Author：Liu Jian

Time：2019-04-27

机器学习4-参数估计、贝叶斯决策论

0 概念和符号说明

概率与统计的关系：概率论是在给定条件（已知模型和参数）下，对要发生的事件（新输入数据）的预测。统计推断是在给定数据（训练数据）下，对数据生成方式（模型和参数）的归纳总结。概率论是统计学的数学基础，统计学是对概率论的应用。**简而言之，概率是已知模型和参数，推数据；统计是已知数据，推模型和参数。**

X：输入随机变量，可以是一维或多维 $X=(X^{(1)},X^{(2)},\cdots,X^{(n)})$，这里是 n 维。当不需要区分输入随机变量与输出随机变量时，用 X 表示所有的随机变量，即 $X=(X,Y)$。

Y：输出随机变量。

$\mathbb{X}$：输入空间，X 的可取值空间，比如，离散点 $\{q_1,q_2,\cdots,q_r\}$ 或连续值。

$\mathbb{Y}$：输出空间，Y 的可取值空间，比如，离散点 $\{v_1,v_2,\cdots,v_s\}$ 或连续值。

武大《应用数理统计》中给出了总体和个体的概念：

> 总体就是一个具有确定概率分布的随机变量(一维或多维)，而一个个体则是随机变量的一次观测值。从一个总体中抽取 N 个个体组成一个样本。样本具有二重性，一般在理论推导中总把样本视为随机变量，而在用理论推导所得出的结论进行具体推断时，样本就成了具体数字 (上述 q_i 、 v_i) 了。

也就是说，总体就是这里我们所说的随机变量 X 、 Y ，常用大写字母表示。个体就是一次抽样的就结果，也就是样本点。样本点具有二重性，抽样前是与总体同分布的随机变量，具有随机性，用于理论推导；抽样后就是一个具体的数值，用于具体推断。**一般地，我们为了区分二重性，采用大写字母表示抽样前/具有随机性/用于理论推导的样本点，采用小写字母表示抽样后/是具体数值/用于具体推断的样本点。**

x ：输入实例， $x \in \mathbb{X}$ ，为输入空间中某个具体数值，也就是抽样后没有随机性的样本点。注意符号属于 $\in$ 和等号 $=$ 的区别。属于某个集合表明它就为某个集合的元素，而等于符号表达的只是此刻二者的值相等，类似于编程语言中的常量和变量的区别。常量和变量的值可以相等，但他们代表的含义是不同的。同样地，可以是一维或多维向量 $x = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})$ ，带括号的上标表示第几个分量。

y ：输出的标记， $y \in \mathbb{Y}$ ，为输出空间中某个具体数值，也就是抽样后没有随机性的样本点。

$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ ：训练数据/样本数据，是抽样后没有随机性的样本， N 为样本容量。 (x_i, y_i) 与上面的 (x, y) 的含义相同，多出的下标表示其为第 i 个训练数据点/样本点。若输入随机变量是多维的，则 $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})$ 。也可以将样本数据分开写为：
$\mathcal{X} = (x_1, x_2, \cdots, x_N)$ 和 $\mathcal{Y} = (y_1, y_2, \cdots, y_N)$ 。

$\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ ：抽样前的样本，具有随机性。

样本空间：样本 $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)\}$ 可能取值的全体。注意，不要把抽样后的结果 $\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ 当作样本空间了。

简单随机样本：满足独立性和代表性的样本(独立同分布采样得到的样本)。

需要补充说明的是 x 、 y 也用作函数的自变量和因变量，此时就为变量而不是数值了。

- 在英文中，似然（likelihood）和概率（probability）是同义词，都指事件发生的可能性。但在统计中，似然与概率是不同的东西。**概率是已知参数，对结果可能性的预测。似然是已知结果，对参数是某个值的可能性预测。** X 为随机变量、 θ 为模型参数：

  $P(X|\theta)$ ：

  1. 若 θ 是确定的， X 是变量，这个函数叫做概率函数 (probability function) ，它描述对于不同的随机变量取值其出现的概率是多少；
  2. 若 X 是确定的， θ 是待定的，这个函数叫做似然函数 (likelihood function)，它描述对于不同的模型，出现同一个确定的数据点的概率是多少。

  $P(\theta)$ ：模型参数的先验概率

  $P(\theta|X)$ ：模型参数的后验概率

1 参数估计

概率模型的训练过程就是参数估计过程。对于参数估计，统计学界的两个学派分别提供了不同的解决方案：**频率主义学派认为参数 θ 为一个未知但客观存在的固定值，因而可通过优化似然函数等准则来确定参数值，对应最大似然估计。贝叶斯学派认为参数 θ 是未观察到的随机变量，其本身也可有分布，因此可假定参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布，对应最大后验估计和贝叶斯估计。**

随机变量 X (这里没有输入随机变量和输出随机变量之分，统统打包为 X)，模型参数 θ 。三种参数估计方法都和贝叶斯公式有关，因此首先从分析**贝叶斯公式**入手：

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \text{i.e.} \quad \text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

posterior：通过 X 得到参数的概率

likelihood：通过参数得到 X 的概率

prior：参数的先验概率，一般是根据人的先验知识来得出的。比如人们倾向于认为抛硬币实验会符合先验分布：Beta分布。当我们选择Beta分布的参数 $a=b=0.5$ 时，代表人们认为抛硬币得到正反面的概率都是 0.5 。

evidence： X 发生的概率，是各种条件下发生的概率的积分，计算公式：

$$P(X)=\int P(X|\theta)P(\theta)\mathrm{d}\theta$$

1.1 最大似然估计 (Maximum Likelihood Estimate，MLE)

上面提到的似然函数针对的是**单个随机变量** X ，而最大似然估计最大化的目标函数是**样本的似然函数(是一个联合概率分布，可以看做是由多个随机变量打包而成的一个新的随机变量** $(X_1,X_2,\cdots,X_N)$ **)**：

$$P(X_1,X_2,\cdots,X_N|\theta)=\prod_{i=1}^{N}P(X_i|\theta)$$

上式中，**等号成立是因为独立同分布采样。**最大化上式等价于最大化对数似然函数：

$$\max_{\theta}L(\theta)=\max_{\theta}\log\prod_{i=1}^{N}P(X_i|\theta)=\max_{\theta}\sum_{i=1}^{N}\log P(X_i|\theta)$$

将对数似然函数对 θ 求导并令导数为 0 即可求解。最大似然估计的求解步骤：

- 确定似然函数
- 将似然函数转换为对数似然函数
- 求对数似然函数的最大值 (求导，解似然方程) 。

关于条件似然函数最大化的评述可参见信息熵的笔记中第7章的最后一部分。

1.2 最大后验估计 (Maximum A Posteriori，MAP)

和最大似然估计不同的是，MAP寻求的是能使**样本的后验概率**最大的值 (不是前面公式中**单个随机变量** X **的后验概率**，而是一个联合概率分布，等于各个样本点后验概率分布的乘积) ：

$$\arg\max_{\theta}P(\theta|X_1,X_2,\cdots,X_N)=\arg\max_{\theta}\frac{P(X_1,X_2,\cdots,X_N|\theta)P(\theta)}{P(X_1,X_2,\cdots,X_N)}$$

$$=\arg\max_{\theta}P(X_1,X_2,\cdots,X_N|\theta)P(\theta)=\arg\max_{\theta}\prod_{i=1}^{N}P(X_i|\theta)P(\theta)$$

$$=\arg\max_{\theta}\sum_{i=1}^{N}\log P(X_i|\theta)+\log P(\theta)$$

上式中，第一个等号成立是根据贝叶斯公式，第二个等号成立是因为分母 $P(X_1,X_2,\cdots,X_N)$ 和 θ 无关，第三个等号成立是因为独立同分布采样，第四个等号是取对数后恒等变形的结果。 θ 的先验分布 $P(\theta)$ 可按照实际情况来选择，比如抛硬币实验，我们就可以选择上面说过的Beta分布。

最后后验估计的求解和最大似然估计类似，都是对目标函数对 θ 求导并令导数为 0 。求解步骤如下：

- 确定参数的先验分布以及似然函数
- 确定参数的后验分布函数
- 将后验分布函数转换为对数函数
  求对数函数的最大值 (求导，解方程)

1.3 贝叶斯估计 (Bayesian Estimation, BE)

BE和MAP类似，都是以最大化后验概率为目的。区别在于：MLE和MAP都是只返回 θ 的估计值就完事了，而BE要计算后验概率分布 (当然最终也要返回一个 θ 值)。因此，MAP在计算后验概率时可以忽略分母 $P(X_1, X_2, \cdots, X_N)$ ，而贝叶斯估计则不能忽略。

这里先引入**共轭分布** (conjugate distribution) 的概念：

还是回到贝叶斯公式：

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

$$P(X) = \int P(X|\theta)P(\theta)\mathrm{d}\theta$$

若我们假设参数的先验分布 $P(\theta)$ 和分布/似然函数 $P(X|\theta)$ ，则可以通过上式计算后验分布 $P(X|\theta)$ 。当先验分布和后验分布都是同一类型的分布，则称**先验分布 $P(\theta)$ 为分布/似然函数 $P(X|\theta)$ 的共轭分布**。可以举几个例子：

- likelihood为高斯分布，prior为高斯分布，则posterior也为高斯分布
- likelihood为伯努利分布/二项分布，prior为Beta分布，则posterior也为Beta分布
- likelihood为多项式分布，prior为Dirichlet分布（Beta分布的一个扩展），则posterior也为Dirichlet分布
- likelihood为指数分布/泊松分布，prior为Gamma分布，则posterior也为Gamma分布

在把后验概率推导为和先验概率一样的分布形式时，分母 $P(X)$ 其实可以看做一个常数，起归一化的作用，因此在证明某个先验分布为某个分布的共轭分布时是无需计算 $P(X)$ 的。

通过选择共轭先验，可以大大简化计算。比如，伯努利分布的共轭先验为Beta分布，则Beta分布的参数从 a ， b 变化为 $a + N\overline{x}$ ， $b + N - N\overline{x}$ （$\overline{x}$ 为样本均值)。显然，使用共轭先验之后，只需调整 a ， b 这两个预先给定的值就可以方便地根据采样数据进行模型更新。

在求得后验分布后，返回 θ 的期望值作为参数的最终结果。BE的基本步骤如下：

- 确定参数 θ 的先验分布 $P(\theta)$
- 确定**样本的似然函数**：

$$P(X_1, X_2, \cdots, X_N|\theta) = \prod_{i=1}^{N} P(X_i|\theta)$$

- 根据贝叶斯公式，求 θ 的后验分布：

$$P(\theta|X_1, X_2, \cdots, X_N) = \frac{P(X_1, X_2, \cdots, X_N|\theta)P(\theta)}{P(X_1, X_2, \cdots, X_N)}$$

- 计算贝叶斯估计值：

$$\hat{\theta} = \int \theta P(\theta|X_1, X_2, \cdots, X_N)\mathrm{d}\theta$$

1.4 MLE、MAP、BE的关系

- 最大似然估计等价于最小化训练集上的经验分布相对于建立在模型上的概率分布之间的交叉熵。
- MAP允许我们把先验知识加入到估计模型中，这在样本很少的时候是很有用的。因为样本很少的时候我们的观测结果很可能出现偏差，此时先验知识会把估计的结果“拉”向先验，实际的预估结果将会在先验结果的两侧形成一个顶峰。通过调节先验分布的参数，比如Beta分布的 a 和 b ，我们还可以调节把估计的结果“拉”向先验的幅度， a 和 b 越大，这个顶峰越尖锐。我们称 a 和 b 这样的参

数为预估模型的“**超参数**”，即参数 θ 的参数 (先验分布 $P(\theta)$ 中的参数) 。MLE可以认为是MAP中把先验概率分布 $P(\theta)$ 视为均匀分布的特例。

- BE相对于MAP的好处在于，BE计算了整个后验概率的分布，从而也能求出其他一些比如分布的方差之类的值来供参考，比如计算出来方差太大的，我们可以认为分布不够好。通过选择合适的**超参数**，可以获得方差更小的分布。因此，分布的方差可以作为超参数选择的一个考虑因素。实际上，BE的估计结果会比MAP的估计结果更加接近先验结果。
- 相比于极大似然估计，最大后验估计和贝叶斯估计引入了先验信息，相当于在进行**正则化**。
- 先验分布 $P(\theta)$ 中的超参数对应着正则项的系数。对于MAP而言，当 $P(\theta)$ 是拉普拉斯分布时，相当于加上 L_1 正则项 (LASSO) ，当 $P(\theta)$ 是高斯分布时，相当于加上 L_2 正则项 (岭回归) 。当然，并不是所有的正则项都对应着一个先验分布的对数。

1.5 条件似然最大化

机器学习研究的是存在输入输出 (X, Y) 的问题，前文阐述的参数估计方法应用到机器学习中时，实际上是令 $X = (X, Y)$ ，即构建的是联合概率模型。实际上，参数估计的方法也可以应用到条件概率模型，即条件似然最大化。

条件似然最大化其实就是将似然函数中的联合概率分布 $P(X, Y)$ 替换为条件概率分布 $P(Y|X)$ ：

$$\max_{\theta} L(\theta) = \max_{\theta} \log \prod_{i=1}^{N} P(Y_i|X_i, \theta) = \max_{\theta} \sum_{i=1}^{N} \log P(Y_i|X_i, \theta)$$

事实上，前文似然函数中的概率分布 $P(X|\theta)$ 以及这里条件似然的概率分布 $P(Y|X, \theta)$ 都是指我们所假设的待求的概率模型，即 P_{model} 。在关于信息熵的笔记中 (第七章最后一部分) 我们证明了最大似然估计等价于最小化交叉熵，并指出条件似然最大化等价于最小化 $P_{data}(X, Y|\theta)$ 和 $P_{model}(Y|X, \theta)$ 的交叉熵，即最小化 $P_{data}(X, Y|\theta)$ 和 $P_{model}(Y|X, \theta)$ 的差异。显然，条件似然最大化在信息熵的框架下看是存在问题的。针对这一点，我们提出了如下三种处理方法：

1. 对样本数据按照 X 的取值进行分类，对 X 每一种可能的取值点分别进行最大似然估计得到对应的概率模型，综合所有点的模型得到最终的条件概率模型。
2. 若 $P_{data}(X)/P_{real}(X)$ 为常数值，即 X 的各种可能取值出现的概率相等，表现在采样数据中就是 X 的各种可取值出现的次数都差不多。那么，可以不对样本数据进行分类，直接进行条件似然最大化。此时，条件似然最大化也等价于最小化 $P_{data}(Y|X, \theta)$ 和 $P_{model}(Y|X, \theta)$ 的交叉熵。

上面提到的这两种处理方式在信息熵的框架下看是合理的，第三种处理方式如下：

3. 不管 X 的各种可能取值出现的概率是否相等，不对样本数据进行分类，直接进行条件似然最大化。这种做法**在最小化交叉熵的视角上来看是有问题的，但在风险最小化的角度上来看又是合理的。条件似然最大化实际上就是最小化条件概率模型 $P_{model}(X|Y, \theta)$ 关于训练数据集的平均损失，即经验风险最小化。** 其中，损失函数为对数损失函数 $-\log P(Y|X, \theta)$ 。

我们把上述现象解释为不同体系对同一问题的看法不同，没有谁对谁错之分。

1.6 参考文献

极大似然估计，最大后验概率估计(MAP)，贝叶斯估计

贝叶斯估计、最大似然估计、最大后验概率估计

正则化可见西瓜书11.4节P252

2 风险最小化

2.1 损失函数

(X, Y) 表示我们所要研究的系统给出的输入输出，$f(X)$ 是我们要构建的一个用以模拟我们所要研究系统的决策函数，给定一个 x'，决策函数产生一个输出 $y^* = f(x')$，当然这和真实系统给出的 y' 是有差距的，这个差距用损失函数 $L(Y, f(X))$ 来衡量。常用的损失函数，如下：

1. $0-1$ loss function

$$L(Y, f(X)) = \begin{cases} 1, & y = f(x) \\ 0, & y \neq f(x) \end{cases}$$

2. quadratic loss function

$$L(Y, f(X)) = (y - f(x))^2$$

3. absolute loss function

$$L(Y, f(X)) = |y - f(x)|$$

4. logarithmic/loglikelihood loss function

$$L(Y, f(X)) = -\log P_{model}(y|x)$$

观察上面的损失函数，其中对数损失函数是没有直接给出决策函数 $f(X)$，而是给出了概率模型 $P_{model}(X|Y)$。可以这样理解对数损失函数，$X = x'$ 时，若由真实系统给出的输出是 $Y = y'$，则说明此时真实系统让 Y 取 y' 的概率比较高 (这种想法类似于最大似然估计)。则将数据点 (x', y') 代入我们要构建的概率模型，得到的概率越高损失越小。而如何通过概率模型给出决策函数呢？这其实是贝叶斯决策理论所解决的问题。根据贝叶斯决策理论，若已知概率模型 $P_{model}(X|Y)$，则决策函数 $f(X) = \arg\min_Y P_{model}(Y|X)$。贝叶斯决策理论的原理会在后文介绍。

对数损失函数只给出了条件概率的形式，或许有人会认为也可以用于联合概率。因为和条件概率模型 $P_{model}(X|Y)$ 的表现类似，若联合概率模型 $P_{model}(X, Y)$ 匹配由真实系统生成的数据点 (x', y') 越好，即 $P_{model}(x', y')$ 给出的概率越高，损失 $-\log P_{model}(x', y')$ 越小。但注意到如下关系：

$$-\log P_{model}(x', y') = -\log P_{model}(y'|x')P_{model}(x') = -\log P_{model}(y'|x') - \log P_{model}(x')$$

显而易见的是，给定输入 $X = x'$，预测 Y 的损失应该是与 x' 出现的概率无关的量，因此损失函数应只含有 $-\log P_{model}(Y|X)$ 项而不应含有 $-\log P_{model}(X)$ 项。但可能又有人会说，当给定 $X = x'$ 时，损失项 $-\log P_{model}(x')$ 为常数，就像可以采用联合概率分布 $P(X, Y)$ 计算对于给定的 X 最有可能出现的 Y 值一样 ($P(X, Y) = P(Y|X)P(X)$，给定 X 后，$P(X)$ 为常数值，因此采用 $P(X, Y)$ 和采用 $P(Y|X)$ 预测最有可能出现的 Y 值是等价的)，采用损失 $-\log P_{model}(X, Y)$ 不会影响期望损失最小化的结果。但事实上，虽然给定 X 后，损失项 $-\log P_{model}(X)$ 不变，可视为常量，但对于 X 的不同取值，损失项 $-\log P_{model}(X)$ 是变化的，不是常量，而期望损失要遍历样本空间中所有可能的点 (X, Y)，显然采用 $-\log P_{model}(X, Y)$ 作为损失函数和采用 $-\log P_{model}(Y|X)$ 作为损失函数进行期望损失最小化并不是等价的。因此，$-\log P_{model}(X, Y)$ 不能作为损失函数。事实上，在计算期望风险时会考虑 $P_{real}(x)$，但不是以损失项 $-\log P_{real}(x)$ 的形式出现，而是被包含在 $P_{real}(X, Y)$ 中以权重的形式去乘以损失函数 (因为期望风险是要计算平均意义下损失，而不是单独某一个点 (x, y) 的损失，所以在计算时每个点的损失都要乘以其出现的概率 $P_{real}(x, y)$ 再求和)。**总而言之，给定 X 的取值，可采用联合概率分布 $P(X, Y)$ 或条件概率分布 $P(Y|X)$ 预测最有可能出现的 Y 值，但给定点 (x, y)，损失函数的计算只能采用条件概率分布 $P(Y|X)$ 而不能采用联合概率分布 $P(X, Y)$。**

2.2 期望风险 (Expected Loss)

损失函数只是针对样本空间中的一个点，为了全面评估决策函数的拟合性能，就需要引入期望风险的概念。期望风险就是决策函数 $f(X)$ 的拟合损失在联合概率分布 $P_{real}(X, Y)$ 下的平均：

$$R_{exp}(f) = \int_{\mathbb{X}\times\mathbb{Y}} L(y, f(x)) P_{real}(x, y) \mathrm{d}x\mathrm{d}y$$

注意，积分是在样本空间中进行的，和采样数据无关，而且此时也根本无需采样数据。假设 (x', y') 由模型 $P_{real}(X, Y)$ 生成，采用 $f(X)$ 对其拟合得到数据 (x', y^*) ,则 (x', y') 出现的概率越高，拟合损失的权重越大，这显然是合理的。

机器学习的目的就是求使期望风险最小的模型 $f(X)$ 。显然，这是一个病态的问题。因为对给定模型 $f(X)$ 若能计算期望风险，则就必须知道联合分布 $P_{real}(X, Y)$ ，而联合分布我们是不知道的。毕竟，若知道真实的联合分布，就不需要进行机器学习了。**为了解决这个问题，我们通过采样，用采样数据的概率分布 $P_{data}(X, Y)$ 来代替真实分布 $P_{real}(X, Y)$ ，这就得到了接下来我们要介绍的经验风险的概念。**

2.3 经验风险 (Empirical Loss)

样本数据集： $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

基于样本集 T 的经验分风险：

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

经验风险也就是模型关于训练样本集的平均损失。**根据大数定律，当样本容量 N 趋于无穷大时，经验风险 $R_{emp}(f)$ 趋于期望风险 $R_{exp}(f)$** 。相关证明在信息熵第七章的最后已有涉及，现明确如下。将样本数据 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ 按随机变量 (X, Y) 的各种可能取值 (q_j, v_j) ($j = 1, \cdots, r$)进行分类，其中第 j 类样本的数量记为 N_j ，得经验风险最小化如下：

$$\min R_{emp}(f) = \min \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) =$$

$$\min \sum_{j=1}^{r} \frac{N_j}{N} L(v_j, f(q_j)) = \min \sum_{j=1}^{r} P_{data}(q_j, v_j) L(v_j, f(q_j))$$

根据附录中提到的伯努利大数定律，当样本容量 N 趋于无穷大时，频率收敛于概率，即 $P_{data}(q_j, v_j) = \frac{N_j}{N} \xrightarrow{P} P_{real}(q_j, v_j)$ ，得证。

期望风险与经验风险的关系：期望风险不涉及数据集，而经验风险基于数据集；期望风险最小化是理论推导，经验风险最小化可看做是期望风险最小化计算的落地，这和参数估计有些类似，参数估计在理论推导时数据集是随机变量，而在计算时则是具体数值，是理论推导的落地。

