \frac{2^{r_i} - 1}{log(i+1)}$$

$$NDCG = \frac{DCG@K(r_i)}{DCG@K(r_j)}$$

其中r_i表示模型检索排序的对应第i个结果的得分，r_j表示真实排序下对应第j个结果的得分。可以发现，最好的模型检索结果即为与真实排序一致，那么DCG值最大，规范化之后为1。

MAP（Mean Average Precision）：信息检索下的评价指标，MAP与DCG不同之处在于对检索结果不考虑排序，而是考虑平均的精确率（每个结果的权重为1）。设定一组检索阈值（比如说检索结果排在前K位分类为正例），对应阈值下会有一个精确率后取均值。

以上，在进行模型选择时，往往在多组数据集上进行测试，同时需要综合一些评价指标，以及特定的需求（比如某些场景下更看重召回率，而精确率却不是特别重要），权衡选择出做好的模型。

六、模型诊断与调参

一、快速搭建

1、数据集准备

2、特征工程

3、模型选择

4、模型评价

在我们开发过程中，我们往往都是摸着石头过河，我们不知道数据应该是什么样子，不知道数据的特征，数据集多少合适，如何进行特征工程，选择什么样的模型最合适，最后我们采用什么指标来评价模型。对于一系列模糊的过程，我们很难确定各个模块该如何处理，如何优化，所以我们需要快速搭建一个模型，通过结果去分析如何优化问题。

二、偏差/方差分析

可以说，决定模型最后性能的就是偏差和方差，如果模型在测试集上很好，在训练集上不好，那么很有可能是模型方差过大（模型过拟合，前提是训练集和测试集分布大致一致）。如果模型在训练集上不好，那么意味着模型偏差过大（模型欠拟合）。针对这两种情况，采用控制变量法去调优模型。

一般而言，在模型调优的过程有如下选择：

1）更多的数据集

2）特征工程

3）模型参数调优

4）优化算法调优

5）换模型，换优化算法

6）数据集分析

如何定位出模型的问题，就需要分析模型在训练集和测试集上误差来大致确定是偏差问题还是方差问题。

偏差问题：

1）优化算法：是否收敛，学习率是否合适，迭代次数是否合适，是否需要换优化算法

2）模型：模型参数选择是否合适，模型的表示能力是否更强，是否需要换模型

3）特征工程：特征选择和特征提取是否做的不够

4）数据是否不够

方差问题：

1）测试集和训练集分布是否一致

2）模型是否加强正则项

3）优化算法是否可以提前收敛

其中最难是不同的模型有不同的调优方式，尤其是参数多的模型，一般采用控制变量的方法固定其他不变来调整其中一个来调优。对于不同的模型，还有特殊的处理技巧，比如深度学习，各种超参，技术都会影响性能。

[1] NG 机器学习PDF材料

[2] 统计学习方法

[3] 机器学习西瓜书

[4] 机器学习公理化

[5] 统计自然语言处理

[6] 模式识别

[1]: