课本的数学符号约定：

向量用小写加粗的罗马字母表示，所有向量都认为是列向量

大写字母T表示向量或矩阵的转置

矩阵用大写加粗的罗马字母表示

# 第一章

线性模型，关于参数是线性的（而不是关于自变量），因此多项式回归模型是一个线性模型，尽管里面的自变量x有高次方，但参数w都是一次的。

## 多项式拟合：

### 过拟合:

- 随着多项式的次数M的增大， 模型的灵活度越来越高，多项式的各阶系数中甚至会出现比较大的正数或负数，来保证拟合曲线能够经过每一个样本点，但由于灵活度较大，拟合曲线也越来越迎合那些噪声点，从而越来越偏离原始的曲线
- 注意到，对于次数高的多项式来说，那些次数较低的多项式是它们【次数高的多项式】的特殊情况，则高次的多项式相对于低次来说应该能产生更好的结果
- 随着样本大小的增大， 过拟合的现象会减弱。意味着：数据集越大，用于拟合数据的模型越复杂（灵活）。从中可以得出一个启发式的方法：样本的大小应不小于模型参数【adaptive parameter】的n倍(如n=5或10)

**解决过拟合的方法**

正则化用于控制过拟合的现象。通过在误差函数中添加一个惩罚项，来防止多项式的系数w取较大的值。通过调整惩罚项的权重lambda来调整惩罚项的重要性。二次的正则化称为岭回归。

# 数据集的划分方法:

训练集：训练模型

测试集：调整模型参数

验证集：对模型复杂度进行优化

# 概率论：

**1. 基本概念**

先验概率，在不知道观察值A的情况下，某件事B发生的概率 P(B)

后验概率，在已知观察值A的条件下，某件事B发生的概率 P(B|A)

离散型随机变量，概率质量函数，P(X=x)

连续型随机变量，概率密度函数，f(x)delta = P(x<X<x+delta)

**2. 贝叶斯的观点 VS 频率学派:**

不同于经典概率论，用贝叶斯的观点来理解概率。有的事件是无法重复多次的，是不确定的，只能通过过去的知识，以及现有的新证据，来预测未来可能发生的概率

- 频率估计 frequentist estimator：极大似然估计 ML
- 贝叶斯估计 bayesian estimator: 极大似然估计ML 或 最大后验概率MAP

  贝叶斯的写法 $p(w|D) = \frac{p(D|w)p(w)}{p(D)}$，这些量都是关于参数w的函数

**3.似然函数likelihood VS 概率probability：**

*概率*，基于已知的参数估计未知的输出 p(w|D)，对于不同的D，得到w的可能性是多少(w通过估计的 方法计算出来，估计的误差则是根据D的分布来计算，比如对于sin(2pix)，可以从中任意取几组样本点，已知这些样本点服从sin(x)的分布，由此来计算w并估计误差。)

*似然*，基于已知的输出估计未知的参数 p(D|w)，对于不同的w，得到D的可能性是多少(只有一个数据集D，而参数w服从某项分布)。而不是关于w的概率分布(not a probability distribution over w)

*后验* 正比于 (似然 * 先验 ) => p(w|D) 正比于 p(D|w)*p(w)

**4. 高斯分布为例**

D维变量的高斯分布 基本性质

- 均值：D维变量的均值
- 协方差矩阵：大小为D*D，是D维变量的协方差矩阵

高斯分布的似然函数：

- sigma为标准差，var为方差 var=sigma^2
- 从同一个高斯分布中随机选取n个数据点，构成数据集x=(x1,x2,...xn)T，这些样本点都是独立同分布的，设高斯分布的参数为(mu, sigma)，即N(mu, sigma)
- $p(\mathbf{X}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}(x_n|\mu,\sigma^2)$

极大似然估计：

- 求解未知的mu和sigma，使得似然函数最大化
- 通常将似然函数求对数，将累乘变为累加，防止计算机计算时数字下溢【若累乘，由于概率都是小数，可能会导致向下溢出】。
- 求出的$\mu_{ML},\sigma^2_{ML}$，E(mu(ML)) = mu, E(var(ML)) = var*(N-1)/N，即ML低估了方差，这是由于它采用的均值是样本的均值而不是该分布的均值

**5.从贝叶斯的角度看待多项式拟合的问题：**

设样本点中自变量为x, 因变量为 t, 估计到的参数为 w，估计到的多项式为y(x,w)

假设噪声符合高斯分布，均值为$y(x,w)-t$，方差为$\beta^{-1}$

- 通过极大似然估计可以求解出$w_{ML}$ 和 $\beta_{ML}$

- 求解 $w_{ML}$ 相当于最小化误差平方和(误差指的是 $y(x,w)-t$)

接下来，估计w的先验概率，假设$w=\mathcal{N}(0,\alpha^{-1})$

则 $p(w|x,t,\alpha,\beta) \propto p(t|x,w,\beta)*p(w|\alpha)$

由此，可以计算出对于给定数据$(x,t)$，最有可能的w，即w的后验概率

**通过最大化后验概率的方法(MAP)来求解**，代入高斯分布式子，变换得，最终的目标函数：

$$\min(\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})$$

此即相当于之前提到的，**最小化带惩罚项的误差函数**，其中惩罚项的权重$\lambda=\alpha/\beta$

**多项式拟合的总结**，

- 多项式的次数相当于参数的自由度，控制着模型的复杂度；
- 在带正则项的误差函数中，正则项的权重$\lambda$ 也控制着模型的复杂度；
- 因此，越复杂的模型，掌管模型复杂度的参数就越多。

模型选择：对于一系列模型（不同参数对应着不同模型），选择在预测新数据时表现最好的。

**6.数据集的划分：**

前述表明，仅使用训练集的准确率作为评估依据是不全面的，由于过拟合的存在。故一般从数据集中独立选取一部分作为验证集

对于一系列模型，（可能是一个方法用不同部分的训练集，训练出不同的模型；也可以是一个模型的参数可以取一定范围内的值，不同的取值对应不同的模型），使用*验证集* 来评估，选择表现最好的一个模型。最后再用*测试集* 来对这个选出来的模型进行评估

**S-折交叉验证**，将数据分成S等份，用S-1份作为训练集，剩下的1份作为验证集，该验证进行S次（即每一份数据都要被拿来作为一次验证集），

# 决策论：

### 1.推理过程：inference step

求解$p(x,t)$ 或 $p(t|x)$ 的过程为推理过程， x为一系列输入变量，t 为输入变量对应标签或值。利用训练数据训练模型来计算p(t|x)。

### 2.决策过程：decision step

对于给定的x，选择最优的t。即根据推理过程计算得到的关于t的后验概率p(t|x)，来做最优分类。

## 为什么要分离为推理和决策两个过程，而不是直接用一个判别函数做决策?

若直接用判别函数做决策，即输入样本x，直接输出其对应的类别，这样我们不会计算到后验概率$p(C_k|x)$

然而计算后验概率是必要的：

- 最小化风险。问题中损失函数的定义可能经常会修改，如果只有判别函数，那么每当修改损失函数就总要重新训练数据，而如果知道后验概率，只需修改与损失函数对应的最小风险决策准则即可。
- 拒绝选项 rejection option。如果给定被拒绝的数据点所占比例，能用后验概率得出最小化(误分类率的拒绝标准)
- 补偿类先验概率。假设对于医疗X光问题，我们开始时收集到的训练数据中，癌症的出现次数很少，而一个平衡的数据集，要求在每个类别中选择数量相等的样本，要想让模型有更好的泛化能力，我们需要补充训练数据，根据贝叶斯定理，原数据的后验=添加新数据后的先验，因此如果知道后验概率，对训练数据做出修改就比较方便。如果直接学习一个判别函数，则无法直接在原来知识的基础上添加新的训练数据，而需要重新训练。
- 组合模型。对于复杂的应用来说，我们可能希望把问题分解成若干个小的子问题，每个子问题可以通过一个独立的模型来解决。

### 3.分类问题讨论决策论

*例子：假设一位病人拍了X光，记X光的信息为输入变量 x，从中判断其有癌症的对应类别标签为C1，否则为C2。为了尽可能保证没有误判，会选择后验概率较大的类别，即选择 p(C1|x)、p(C2|x) 中的较大者。*

- 最小化误分类概率：
- 将输入空间划分为多个决策空间，每个决策空间的边界称为决策边界或决策面
- 决策空间不一定必须是连续的，也可以是由多个独立的子空间组成
- 目标是最小化错误的概率。

  对于例子，一共有两个决策空间：$R_1$ ，$R_2$。当输入变量落在$R_1$时，对应着有癌症；否则，落在$R_2$对应着没有癌症。$R_k$表示决策类别，$C_k$表示实际类别。则错误的概率为

$$p(mistake) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) = \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1$$

*推理过程：即计算所有的p(Ck|x)*

*决策过程：根据各个Ck的后验概率选择使评判标准最优的Ck*

事实上，在用决策论解决分类问题的时候，有3种完全不同的思路：（复杂度由高到低）

- **生成模型（generative models）**：对输入数据和输出数据进行建模，因此，我们可以根据模型生成一些新的输入数据点；

  a、首先对每一类都要计算一个p(x|Ck) 和 p(Ck)；

  b、使用贝叶斯计算后验概率：p(Ck|x)；

  特点：比较费劲，涉及到x和Ck的联合概率，但是，我们可以从中获取一些额外的信息；比如可以通过归一化得到p(x)，从而了解一个待测样本点是噪声点的可能性有多大（噪声检测）。
- **判别模型（discriminative models）**：

  a、对p(Ck|x)建模；

  b、直接指定输入x的类别（由于x已经作为条件概率中的条件）；
- **判别函数（discriminant function）**：

就是一个映射函数，输入一个x，输出一个label。

- *生成VS判别，例子*：假设x是特征，y是标记，x取1或2，y取0或1。则生成模型学习p(x,y)，判别模型学习p(y|x)。假设样本为(1,0)、(1,0)、(1,1)、(2,1)。

  生成模型：

| **x\y** | **0** | **1** |
|---|---|---|
| **1** | 1/2 | 1/4 |
| **2** | 0 | 1/4 |

  判别模型：

| **x\y** | **0** | **1** |
|---|---|---|
| **1** | 2/3 | 1/3 |
| **2** | 0 | 1 |

  *在实际分类问题中，判别模型可直接用来判断分类情况；而生成模型需要基于贝叶斯法则，再应用到分类中（即需要通过联合分布计算条件概率?）然而，生成模型可以还有其他应用，即生成模型更一般更普适；而判别模型更简单直接*

**4.回归问题讨论决策论**

- 假设对于每个输入 x， 其目标为 t，对 t 的估计为y(x)。
- 损失记为$L[t,y(x)]$， 平均损失$E[L]=\int\int L[t,y(x)]p(x,t)dxdt$
- 一般定义损失函数为平方损失，即$L[t,y(x)]=\{y(x)-t\}^2$
- 代入求得平均损失$E[L]=\int\int\{y(x)-t\}^2p(x,t)dxdt$
  - 最小化$E[L]$，对$y(x)$求导，得出 $\frac{\delta E[L]}{\delta y(x)}=2\int\{y(x)-t\}p(x,t)dt$
  - $y(x)=\frac{\int tp(x,t)dt}{\int p(x,t)dt}=\frac{\int tp(x,t)dt}{p(x)}=\int tp(t|x)dt=E_t[t|x]$
  - 即 在x的条件下，t 的条件均值， y(x)称为回归函数
  - **最小化了 期望平方损失的回归函数 y(x) 由条件概率分布 p[t|x] 的均值 给出**