2.4 结构风险 (Structural Risk)

结构风险就是在经验风险最小化的基础上加上了正则项 $J(f)$ ：

$$R_{srm}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

其中， λ 为人为设置的权重。

3 贝叶斯决策论与朴素贝叶斯

3.1 贝叶斯决策论

贝叶斯决策论处理的问题是若已知系统的概率模型 $P_{real}(X, Y)$ 或 $P_{real}(Y|X)$ ，如何确定决策函数 $f(X)$ 。注意，其任务不是构建概率模型，构建概率模型是参数估计和经验风险最小化 (也可以加上最小化交叉熵，貌似不是很常用) 的任务。总的来说，我们先假设一个用来近似真实模型 P_{real} 的含有待定参数的概率模型 P_{model} ，再通过相关策略，如参数估计、经验风险最小化等确定这些待定参数；得到概

率模型后，即可根据贝叶斯决策论确定决策函数用于预测。

贝叶斯决策论用公式表示如下；

$$f(X) = \arg\min_{y} P(y|x)$$

即，已知概率模型 $P(Y|X)$ ，若给定输入 X 的值，则输出 Y 取使概率模型达到最大的值。**贝叶斯决策论关于 Y 值的选择看起来是显而易见的，同时，我们也指出贝叶斯决策论也可由 $0-1$ 损失的期望风险最小化得到，证明如下：**

决策函数 f 应使得期望风险最小：

$$\arg\min_{f} \sum_{i=1,j=1}^{r,s} L(f(q_i), Y=v_j) P_{real}(Y=v_j, X=q_i)$$

$$= \arg\min_{f} \sum_{i=1,j=1}^{r,s} L(f(q_i), Y=v_j) P_{real}(Y=v_j|X=q_i) P_{real}(X=q_i)$$

$$= \arg\min_{f} \sum_{i=1}^{r} \left(P_{real}(X=q_i) \sum_{j=1}^{s} L(f(q_i), Y=v_j) P_{real}(Y=v_j|X=q_i) \right)$$

$$\Leftrightarrow \arg\min_{f(q_i)\in\mathbb{Y}} P_{real}(X=q_i) \sum_{j=1}^{s} L(f(q_i), Y=v_j) P_{real}(Y=v_j|X=q_i)$$

$$= \arg\min_{f(q_i)\in\mathbb{Y}} \sum_{j=1}^{s} L(f(q_i), Y=v_j) P_{real}(Y=v_j|X=q_i) \ \ (\text{for all } i)$$

上式第一行即为期望风险最小化的含义，对样本空间中的每一个点 (q_i, v_j) ，决策函数给出预测结果 $f(q_i) \in \{v_1, v_2, \cdots, v_s\}$ ，期望风险计算样本空间中每个点加权损失之和；由全概率公式可得上式第二行，恒等变形得上式第三行；而求决策函数 f 等价于求 $f(q_1), f(q_2), \cdots, f(q_r)$ 的值，由上式第三行，最外层求和的每个被加项只是 $f(q_i)$ 的函数，彼此之间没有影响，因此我们只需对每个被加项逐个最小化即可，即得上式第四、五行。若损失函数取 $0-1$ 损失，则上式等价于：

$$\text{for all } i, \quad \arg\min_{f(q_i)\in\mathbb{Y}} \sum_{j=1}^{s} I(f(q_i) \neq v_j) P_{real}(Y=v_j|X=q_i)$$

$$= \arg\min_{f(q_i)\in\mathbb{Y}} P_{real}(Y \neq f(q_i)|X=q_i)$$

$$= \arg\min_{f(q_i)\in\mathbb{Y}} 1 - P_{real}(f(q_i)|q_i)$$

$$= \arg\max_{f(q_i)\in\mathbb{Y}} P_{real}(f(q_i)|q_i)$$

上式第一行的记号 $I(f(q_i) \neq v_j)$ 表示，若预测的标记 $f(q_i)$ 与 v_j 不相等，则未加权损失为 1 ，否则为 0 。 $f(q_i)$ 为 Y 样本空间中某个待定的值，即 $f(q_i) \in \{v_1, v_2, \cdots, v_s\}$ 。则对于 v_j 而言，遍历 Y 的样本空间 $j = 1 \sim s$ ，当且仅当其与 $f(q_i)$ 相等时，损失才为 0 ，不相等时，加权损失为 $P_{real}(v_j|q_i)$ 。所有的加权损失之和即为 $P_{real}(Y \neq f(q_i)|X=q_i)$ ，即得上式第二行。恒等变形可得上式最后一行，即 $f(q_i) = \arg\max_{f(q_i)\in\mathbb{Y}} P_{real}(f(q_i)|q_i) = \arg\max_{y\in\mathbb{Y}} P_{real}(y|q_i)$ 。由 q_i $(i = 1, \cdots, r)$ 的任意性，可知决策函数 $f(X)$:

$$f(X) = \arg\max_{Y} P_{real}(Y|X)$$

至此，得证贝叶斯决策论与损失函数为 $0-1$ 损失的期望风险最小化的等价性。

3.2 朴素贝叶斯

上一节贝叶斯决策论解决的问题是若已知概率模型如何构建决策函数，而朴素贝叶斯针对的问题是如何构建概率模型。我们知道随机变量 $X=(X^{(1)},X^{(2)},\cdots,X^{(n)})$ 的概率实际上可看作各分量的联合概率。随着分量的增多，需要的训练数据呈爆发性增长，计算量会非常大。为此，朴素贝叶斯基于一个很强的假设条件，通过牺牲准确性来降低构建概率模型时的计算量。朴素贝叶斯假设在给定 Y 后，随机变量 X 的各分量 $(X^{(1)},X^{(2)},\cdots,X^{(n)})$ 相互独立 (属性条件独立性假设，注意这里讨论的是随机变量的分量，没有涉及到样本数据)，进而采用参数估计或经验风险最小化等策略估计概率模型，最后根据贝叶斯决策论进行决策。

