概率论与数理统计笔记

第一章 导论

统计学习（statistical learning）是一套以理解数据为目的庞大工具集，可分为**监督式（supervised）**学习和**非监督式（unsupervised）**学习。

第二章 统计学习

2.1 相关概念

1. 统计学习是关于估计$f(\cdot)$的一系列方法，其中$f(\cdot)$为一个定量的响应变量Y和p个不同的预测变量$X=(X_1,X_2,\ldots,X_p)$之间的关系，一般形式如下：

   $$Y=f(X)+\epsilon$$

   其中，ϵ是随机误差项（error term），与X独立，且均值为0.

   误差项包含了一下因素：

   - 真实的关系可能不是$f(\cdot)$，例如在简单线性回归估计中，实际关系可能并不是线性的；
   - 可能是其他变量导致了Y的变化；
   - 可能存在测量误差。
2. 估计$f(\cdot)$的主要原因可分为**预测（prediction）**和**推断（inference）**，其中：
   - **预测**

     关注预测的结果，不关注模型的可解释性和变量之间的关系，可表示为：

     $$\hat{Y}=\hat{f}(X)$$

     其中$\hat{Y}$表示Y的预测值，依赖于两个量，**可约误差（reducible error）**和**不可约误差（irreducible error）**，可约误差可通过改进统计学习方法降低，而不可约误差ϵ是无法降低的，所以即使得到一个f的精确估计，预测仍然存在误差，预测的均方误差可表示为

     $$\begin{aligned}E(Y-\hat{Y})^2 =& E[f(X)+\epsilon-\hat{f}(X)]^2\\ =& \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{可约误差}}+\underbrace{Var(\epsilon)}_{\text{不可约误差}}\end{aligned}$$

- **推断**

  目标不是为了预测Y，而是想明白X和Y之间的关系，可以描述为以下问题：

  - 哪些预测变量与响应变量相关？
    - 响应变量与每个预测因子之间的关系是什么？
    - Y与每个预测变量的关系是否能用一个线性方程概括，还是需要更复杂的形式？

3. 估计$f(\cdot)$的方法可分为**参数方法**和**非参数方法**：

   - **参数方法**

     参数方法指有一定的形式或形状的模型，如假设$f(\cdot)$是线性的，则具有如下形式：

     $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

     在模型选完后则需要使用训练数据去**拟合（fit）**或**训练（train）**模型，即估计参数$\beta_0, \beta_1, \ldots, \beta_p$。参数方法**最大的优势**就是可以将$f(\cdot)$假设为具体的参数形式可简化估计。然而**缺陷**则是选定的模型并非与真正 的$f(\cdot)$在形式上是一致的。**非参数方法适合推断的问题。**

   - **非参数方法**

     非参数方法不需要对函数f的形式事先做明确的假设。**优势**是不限定函数$f(\cdot)$的具体形式，可能在更大的范围选择更适宜$f(\cdot)$，然而有**最致命的缺陷**即无法将估计$f(\cdot)$的问题简化成对参数的估计，需要大量的数据（远远超出参数方法所需要的）。

4. **监督学习**和**非监督学习**的区别在于**前者有响应变量（标签）**，形如$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$而**后者无响应变量（标签）**，形如$\{x_1, x_2, \ldots, x_n\}$。
5. 根据变量的**定量（连续）**和**定性（离散）**类型，可将任务分为**回归**和**分类**问题，前者如对GDP、PM2.5的预测，后者如对动物、生病与否的识别。

2.2题目答案

1. (a) 当**样本量n非常大**，**预测变量数p很小**时，这样容易欠拟合，所以一个**光滑度更高的学习模型更好**。

   (b) 当**样本量n非常小**，**预测变量数p很大**时，这样容易过拟合，所以一个**光滑度更小的学习模型更好**。

   (c) 当预测变量与响应变量之间的关系是**非线性**时，说明光滑度小的模型会容易欠拟合，所以**光滑度高的模型更适合**。

   (d) 当**误差项的方差**$\sigma^2 = Var(\epsilon)$**极大**时，因为方差是指用一个不同的训练数据集估计f时，估计函数的改变量。一般来说，**光滑度越高的统计模型有更高的方差**，所以这里适合**光滑度小的模型**。

2. (a) 收集了美国500强公司的数据。每个公司都记录了利润、员工人数、产业类型和CEO的工资。（回归，推断）

   (b) 考虑研发一个新产品，希望知道它会成功还是失败，收集了先前研发的20个相近产品的数据，并记录它们成功或失败的状态，以及其他若干变量。（分类，预测）

   (c) 兴趣在于预测美元的百分比变化率随全球股市周变动的变化规律，为此收集了2012年所有的周数据。（回归，预测）

3. **偏差**：度量了学习算法的期望预测与真实结果偏离程度，即刻画了学习算法本身的拟合能力。

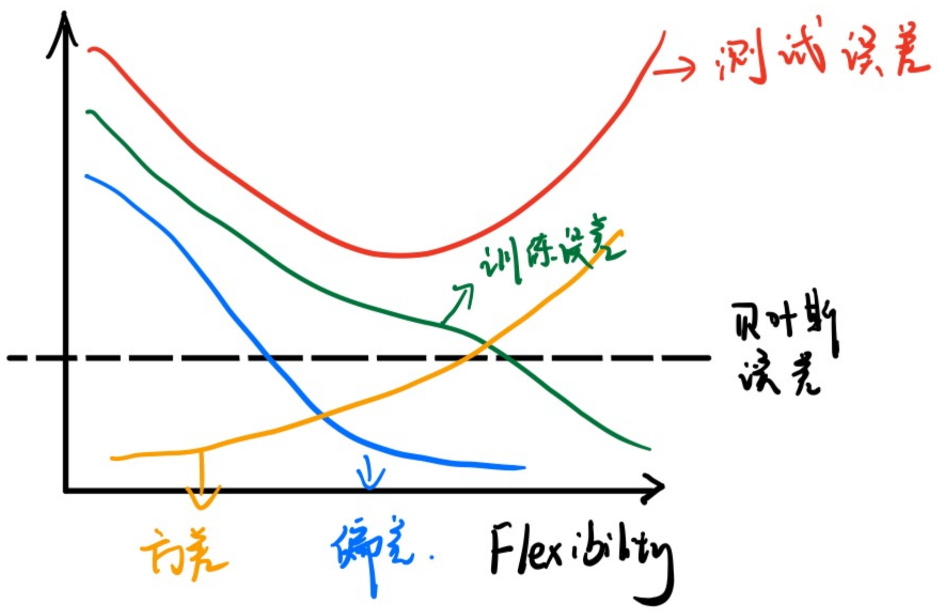
   方差：度量了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响，或者说学习算法的稳定性。

   训练误差：模型在训练集上的误差。

   测试误差：测试集上的误差。

   贝叶斯（或不可约）误差：贝叶斯误差也叫最优误差，通俗来讲，它指的就是现有技术下人和机器能做到最好的情况下，出现的误差。比如图像识别和语音识别这类处理自然数据的任务，人类水平和贝叶斯水平相差不远，通常用人类水平来近似成贝叶斯水平，也就是说人的误差可以近似地看成贝叶斯误差。

   偏差 = 贝叶斯误差 + 可避免偏差



4. 略
5. 一个光滑度高的回归模型或者分类模型，能够更好的拟合非线性模型，偏差更小。但是模型越光滑，所需要计算的参数就越多，而且容易过拟合，方差更大。当我们更想预测，而不是推断的时候，我们优先考虑光滑度高的模型。光滑度低的模型相反。
6. (a) 参数方法是一种基于模型估计的两阶段方法。优点：它把估计$f(\cdot)$的问题简化到估计一组参数，对f假设一个具体的参数形式将简化对$f(\cdot)$的估计，因为估计参数是更为容易的，不需要拟合任意一个函数$f(\cdot)$。缺点：选定的模型并非与实际的f形式上一致，而且还有过拟合的可能情况。

   (b) 非参数方法不需要对函数f的形式实现做明确说明的假设。相反，这类方法追求的接近数据点的估计，估计函数在去粗和光滑处理后尽量可能与更多的数据点接近。优点：不限定函数$f(\cdot)$的具体形式，可以更大的范围选择更适宜的$f(\cdot)$形状的估计。缺点：无法将估计$f(\cdot)$的问题简单到对少数参数进行估计的问题，所以往往需要大量的观察点。

