一、基本概念

- 监督学习关于数据的**基本假设**：监督学习假设输入和输出的随机变量X和Y遵循联合概率分布$P(X,Y)$。在学习过程中，**假定这一联合概率分布存在，但是目前是未知的。训练数据与测数据被看作是联合概率分布$P(X,Y)$独立同分布产生的。**
  - $P(X,Y)$表示分布函数，或者分布密度函数
  - 独立同分布(iid, independently identically distribution)。指在随机过程中，任何时刻的值都为随机变量，服从统一分布且互相独立。独立指：随机变量X_1的取值并不会影响随机变量X_2的取值，X_2的取值也不影响X_1的取值
- 假设空间：假设空间指的是学习的范围，即由输入空间到输出空间所有可能的映射/模型所组成的空间
- Semi-supervised learning - 半监督学习指的是利用少量标注数据和大量未标注数据进行**辅助标注数据**，进行监督学习，以较低的成本达到较好的学习效果
- Active learning - 机器不断主动给出实例让教师进行标注
- 参数化模型与非参数化模型：参数化模型参数维度固定，非参数化模型参数维度不固定。**参数化模型适合问题简单的情况，现实中问题往往比较复杂，参数维度较大，适用于非参数化模型**
- 在线学习（Online Learning）与批量学习（Batch Learning）：在线学习一次接受一个样本，批量学习一次接受所有数据
- 经验风险最小化（ERM）和结构风险最小化（SRM）：监督学习问题本质就变成了经验风险或结构风险函数的最优化问题
  - 理论损失/期望损失：即假设我们知道输入X和输出Y的联合概率分布时，我们可以计算出Loss Function的理论损失
  - 经验损失/经验风险：$R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}Loss(y_i,f(x_i))$。根据**大数定律**，当样本容量$n\to+\infty$时，**经验损失趋于理论损失**。因此，当数据量较大时，经验风险最小化效果优秀
  - 结构损失/结构风险/正则化：$R_{srm}(f)=R_{emp}(f)+\lambda J(f)$，「在经验风险的基础上加上了一个正则化项」，$J(f)$代表模型复杂度，$\lambda\geq 0$是系数。可以看出，如果模型复杂度较大，则容易导致**overfitting**，此时经验损失较小，但并不代表我们的模型优秀。这时我们用$\lambda J(f)$约束模型复杂度，让模型复杂度不至于过大，从而防止**overfitting**的现象
- 正则化/SRM：正则化符合奥卡姆剃刀原理，即在所有可能选择的模型中，能够很好地解释已有数据并且十分简单复杂度较低的模型才是最好的模型

二、疑难解释

> 条件概率分布$P(y|x)$最大化后得到函数，函数$y=f(x)$归一化后得到条件概率分布

1. 条件概率分布：对于输入x和输出y，如果我们使用概率模型建模，可以得到$P(y|x)$。在这个概率分布中，我们取最大值，则可以得到一个确定的输出$\hat{y}$，所以这里实际上建立了输入到输出的一种确定性函数，即：$\hat{y}=max(P(y|x))=f(x)$
   - 例如：对于MNIST数据集，我们最后会输出输入属于每一个类别的概率。比如是数字5的概率为0.9，是数字4的概率为0.05，则如果我们规定概率最大的结果即为答案，则成功的将一个概率分

布转换为了一个确定性函数。

2. 函数：函数归一化指将函数的输出通过某种标准化运算归一到某个范围中。归一化后的函数在$(-\infty,+\infty)$的积分为1。
    - 例如：Model所代表的决策函数的结果通常是有限点集，我们可以将这些点集归一化后结果落到了[0,1]区间，则我们就可以对其进行概率建模。