由贝叶斯决策论：

$$f(X)=\arg\max_Y P(Y|X)=\arg\max_Y \frac{P(Y)P(X|Y)}{P(X)}=\arg\max_Y P(Y)P(X|Y)$$

又由 X 分量的条件独立性假设：

$$P(X|Y)=\prod_{k=1}^{n}P(X^{(k)}|Y)$$

我们要估计的概率为 $P(Y)$ 和 $P(X^{(k)}|Y)$ ($k=1,\cdots,n$)。现记分量 $X^{(k)}$ ($k=1,\cdots,n$) 的可取值空间为 $\{\hat{q}_1^k,\hat{q}_2^k,\cdots,\hat{q}_{r_k}^k\}$ (为了和 X 的可取值点保持一致性又能区分开来，字母 q 上加了帽子变为 $\hat{q}$，上标 k 是为了区分各个分量，下标 r_k 中还含有下标 k 是因为各个分量的可取值个数可能不同；X 的可取值空间 $\{q_1,q_2,\cdots,q_r\}$ 由各个分量的可取值空间组合而成)。按分量 $X^{(k)}$ 为离散型或连续型随机变量，分两种情形讨论。

1. 若 $X^{(k)}$ 为离散型随机变量。则：

$$P(Y=v_j)=\frac{C_j}{N}$$

$$P(X^{(k)}=q_i^k|Y=v_j)=\frac{C_{ji}^k}{C_j}$$

上式中 N 如前所述为样本个数，C_j 表示对样本数据中标记 $Y=v_j$ 的样本点进行计数，C_{ji}^k 表示对样本数据中标记 $Y=v_j$ 且第 k 个输入分量 $X^{(k)}=q_i^k$ 的样本点进行计数。书上强调，上述第一个式子由大数定理得到，第二个式子由最大似然估计得到。事实上，大数定理或最大似然估计均可得上述两式，但大数定理除了可以得到上述两式外，还可表明当样本数量趋于无穷大时，上述两式依概率收敛于真实概率，可参见附录中大数定理的内容。

也可采用如下公式计算：

$$P(Y=v_j)=\frac{C_j+\mu}{N+\mu s}$$

$$P(X^{(k)}=q_i^k|Y=v_j)=\frac{C_{ji}+\mu}{C_j+\mu r_k}$$

其中 s 如前所述为 Y 可取值个数，r_k 如前所述为第 k 个输入分量 $X^{(k)}$ 的可取值个数，$\mu\geqslant 0$ 为给定参数。相比前面采用大数定理/极大似然估计的结果，此处的计算公式引入了数据的先验，属于贝叶斯估计的范畴。当 $\mu=0$ 时就为前面极大似然估计的计算公式；当 $\mu=1$ 时，就为所谓的**拉普拉斯平滑/拉普拉斯修正**，此时，实际上是在假设标记和输入分量属性值的先验分布为均匀分布。我们可以注意到，使用前面极大似然估计的公式时，会因采样不够多而导致某些输入分量的属性值没有出现而将其概率判断为 0。而采用**拉普拉斯平滑/拉普拉斯修正**可避免这一情况。此外，当样本数量足够多时，平滑/修正过程引入的先验的影响也会逐渐变得可忽略，从而趋于真实概率。

2. 若 $X^{(k)}$ 为连续型随机变量，为计算 $P(X^{(k)}|Y)$ 需假设连续型的概率模型。参见西瓜书P151页的内容，可先对样本数据按照 Y 的各种可能取值进行分类，对每一类标记构建第 k 个输入分量 $X^{(k)}$ 的正态分布模型 (可采用极大似然估计)，构建完此类标记下输入 X 的所有分量的正态分布模型

后，累乘即得概率模型 $P(X|Y=y)$（y 表示某类标记的值)。综合 s 个这样的概率模型，即得模型 $P(X|Y)$ 。

3.3 半朴素贝叶斯与贝叶斯网

由贝叶斯决策论可知，决策函数 $f(X)=\arg\min_Y P(Y)P(X|Y)$，问题主要是要求解概率 $P(Y)$ 和 $P(X|Y)$ 。概率 $P(Y)$ 好求，但概率 $P(X|Y)$ 会因为输入 X 的分量过多使得这样一个联合概率的求解面临组合爆炸的问题而不可行。为此，朴素贝叶斯假设各个输入分量 (或称之为属性) 在给定标记 Y 后条件独立，而这里将要介绍的半朴素贝叶斯的基本想法是在朴素贝叶斯的基础上适当考虑一部分输入分量间的相互依赖信息，从而既避免了完全联合概率巨大的计算量，又不至于彻底忽略了比较强的输入分量间的依赖关系。进一步，可采用贝叶斯网/信念网 (一种有向无环图) 来刻画输入分量间较复杂的依赖关系。

半朴素贝叶斯的常用的模型有：

1. 独立依赖分类器 (One-Dependent Estimator, ODE)，核心假设是：

$$P(X|Y)=\prod_i^n P(X^{(i)}|Y,pa_i)$$

ODE假设每个输入分量 $X^{(i)}$ 除了依赖于标记 Y 以外，最多仅依赖于一个其他的输入分量 pa_i 。输入分量 pa_i 称为输入分量 $X^{(i)}$ 的父属性，记号中的 pa 就来自于英文单词"parent"。不同的父属性选择策略可得不同的独立依赖分类器:

- 通过将所有分量的父属性 (pa_i for all i) 设定为某一个属性 (这一属性称为超父属性，可采用交叉验证选择超父属性为最优的模型。**从这里我们可以看到，交叉验证的目的是模型选择，而参数估计的目的是在模型形式给定的情况下，确定模型中的待定参数，要注意二者之间的区别**)，可得 SPODE (Super-Parent ODE)；进一步，基于SPODE进行集成学习，可得 AODE (Averaged One-Dependent Estimator) (AODE对所有超父属性不同的SPODE进行集成，因此没有模型选择的过程);
- TAN (Tree Augmented naive Bayes) 通过如下的策略确定每个输入分量的父属性：TAN 首先计算每个属性间的条件互信息，用以衡量属性间的依赖关系；再以属性为结点构建完全图，两个结点间边的权重为对应的条件互信息；基于完全图构建最大带权生成树，并挑选根变量，将边置为有向，由此可得各个输入分量/属性结点的父属性。

2. 考虑属性间的高阶依赖，即每个输入分量除了依赖于标记以外，还依赖于多个其他的输入分量。比如，可将 ODE 中一个属性 pa_i 改为包含有 k 个属性的集合，从而将 ODE 扩展为 kDE。

基于上述半朴素贝叶斯对概率模型的假设，接下来就可采用参数估计等方法求解概率。

贝叶斯网是一种概率图模型： $B=\langle G,\theta\rangle$ ，其中 G 为贝叶斯网的网络结构，是一个有向无环图，表示随机变量各分量间的依赖关系， θ 为贝叶斯网的参数，用于定量描述各种依赖关系，也就是各种条件概率。显然，参数 θ 是依赖于网络结构 G 的。**那么对贝叶斯网的学习就涉及到两方面：1) 对贝叶斯网结构 G 的学习；2) 在已知结构 G 的条件下对参数 θ 进行估计。这和贝叶斯模型选择类似，首先要基于 BME 选择模型，再基于选择的模型对待定参数进行估计。**

西瓜书 7.5 节从信息论的角度出发，基于“最小描述长度” (Minimal Description Length, MDL) 准则提出根据评分函数进行贝叶斯网的学习：

$$s(B|D)=mg(\theta)-\log p(D|B)$$

其中， D 为样本容量为 N 的样本数据； m 表示参数 θ 的个数， $g(\theta)$ 表示描述每个参数所需要的编码位数，则 $mg(\theta)$ 表示编码贝叶斯网 B 所需要的编码位数； $-\log p(D|B)$ 为样本数据的负对数似然，按照信息论的观点，表达了基于概率模型 B 对样本数据 D 进行编码所需要的位数。**评分函数越小的贝叶斯网络越好。可以看到，最小描述长度准则就是寻找一个自身编码长度小，对样本数据进行描述时编码长度也小的模型。**

注意评分函数和 KL 散度 (相对熵) 的区别。 $KL(P_{data}||P_{model}) = H(P_{data}, P_{model}) - H(P_{data})$，其中 $-\log p(D|B)$ 对应交叉熵 $H(P_{data}, P_{model})$，但 $mg(\theta)$ 却不对应信息熵 $H(P_{data})$ 而是信息熵 $H(P_{model})$。

从风险最小化的角度上来看，上式就等价于结构风险最小化，$-\log p(D|B)$ 为对数损失函数，$mg(\theta)$ 为正则化项。

从贝叶斯模型选择的角度上来看：

- 若 $g(\theta) = 1$，可得 AIC：$\frac{1}{2}AIC = m - \log p(D|B) = m - \log p(D|G, \theta)$
- 若 $g(\theta) = \frac{\log N}{2}$，可得 BIC：
  $\frac{1}{2}BIC = m\frac{\log N}{2} - \log p(D|B) = m\frac{\log N}{2} - \log p(D|G, \theta)$

但是，贝叶斯模型选择中的参数 θ 是为给定结构 G 下的极大似然估计值 $\tilde{\theta}$ 或最大后验估计值 $\hat{\theta}$，而这里的 θ 为变量。**注意到，当网络结构 G 给定时，评分函数的第一项为常量，最小化评分函数就相当于在进行极大似然估计。可见，最小化评分函数相当于在同时进行模型选择和参数估计。而贝叶斯模型选择是先假设模型结构并进行参数估计，再回过头来基于参数估计值计算模型选择指标 BIC、KIC 等以评估模型结构的好坏。**

贝叶斯网的推断：最理想的是根据贝叶斯网定义的联合概率分布精确计算后验概率，但这是“NP”难的，**一般采用近似推断 (如吉布斯采样) 或变分推断的方法。**

朴素贝叶斯、半朴素贝叶斯、贝叶斯网都是在对随机变量分量间的依赖关系进行假设，从而简化联合概率模型便于对模型进行参数估计。

4 最小化交叉熵、参数估计、风险最小化、贝叶斯决策论、朴素贝叶斯、贝叶斯网之间的关系

上述几个概念所属层面如下：

- **决策函数层面**：贝叶斯决策论是根据已知的概率模型确定决策函数。
- **模型选择层面**：朴素贝叶斯、贝叶斯网是对概率模型中输入分量间依赖关系的假设。
- **参数估计层面**：最小化交叉熵、参数估计、风险最小化是构建模型的三种策略。

模型选择可看做上升一个尺度的参数估计，但参数估计和模型选择并没有明显的界限。个人认为，参数估计偏向于更好地拟合样本数据，而模型选择还考虑了模型复杂度、过拟合、和泛化能力等方面。基于目前的学习，可从概率模型、统计学习和信息论这三个体系来进行机器学习：

视角	拟合	正则化	模型选择 (局限于训练数据)	模型选择 (存在预测数据)
概率模型	极大似然估计	最大后验估计	BME (BIC、KIC)	交叉验证
统计学习	经验风险最小化	结构风险最小化	--	交叉验证
信息论	最小化相对熵/交叉熵	最小描述长度、 AIC	最小描述长度、AIC	交叉验证