第三章 线性回归

3.1 相关概念

3.1.1 简单线性回归

1. **简单线性回归（Simple linear regression）**假定X和Y之间存在线性关系，其形式为：

$$Y \approx \beta_0 + \beta_1 X$$

表示Y**对**X**的回归**，其中β_0和β_1分别表示为模型的截距和斜率，被称为模型的**系数（coefficient）**或**参数（parameter）**。在给定数据时，也可表示为：

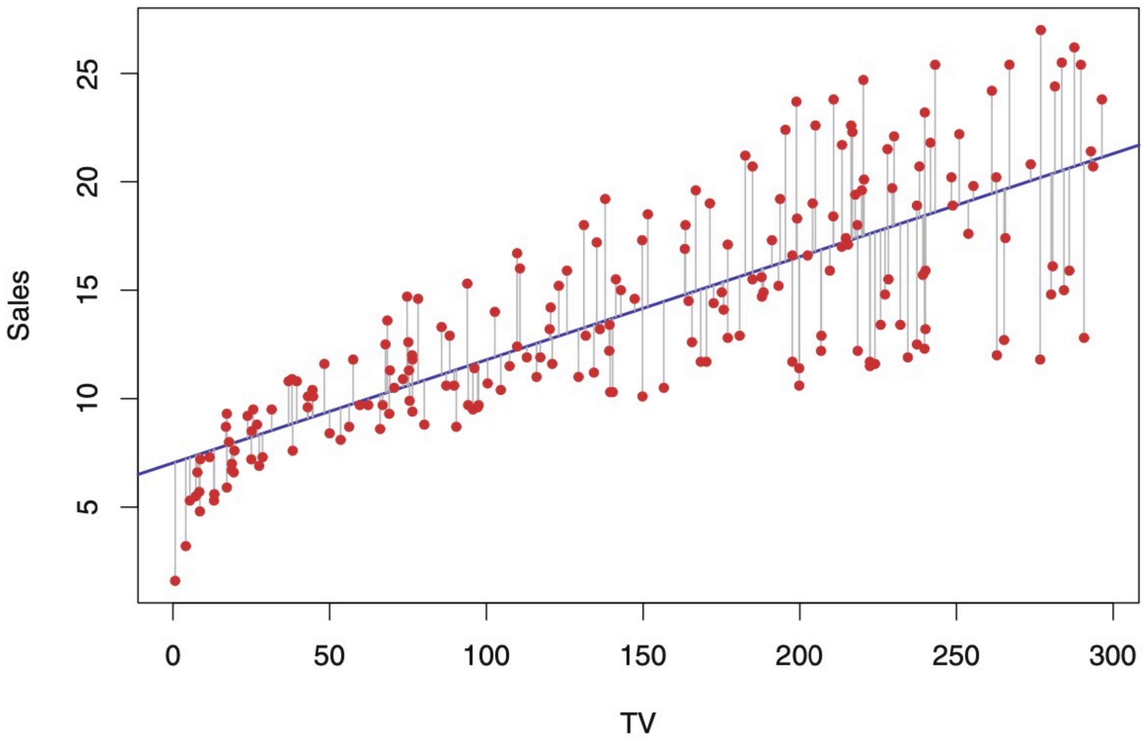
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

其中，$\hat{y}$表示$X = x$的基础上对Y的预测。

2. 评价模型拟合效果可通过测量**接近程度（closeness）**，常用观测的相应值和预测的相应值之间的差距作为参考，定义**残差平方和（Residual sum of square，RSS）**为：

$$\begin{aligned} RSS =& e_1^2 + e_2^2 + \ldots + e_n^2 \\ =& (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \\ =& \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \end{aligned}$$

3. **最小二乘估计（Least squares coefficient estimate）**期望将模型的RSS达到最小，如图所示，其中每条线段代表一个残差。



可通过微积分运算，使简单线性回归的RSS达到最小的参数估计为：

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

其中，$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$和$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$是样本均值。

> **最小二乘估计的推导**
>
> $$RSS = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$
>
> $$\stackrel{\text{对参数求偏导}}{\Rightarrow} \begin{cases} \frac{\partial RSS}{\partial \hat{\beta}_0} & = 2\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \frac{\partial RSS}{\partial \hat{\beta}_1} & = 2\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) \end{cases}$$
>
> $$\stackrel{\text{令其为}0}{\Rightarrow} \begin{cases} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) & = 0 \\ \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) & = 0 \end{cases}$$
>
> $$\stackrel{\text{化简得}}{\Rightarrow} \begin{cases} n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x} & = 0 \\ \sum_{i=1}^n x_i y_i - n\hat{\beta}_1 \bar{x} - \hat{\beta}_1 \sum_{i=1}^n x_i^2 & = 0 \end{cases}$$
>
> $$\begin{aligned} \stackrel{\text{化简得}}{\Rightarrow} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\ &= \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n \bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}^2} \\ &= \frac{\sum_{i=1}^n (x_i y_i - yi\bar{x} - x_i \bar{y} + \bar{x}\bar{y})}{\sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

4. 线性回归是**无偏估计**，同时也遵从估计的**相合性**（遵循格里纹科定理）原则，即如果在特定数据集的基础上估计β_0和β_1，则估计值不会恰等于β_0和β_1，但是，如果对从大量数据集上得到的估计值求平均，他们的均值恰为真值。
5. 一般的估计问题中，可以使用**标准误差（Standard error，写作$SE(\hat{\mu})$）**评价估计的准确性，表示估计$\hat{\mu}$偏离μ的实际值的平均量，形式为：

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

其中，σ是变量Y的每个现实值y_i的标准差。同理，也可以探究$\hat{\beta}_0$和$\hat{\beta}_1$与真实值β_0和β_1的接近程度，形如：

$$SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$$
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

> **系数标准误差的推导：**
>
> 暂略，后续补

6. 标准误差可用于计算**置信区间（Confidence interval）**，对于线性回归模型，β_1和β_0的95%置信区间约为$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$和$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$.
7. 标准误差可用对系数进行**假设检验**，其中最常用的检验包括零假设（X和Y之间没有关系）和备择假设（X和Y之间有一定的关系），使用t统计量测量$\hat{\beta}_1$偏离0的标准偏差，其形式为：

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

对于零假设，即假设$\beta_1 = 0$，计算任意观测值大于等于$|t|$的概率即可，该概率为p**值**，可以解释为：一个很小的p值表示，在预测变量和相应变量之间的真实关系未知的情况下，不太可能完全由于偶然而观察到预测变量和相应变量之间的强相关。因此，**如果p值很小，可以推断预测变量和相应变量之间存在关联，即可拒绝零假设**，典型的拒绝零假设的临界p值是5%或1%.

8. 评价模型的准确性有两个指标，一是**残差标准误（Residual standard error，RSE）**，是对模型中ϵ的标准偏差的估计，形式为：

$$RSE = \sqrt{\frac{1}{n-1}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

RSE被认为是对模型**失拟（lack of fit）**的度量，$\hat{y}_i$与y_i相差很大，那么RSE可能是相当大的，这表明该模型未能很好地拟合数据。

9. 另一是R^2**统计量**，相比较于RSE对数据失拟的绝对测度方法，R^2统计量采取比例（被解释方差的比例）形式，其形式为：

$$R^2=\frac{TSS-RSS}{TSS}=1-\frac{RSS}{TSS}$$

其中，TSS是总平方和，测量了相应变量Y的总方差，可认为是在执行回归分析之前相应变量中的固有变异性，RSS测量的是进行回归后仍无法解释的变异性，因此$TSS-RSS$测量的是相应变量进行回归之后被解释（或被消除）的变异性，则R^2**测量的是Y变异中能被X解释的部分所占比例。**