> 贝叶斯估计和极大似然估计的比较

- 对概率看法不同的两大派别：频率学派和贝叶斯派
    - 频率学派：频率学派认为世界是确定的，也就是说如果事件在多次重复实验中趋于一个稳定的值p，那么这个p就是该事件的概率。从而诞生了频率学派的参数估计方法，**极大似然估计（MLE）**
    - 贝叶斯派：贝叶斯派认为世界是不确定的，我们对某个事件建模需要对它有一个预先的估计（先验概率），然后不断获取信息来调整之前的估计。这种方法在先验假设较为靠谱的情况下效果显著，且随着数据量的增加，先验假设对于模型参数的主导作用会逐渐削弱，而真实数据对模型的指导作用会大大提高。在数据量较大的情况下，先验假设的作用远不如真实数据对模型的指导作用，此时贝叶斯估计就近似等同于最大似然估计
- 两种参数估计方式
    - 极大似然估计：假设模型待估计参数θ是确定存在的，只是目前是未知的。当满足$\theta = \hat{\theta}_{mle}$时，该组观测样本更容易被观测到，也就是说，$\hat{\theta}_{mle}$是使事件发生的可能性最大的值。
        - 举例来讲：对于一个抛硬币的事件，假设我们未知硬币向上的概率，我们经过不断的观测发现硬币向上的次数是总次数的一般，所以我们认为0.5就是该模型的参数。（因为我们观测到的值就是向上向下各自一半，其他概率如0.4，0.6都没有办法很好地描述我们的观测值）
    - 贝叶斯估计：假设模型待估计参数θ是随机的，和一般的随机变量是没有区别的，所以我们无法估计θ的值，我们只能估计θ的分布。因此，根据贝叶斯公式

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

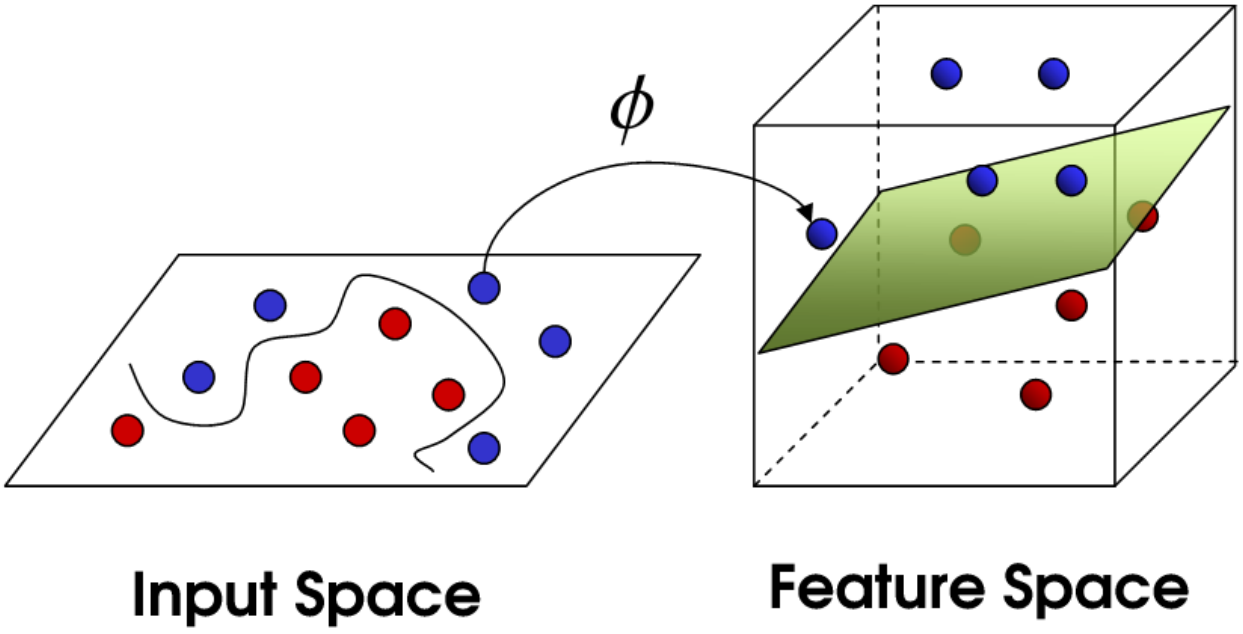
$P(\theta)$是参数θ的先验分布，表明了我们对待估计参数的主观认识，是我们人为定义的而并非从样本中学习到的。

$P(X|\theta)$是似然函数，表明在参数θ遵循我们假定的先验分布的前提下，观察到X的概率。也就是说，如果我们的先验函数定义的不好，则$P(X|\theta)$的值就会较小，从而导致$P(\theta|X)$较小，这里主要起到了一个修正作用。