相关结论如下：

1. 参数估计与风险最小化：
    - 当损失函数是对数损失函数 (前文已经说明，对数损失函数对应的模型只能是条件概率分布而不能是联合概率分布，下同)时，经验风险最小化 (也就是无正则化) 等价于极大似然估计；(证明见下文)
    - 当损失函数是对数损失函数，模型复杂度由模型的先验概率表示时 (引入先验概率进行正则化)，结构风险最小化等价于最大后验概率估计；(证明和极大似然估计的情形类似，略)
    - 最小化均方误差等价于对高斯模型进行极大似然估计；(证明见 BSPCE 笔记)
    - 最小化绝对误差等价于对拉普拉斯模型进行极大似然估计 。(证明见 BSPCE 笔记)
2. 极大似然估计与最小化交叉熵等价 (证明可参见信息熵的笔记)。

再次强调，各个体系的观念相互间并不是完全等价的。比如，从最小化交叉熵的角度来看极大条件似然是不合理的，极大联合似然才是合理的；但从经验风险最小化的角度上来看，极大条件似然是合理的，极大联合似然是不合理的 (因为损失函数中的概率模型使用联合概率分布是不合理的)。

基于对数损失函数的经验风险最小化和极大条件似然估计的证明：

概率模型： $P_{model}(Y|X)$ ，损失函数： $-\log P_{model}(Y|X)$ ，样本数据集： $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ ，经验风险最小化：

$$\min R_{emp}=\min \frac{1}{N}\sum_{i=1}^{N}-\log P_{model}(y_i|x_i)$$

$$\Longleftrightarrow \max \sum_{i=1}^{N}\log P_{model}(y_i|x_i)$$

等价于极大化条件似然。

5 附录

切比雪夫不等式、大数定理与中心极限定理。

5.1 切比雪夫不等式

$$P(|X-\mathbb{E}(X)|\geqslant \varepsilon)\leqslant \frac{\mathbb{D}(X)}{\varepsilon^2}$$

用切比雪夫不等式估计概率大小，实际上就是计算 $\mathbb{E}(X)$ 和 $\mathbb{D}(X)$ 。

5.2 大数定理

大数定理讲的是随机变量序列的算数平均依概率收敛于它们期望的算术平均。切比雪夫大数定理、新钦大数定理都可推得伯努利大数定理。考研时，看到依概率收敛的题，只能用大数定理。

切比雪夫大数定理：随机变量 X_i ($i=1,2,\cdots,N$) **两两不相关**，且**各方差有界**，则:

$$\frac{1}{N}\sum_{i=1}^{N}X_i\xrightarrow{P}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i)$$

$$\text{i. e. }\lim_{N\to\infty}P\left(\left|\frac{1}{N}\sum_{i=1}^{N}X_i\xrightarrow{P}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(X_i)\right|<\varepsilon\right)=1$$

辛钦大数定理：随机变量 X_i ($i=1,2,\cdots,N$)**独立同分布**，且**期望 μ 存在**，则：

$$\frac{1}{N}\sum_{i=1}^{N}X_i \xrightarrow{P} \mu$$

$$\text{i.e.}\lim_{N\to\infty} P\left(\left|\frac{1}{N}\sum_{i=1}^{N}X_i \xrightarrow{P} \mu\right| < \varepsilon\right) = 1$$

切比雪夫大数定理与辛钦大数定理相比，随机变量两两不相关比独立同分布条件要弱，但方差有界 (意味着期望也存在) 比期望存在要强。

特别地，可将上述随机变量序列 X_i（$i = 1, 2, \cdots, N$）视为进行了 N 次独立重复试验， N_A 表示事件A发生的频率， P_A 表示事件A发生的概率，则：

$$\frac{N_A}{N} \xrightarrow{P} P_A$$

$$\text{i.e.}\lim_{N\to\infty} P\left(\left|\frac{N_A}{N} - P_A\right| < \varepsilon\right) = 1$$

即伯努利大数定律。**伯努利大数定理表明可用某个事件的频率近似代替其出现的概率，这一结论与采用最大似然估计计算的结果相同。**记 $P_A = q$，则 $1 - q$ 为其他事件发生的概率。似然函数：

$$L(\theta) = \log q^{N_A}(1-q)^{N-N_A} = N_A \log q + (N - N_A)\log(1-q)$$

对 q 求导并令之为 0，解得 $q = \dfrac{N_A}{N}$ 。**但采用最大似然估计无法得到的结论是，当 N 趋于无穷时，频率收敛于概率。因此，需注意两个定理的内核是不同的。**

5.3 中心极限定理

中心极限定理讲的是随机变量序列的极限分布是正态分布。棣莫佛-拉普拉斯中心极限定理是列维-林德伯格中心极限定理 (也叫独立同分布的中心极限定理) 的特殊情况，只需记后者。

独立同分布的中心极限定理(列维-林德伯格中心极限定理)：随机变量序列 X_i（$i = 1, 2, \cdots, N$）独立同分布，方差存在，则：

$$\lim_{N\to\infty}\frac{\sum_{i=1}^{N}X_i - N\mu}{\sqrt{N}\sigma} \sim \mathcal{N}(0,1)$$

即 $\sum_{i=1}^{N}X_i$ 近似服从正态分布 $\mathcal{N}(N\mu, N\sigma^2)$ 。棣莫佛-拉普拉斯中心极限定理就是上述随机变量分布为二项分布的特殊情况，即二项分布的极限分布为正态分布。

独立不同分布的中心极限定理：林德伯格中心极限定理、李雅普诺夫中心极限定理。这些定理表明，大量“微小”的相互独立的随机变量 X_i（$i = 1, 2, \cdots, N$）叠加组成的随机变量 $\sum_{i=1}^{N}X_i$ 的极限分布为正态分布。