> **注：** 在简单线性回归中，R^2统计量等价于X和Y的相关系数，即$R^2=r^2$，证明如下：
>
> $$Cor(X,Y)=\frac{\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n(y_i-\bar{y})^2}}$$
>
> $$R^2=\frac{\sum_{i=1}^n(y_i-\bar{y})^2-\sum_{i=1}^n(y_i-\hat{y})^2}{\sum_{i=1}^n(y_i-\bar{y})^2}$$
>
> $$\begin{aligned}
> 令A&=\sum_{i=1}^n(y_i-\bar{y})^2-\sum_{i=1}^n(y_i-\hat{y})^2\\
> &=\sum_{i=1}^n[(y_i-\bar{y})^2-(y_i-\hat{y})^2]\\
> &=\sum_{i=1}^n(y_i-\bar{y}-y_i+\hat{y}_i)(y_i-\bar{y}+y_i-\hat{y}_i)\\
> &=\sum_{i=1}^n(\hat{y}_i-\bar{y})(2y_i-\bar{y}-\hat{y}_i)\\
> &=\sum_{i=1}^n(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})(2y_i-\bar{y}-\hat{\beta}_0-\hat{\beta}_1x_i)\\
> &=\sum_{i=1}^n\hat{\beta}_1(x_i-\bar{x})[2y_i-2\bar{y}-\hat{\beta}_1(x_i-\bar{x})]\\
> &=\hat{\beta}_1[2\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})-\hat{\beta}_1\sum_{i=1}^n(x_i-\bar{x})^2]\\
> &=\hat{\beta}_1\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})\\
> &=\frac{[\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})]^2}{\sum_{i=1}^n(x_i-\bar{x})^2}
> \end{aligned}$$
>
> $$\therefore R^2=\frac{A}{\sum_{i=1}^n(y_i-\bar{y})^2}=\frac{[\sum_{i=1}^n(x_i-\bar{x})(y_i-\bar{y})]^2}{\sum_{i=1}^n(x_i-\bar{x})^2\sum_{i=1}^n(y_i-\bar{y})^2}=Cor^2$$

3.1.2 多元线性回归

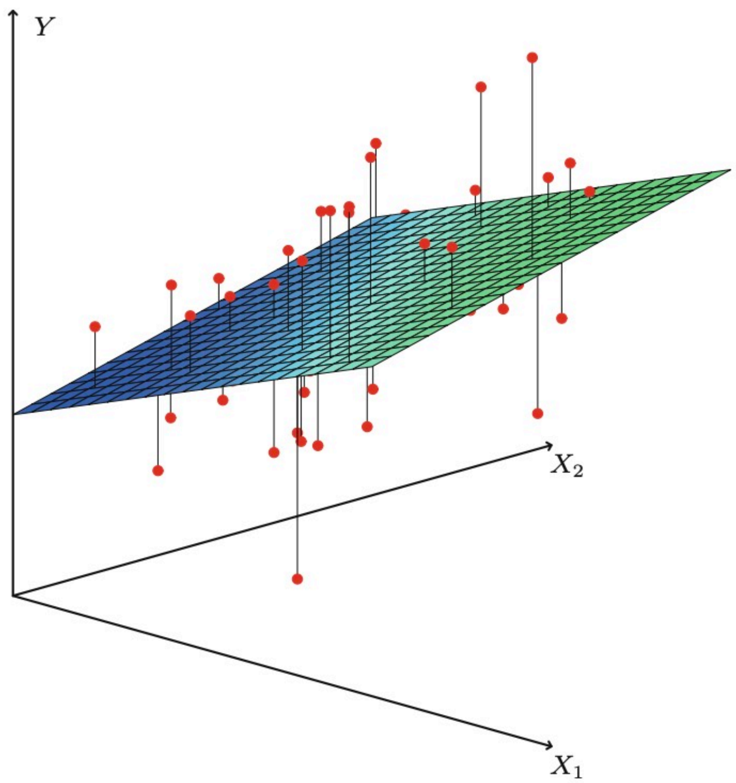
1. 多元线性回归涉及了p个不同的预测变量，该模型形式为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

其中，X_j代表第j个预测变量，β_j代表第j预测变量和相应变量之间的关联。在给定数据时，其参数估计形式为：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

2. 多元线性回归的最小二乘估计原理同简单线性回归一样，期望将模型的RSS达到最小，如图所示



但是需要用矩阵代数形式表示：

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

其中，$\hat{\boldsymbol{\beta}}$为参数向量，$\boldsymbol{X}$为预测变量矩阵，$\boldsymbol{y}$为响应变量向量。