$P(X)$是一个定值，因为X是给定的，所以我们一般并不关注坟分母

$P(\theta|X)$是我们期望求解的值，即给定数据集X的情况下，我们估计到的模型参数θ的分布

> 核方法是使用**核函数**表示和学习非线性模型的一种机器学习方法，核函数指的是映射之后在特征空间的内积



- 核函数的基本思路：基于这样一种假设，”在低维空间中**不能线性分隔**的点集，转化到高维空间中很有可能变为**线性可分**的。例如上图中，在Input Space中我们的数据集显然无法进行线性分隔，但是经过映射$\phi()$后，在高维空间中我们就可以使用一个超平面来分隔数据集
- 常规的方法与核函数的方法：
    1. 常规的方法：寻找满足条件的，从Input Space到Feature Space的映射ϕ，在映射完成后，通过在高维空间中（通过内积）比较两者的相似度来进行分割。例如，对于输入x_1, x_2，我们期望找到一种$\phi(x_1), \phi(x_2)$，使它们在Feature Space中可以有效分割。分割的基础建立在向量相似度，内积公式$< \phi(x_1), \phi(x_2) >$上
    2. 核函数：在核函数中我们并不对空间映射ϕ建模，也就是说我们不关注高维空间的具体形式，我们也不关心如何映射得到对应的Feature Space。我们直接对最后的内积结果建模，即$K(x_1, x_2) =< \phi(x_1), \phi(x_2) >$，**这样可以简化计算，并且让我们无需关心高维度空间的形式**
- 核函数的存在性和如何构造：**Mercer定理**

> 泛化误差上界推导 - Generalization Error Bound

- 泛化误差（generalization error），用来描述Model的泛化能力，以Model对未知数据预测的误差来定义

$$R_{exp}(\hat{f}) = E_P(Loss(Y, \hat{f}(X))) = \int_{X \times Y} Loss(y, \hat{f}(x)) P(x, y) dxdy$$

- 如何衡量泛化误差：模型的泛化能力分析往往是通过分析**泛化误差的上界**来进行的，也就是说，通常我们比较两种学习方法的**泛化误差的上界**来比较他们的优劣
- 泛化误差上界的**推导**：

> **定理1.1（泛化误差上界）** 对二分类问题，当假设空间是有限个函数的集合$F=(f_1,f_2,\cdots,f_d)$时，对于任意一个函数$f\in F$，至少以概率$1-\delta,0<\delta<1$，以下不等式成立
>
> $$R(f)\leq\hat{R}(f)+\epsilon(d,N,\delta)$$
>
> 其中，
>
> $$\epsilon(d,N,\delta)=\sqrt{\frac{1}{2N}(\log d+\log\frac{1}{\delta})}$$

解释：不等式左侧的$R(f)$是泛化误差，右侧即为**泛化误差上界**。在右侧泛化误差上界中，第一项$\hat{R}(f)$指的是训练误差，训练误差越小，则泛化误差也就越小。第二项$\epsilon(d,N,\delta)$是样本容量N的单调递减函数，且是$\sqrt{\log d}$阶的函数，这里d代表假设空间的有限个函数集合的函数数量。

- 由此，我们可以看出，假设空间F包含的函数越多，$\epsilon(d,N,\delta)$值越大。样本容量越大，$\epsilon(d,N,\delta)$值越小。
- 我们可以将该公式理解为：（1）训练误差越小，泛化误差越小（2）样本容量N越大，则训练误差与泛化误差越接近（也就是说，只要样本足够多，我们训练出来的模型在我们已有数据集上跑出来的效果和在未知数据集上跑出来的效果是越接近的）（3）假设空间中函数越多，则泛化误差上界越大

证明：证明过程需要使用**Hoeffding**不等式，如下，

> 设$X_1,X_2,\cdots,X_N$是独立随机变量，且$X_i\in[a_i,b_i]$，$i=1,2,\cdots,N$；$\overline{X}$是$X_1,X_2,\cdots,X_N$的经验均值，即$\overline{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$，则对$\forall t>0$，以下不等式成立：
>
> $$P(\overline{X}-E(\overline{X})\geq t)\leq\exp(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i-a_i)^2})$$
> $$P(E(\overline{X})-\overline{X})\geq t)\leq\exp(-\frac{2N^2t^2}{\sum_{i=1}^{N}(b_i-a_i)^2})$$