同样，在解决回归问题时，也有三种思路：

- 首先确定联合概率密度 p(t , x)的推断问题，之后计算条件概率密度p(t|x)，最后求条件均值 E[t|x]，结果即为回归函数
- 首先解决条件概率密度 p(t|x)的推断，之后求条件均值，结果即为回归函数
- 直接从训练数据中寻找一个回归函数 y(x)

**5.停止决策的阈值 rejection region**

当某个分类的概率达到或大于$\theta$ 时，直接用规则判断，而小于$\theta$ 时(不确定性比较高，即无法明确地进行分类)，需要人为即专家判断

# 信息论：

**1.定义**

一个随机变量x的取值 的信息量为h(x)，取值的概率分布为p(x)

若p(x)=1,则意味着事件一定会发生，它蕴含的信息量为h(x)=0，因为毫无悬念

- 假设 x, y 两个变量独立
- 则根据$h()$的概念，应有 $h(x,y)=h(x)+h(y)$，而$p(x,y)=p(x)*p(y)$
- 因此$h(x)$应该是 $p(x)$的对数
- $h(x)=-\log_2 p(x)$

随机变量x的 熵 $H(x)=E[h(x)]=-\sum_x p(x)\log_2 p(x)$

**2. 熵**

在编码论、统计物理、机器学习中的重要概念

*例子1*

假设离散随机变量x 有 8 个可能的取值，需要多少位(bit)来传送x的状态?

answer：假设每个取值都是等概率的，则随机变量的熵
$H(x) = -8 * \frac{1}{8}\log_2 \frac{1}{8} = 3$

*例子2*

| x | a | b | c | d | e | f | g | h |
|---|---|---|---|---|---|---|---|---|
| p(x) | 1/2 | 1/4 | 1/8 | 1/16 | 1/64 | 1/64 | 1/64 | 1/64 |
| code | 0 | 10 | 110 | 1110 | 111100 | 111101 | 111110 | 111111 |

$H(x) = 2$ bits

平均编码长度 = 1/2 * 1 + 1/4 * 2 + 1/8 * 3 + 1/16 * 4 + 4 * 1/64 * 6 = 2 bits

*例子3*

[对离散分布而言]分布越宽广(broader distribution)，熵的值越大。当分布为均匀分布时，熵最大。

**3.熵的类型**

在离散分布的情况下，最大熵对应于变量的所有可能状态的均匀分布

- **微分熵** differential entropy，$-\int p(x)\ln p(x)dx$ 对熵左右乘小区间并取极限。离散变成连续。

  连续变量下，最大微分熵的分布是高斯分布

- **条件熵** conditional entropy $H[y|x] = -\int\int p(y,x)\ln p(y|x)dydx$

  $H[x,y] = H[y|x] + H[x]$ 若x和y互相独立，则$H[x,y] = H[x] + H[y]$

  *理解：假设有一个联合概率分布 p(x,y)，从中抽取了一对(x,y)，如果x的值已知，那么需要确定对应的y值所需的附加信息就是 -ln p(y|x)*

- **相对熵** KL散度
  $D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i)\cdot(\log p(x_i) - \log q(x_i)) = E_p[\log p(x) - \log q(x)]$

  *理解：有一个未知分布p(x)，假设用一个近似的分布q(x)来表示这个分布，KL散度度量的是信息损失，即真实和近似的差异。*

  **注意**：散度并非距离，因为其**不满足对称性** 即 $D_{KL}(p||q) \neq D_{KL}(q||p)$

- **互信息** mutual information
  $I[x,y] = KL(p(x,y)||p(x)p(y)) = -\int\int p(x,y)\ln(\frac{p(x)p(y)}{p(x,y)})dxdy$