> **最小二乘估计的推导**

> $$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} \\ RSS &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}^T\boldsymbol{W}^T\boldsymbol{y}+\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} \end{aligned}$$
>
> $$\text{由}\ \frac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^T\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}, \quad \frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = (\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{x}\quad \text{得}$$
>
> $$\begin{aligned} \frac{\partial RSS}{\partial \hat{\boldsymbol{\beta}}} &= 0-\boldsymbol{X}^T\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{y}+(\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{X}^T\boldsymbol{X})\hat{\boldsymbol{\beta}} \\ &= 2\boldsymbol{X}^T(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{y}) \end{aligned}$$
>
> $$\begin{aligned} \text{当满秩时，令}\ \frac{\partial RSS}{\partial \hat{\boldsymbol{\beta}}} &= 0 \\ \therefore \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}^T\boldsymbol{y} &= 0 \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \end{aligned}$$

3. 虽然响应变量和预测变量在简单线性回归中具有较高的R^2，但**若增加其他预测变量后，原始的预测变量可能将不具有统计意义（p值低）**，这是因为简单回归模型忽视了预测变量之间的相互关系，可能存在着内部联系。
4. 多元线性回归将关注一下几个重要问题：
   - 预测变量$X_1, X_2, \ldots, X_p$中是否至少有一个可以用来预测响应变量？
   - 所有预测变量都有助于解释Y吗？或仅仅是预测变量的一个子集对预测有用？
   - 模型对数据的拟合程度如何？
   - 给定一组预测变量的值，响应值应预测为多少？所作预测的准确程度如何？
5. 在简单线性回归中，可以使用p值衡量模型的有效性，但p值是针对每一个预测变量的，且**在实际中，即使任何预测变量与响应变量都不相关，但仍有很小的几率使得部分p值小于0.05，因此单独使用t统计量和p值将很有可能错误地得出相关性的结论**。因此需要计算模型整体的评价指标，F统计量：

   $$F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)}$$

   因为$E\{RSS/(n-p-1)\} = \sigma^2$，若零假设为真，$E\{TSS-RSS)/p\} = \sigma^2$，所以**当响应变量和预测变量无关时，$F$统计量应接近1**。一个较大的$F$统计量表示，至少有一个预测变量与响应变量相关，**若n很大，即使F统计量只是略大于1，可能也仍然提供了拒绝零假设的证据，相反，若n很小，则需要较大的F统计量才能拒绝零假设。**
6. 对于本章模型，若$p < n$，则可以使用相应的评价指标，如F统计量等，相反，不可以使用上述的指标，因为无法使用最小二乘法估计模型参数。
7. 在多元线性回归模型中，**最常见的情况是响应变量仅与预测变量的一个子集相关。**因此，需要对预测变量进行筛选，当预测变量很少时，可一个一个迭代筛选，但是数量较多时，则需要以下方法：
   - **向前选择**：从**零模型**（只包含截距）开始，对所有预测变量与响应变量建立简单线性回归模型，并将RSS最小的预测变量纳入零模型中，直到满足某种规则时停止。
   - **向后选择**：从包含所有变量的模型开始，依次删除p最低的变量，循环操作，知道满足某种规则时停止。

- **混合选择**：先做向前选择，在做向后选择。

> **注**：当$p > n$时，不能使用向后选择，而向前选择在各种情况下都适用。向前选择基于贪婪的模式，可将对模型没有“贡献”的变量纳入其中，可使用混合选择方法修正该问题。

8. 在简单回归中，R^2是响应变量和预测变量的相关系数的平方，在多元线性回归中，$R^2 = Cor(Y, \hat{Y})^2$，即是响应值和线性模型拟合值的相关系数的平方。（其实本质上是一样的，在简单线性回归中，$\hat{Y}$只是X的线性变换，并不影响相关系数。）
9. 在多元线性回归中，预测变量间可能存在**协同效应（synergy）**或**交互作用（interaction）**，即组合使用这些预测变量比单独使用预测变量效果更好。