对任意存在于假设空间内的函数$f\in F$，$\hat{R}(f)$是N个对立的随机变量$Loss(Y,f(X))$的样本均值，$R(f)$是随机变量$Loss(Y,f(X))$的期望值。如果Loss Function $Loss(Y,f(X))$取值于区间$[0,1]$，则根据**Hoeffding**不等式中的第二个公式，有如下对应关系：

- 随机变量X对应这里的随机变量Loss Function $Loss(Y,f(X))$
- $\overline{X}$是经验均值，对应这里的$\hat{R}(f)$
- $E(\overline{X})$是期望，对应这里的$R(f)$
- $\forall t>0$对应这里的$\forall\epsilon>0$

- $[a_i, b_i]$对应这里的区间$[0, 1]$

代入公式得到关系式：

$$P(R(f) - \hat{R}(f) \geq \epsilon) \leq \exp(-\frac{2N^2\epsilon^2}{\sum(b_i - a_i)^2}) \leq exp(-2N\epsilon^2) \quad (1)$$

其中第一步不等式是使用公式的结果，第二步不等式是因为$[a_i, b_i] == [0, 1]$，所以分母$\sum(b_i - a_i)^2$小于1，故进行放缩。

因为这里$F = \{f_1, f_2, \cdots, f_d\}$是一个有限的集合，故

$$\begin{aligned}
&P(\exists f \in F, R(f) - \hat{R}(f) \geq \epsilon) \\
&= P(R(f_1) \geq \hat{R}(f_1) + \epsilon) \cup P(R(f_2) \geq \hat{R}(f_2) + \epsilon) \cup \cdots \cup P(R(f_d) \geq \hat{R}(f_d) + \epsilon) \\
&= P(\cup_{f \in F} R(f) \geq \hat{R}(f) + \epsilon) \\
&\leq \sum_{f \in F} P(R(f) \geq \hat{R}(f) + \epsilon) \\
&\leq d \exp(-2N\epsilon^2)
\end{aligned}$$

其中第一个不等式使用了基本的概率不等式（多个事件的并集概率不大于各事件概率的和），第二个不等式使用了前文的公式(1)

上述公式可以总结为

$$P(\exists f \in F, R(f) - \hat{R}(f) \geq \epsilon) \leq d \exp(-2N\epsilon^2)$$

或者可以写作其的逆否命题

$$\forall f \in F, P(R(f) - \hat{R}(f) < \epsilon) \geq 1 - d \exp(-2N\epsilon^2)$$

这里令$\delta = d \exp(-2N\epsilon^2)$，则公式化简为

$$\forall f \in F, P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$$

因此我们证明得到，至少以概率$1 - \delta$，有$R(f) < \hat{R}(f) + \epsilon$，其中我们定义了$\delta = d \exp(-2N\epsilon^2)$，则我们可以求解出 ϵ 的表达式

$$\epsilon = \sqrt{\frac{1}{2N}(\log d + \log \frac{1}{\delta})}$$

综上，我们证明了**定理1.1**，值得注意的是，**以上讨论的只是假设空间包含有限个函数情况下的泛化误差上界，对一般的假设空间要找到泛化误差上界则非常复杂。**

> 标注（tagging）问题和分类（classification）问题的区别

- 标注问题也是一个监督学习问题，可以认为标注问题是分类问题的一个推广
- 标注问题的输入是一个观测序列，输出的是一个标记序列。也就是说，分类问题最后的输出是一个值（即输入属于哪一类），而标注问题输出的是一个向量，向量的每个值都属于一种标记类型

三、习题

> **1.1** 说明伯努利模型的极大似然估计以及贝叶斯估计中的统计学习方法三要素。伯努利模型是定义在取值为0与1的随机变量上的概率分布。假设观测到伯努利模型n次独立的数据生成结果，其中k次的结果为1，这时可以用极大似然估计或贝叶斯估计来估计结果为1的概率