3.1.3注意事项

1. **定性预测变量**。大多数预测变量都是定量的（或者说是**连续型数据**），但有时预测变量会是定性的（或者说是**离散型数据**），如性别（男女）、种族（黄人、白人、黑人）等。以最简单的**二值预测变量**为例，可以将其创建**哑变量**（dummy variable），如基于性别变量创建新变量：

$$x_i = \begin{cases} 1 & \text{女性} \\ 0 & \text{男性} \end{cases}$$

   回归模型可以表示为：

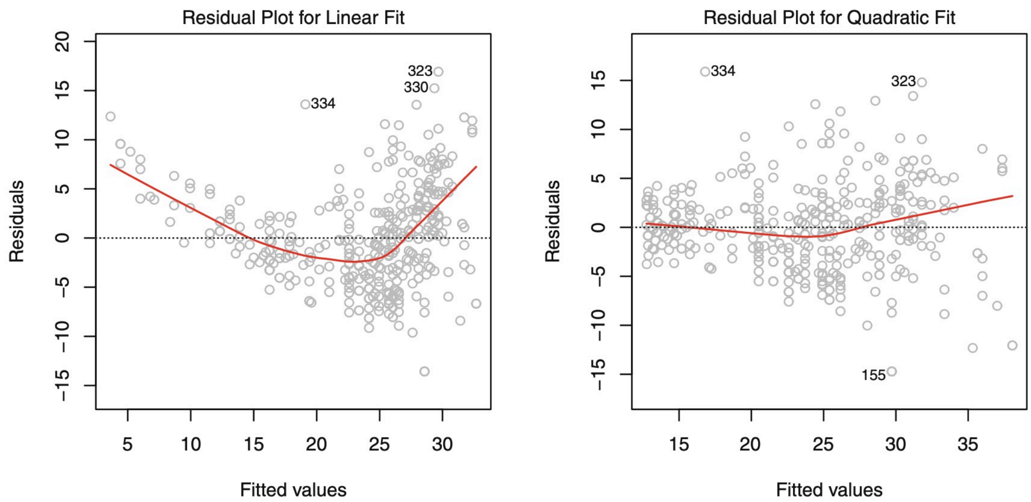
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{女性} \\ \beta_0 + \epsilon_i & \text{男性} \end{cases}$$

   其中，β_0可以解释为男性的平均值，$\beta_0 + \beta_1$为女性的平均值，β_1是男性和女性之间的差异值。

2. 标准的线性回归模型有两个重要的假设，即预测变量和相应变量是**可加**和**线性**的。前者假设预测变量之间是互相**独立**分布的，后者假设无论预测变量取何值，该预测变量引起相应变量的变化是**恒定**的。

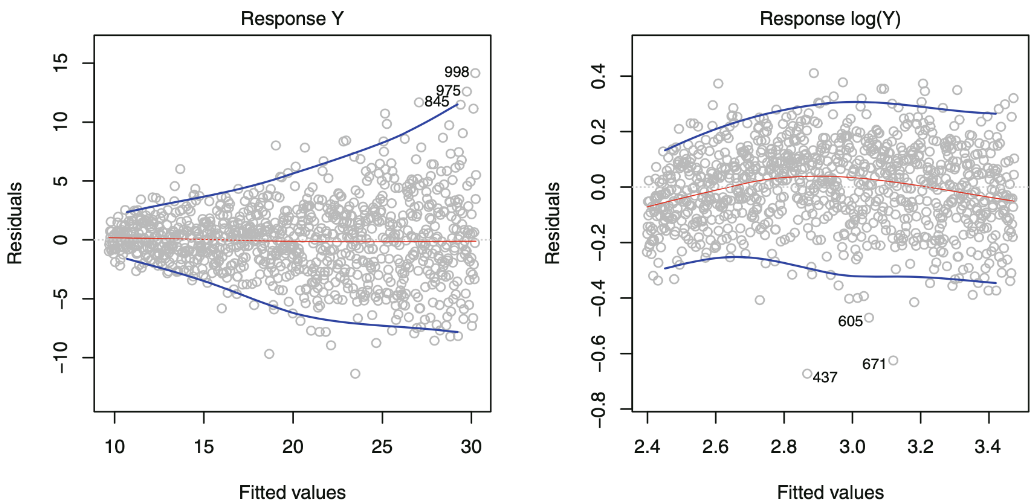
   但是，在现实中并不满足这样的假设，比如预测变量之间**高度相**关，存在协同效应或交互作用，亦或预测变量和相应变量之间的真实关系并是**非线性**的，针对前者需要对变量进一步筛选或降维，后者则需要将模型假设修正为非线性。