- 统计学习方法三要素：模型 + 策略 + 算法

1. 伯努利模型的极大似然估计
    1. 模型：$F = \{f|f_p(x) = p^x(1-p)^{1-x}\}$
    2. 策略：最大似然函数
    3. 算法：$\arg\max_p L(p)$

    Bernoulli的Probability mass function可以写作$f(X) = p^X(1-p)^{1-X}$，这个公式代表着取$X=1$的概率为p，取$X=0$的概率为$1-p$

    其对应的最大似然函数为：

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

    这里的i代表第i次实验，我们期望找到最大化似然函数的p，则将上述公式改写为对数似然函数，其中$Y = \sum_{i=1}^{n} X_i$

$$LL(p) = \sum_{i=1}^{n} \log p^{X_i}(1-p)^{1-X_i}$$
$$= \sum_{i=1}^{n} X_i(\log p) + (1-X_i)\log(1-p)$$
$$= Y\log p + (n-Y)\log(1-p)$$

为了求得最大似然，我们将$LL(p)$进行求导，并令其等于0来求极值点

$$\frac{\partial LL(p)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0$$

解得

$$p = \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1*k + 0*(n-k)}{n} = \frac{k}{n}$$

最后求得结果为1的概率$P(Y=1) = p = \frac{k}{n}$

2. 贝叶斯估计中的极大似然估计
    1. 模型：$\arg\max_\theta P(X_1, X_2, \cdots, X_n|\theta)P(\theta)$
    2. 策略：最大后验概率
    3. 算法：最大后验概率，在预测时计算数据对后验概率分布的期望值

    根据贝叶斯公式，我们可以得到如下推断

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

我们的目标是最大化该贝叶斯公式的后验概率，因为$P(X)$为常量，所以我们只需要最大化分母即可。

$$\arg\max_\theta P(X_1, X_2, \cdots, X_n|\theta)P(\theta)$$
$$= \arg\max_\theta \prod P(X_i|\theta)P(\theta)$$

其中$P(X_i|\theta)$为似然函数，我们可以直接使用**Bernoulli(p)**对应的似然函数公式，$P(\theta)$我们选择Beta Distribution，Beta Distribution的最主要作用就是为**某项实验的成功概率建模**，其概率密度函数为

$$Beta(\theta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

因为分母$B(\alpha,\beta)$在求解$\arg\max$过程中实际上是无关紧要的变量，所以我们只要最大化分子项即可，所以原式变为：

$$\begin{aligned}&\arg\max_{\theta}P(X_1,X_2,\cdots,X_n|\theta)P(\theta)\\&=\arg\max_{\theta}\prod P(X_i|\theta)P(\theta)\\&=\arg\max_{\theta}\theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1}\end{aligned}$$

求解得到

$$\theta=\frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$$

其中，α,β是Beta分布的两个参数，我们可以根据不同情况进行赋值。

因为这里我们估计的参数θ实际上就是对应的概率，所以我们可以得到

$$P(Y=1)=\frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$$

> **1.2** 通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计

- 这里假设模型的条件概率分布是$P_\theta(Y|X)$，损失函数为对数似然损失，如下

$$Loss(Y,P_\theta(Y|X))=-\log P_\theta(Y|X)$$

经验风险最小化即让损失函数越小越好（不考虑正则化项得影响），即

$$\begin{aligned}&\arg\min_\theta \sum(-\log P_\theta(Y|X))\\&=\arg\max_\theta \sum(\log P_\theta(Y|X))\\&=\arg\max_\theta \prod \log P_\theta(Y|X)\\&=\arg\max_\theta \log L(\theta)\end{aligned}$$

倒数第二行中的$\prod \log P_\theta(Y|X)$就是最大似然函数$L(\theta)$，故我们证明了经验风险最小化等价于极大似然估计

- 反过来证明，我们可以知道极大似然估计的似然函数为

$$L(\theta)=\prod P_\theta(Y|X)$$

我们取其对数似然估计

$$LL(\theta)=\sum \log P_\theta(Y|X)$$

则我们极大似然估计的过程如下

$$\begin{aligned}\arg\max_\theta LL(\theta)&=\arg\max_\theta \sum \log P_\theta(Y|X)\\&=\arg\min_\theta \sum(-\log P_\theta(Y|X))\end{aligned}$$

等价于经验风险最小化公式