3. 下面将介绍线性模型遇到的常见几个问题，分别是**数据本身存在非线性**、**误差项自相关**、**误差项方差非恒定**、**离群点**、**高杠杆点**和**共线性**。
   - **数据本身存在非线性**
     - 实际情况中，很少数据是满足线性的，可以根据**残差图**（Residual plot）识别非线性，如下图所示，左图的残差趋势为U形，表明真实的关系应该是非线性的，当将模型修正为非线性时，残差呈现随机分布，表明该修正提升了模型对数据的拟合度。



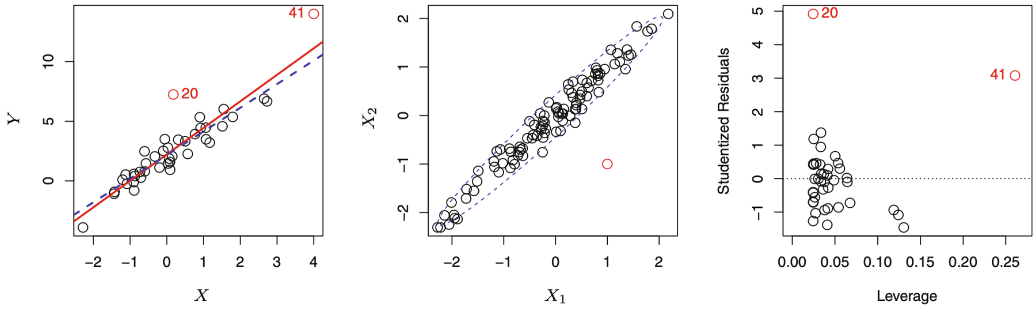
- **措施**：对预测变量使用费线性变换，如$logX,\sqrt{X},X^2$，或者使用更先进的非线性方法。

- **误差项自相关**
  - 如果误差项存在相关性，那么标准误的估计往往会低估了真实标准误。以时间序列或者是地理数据为例，这两类数据最为明显，相邻的观测呈现误差正相关的关系。
  - **措施**：差分法等。
- **误差项方差非恒定**
  - 线性回归模型的另一个重要假设是误差项的方差是恒定的，$Var(\epsilon_i)=\sigma^2$.但通常情况下，误差项的方差并不恒定，可能随着相应值的增加而增加。如图所示，左图残差图呈漏斗形，表明误差方差非恒定。



  - **措施**：使用凹函数对相应值做变换，如$logY$和$\sqrt{Y}$，结果如上图右边所示。
- **离群点**
  - 离群点是指y_i远离模型预测值的点，可通过残差图识别离群点，但是难以使用定量化的方法描述离群点，为解决该问题，引入了**学生化残差**，其由残差e_i除以它的估计标准误得到。**学生化残差绝对值大于3的观测点可能是离群点**。
  - **措施**：直接剔除此观测点。**但是，一个离群点可能不是由失误导致的，而是暗示模型存在缺陷，比如缺少预测变量。**
- **高杠杆点**

- 高杠杆表示观测点x_i是异常的，如下图左边，观测点41具有高杠杆值，因为它的预测变量值比其他观测点都要大。**高杠杆点的观测往往对回归直线的估计有很大的影响（比离群点还大）**。



虽然在各预测变量的取值都在正常范围内，但从整体预测变量集的角度来看，它却是不寻常的，如上图中部所示。可以计算**杠杆统计量**：

$$h_i=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i'=1}^{n}(x_{i'}-\bar{x})^2}$$

一个大的杠杆统计量对应一个高杠杆点，杠杆统计量h_i的取值总是在$\frac{1}{n}$和1之间，且所有观测的平均杠杆值总是等于$(p+1)/n$。

- **措施**：剔除。

- **共线性**
  - 共线性指两个或更多的预测变量高度相关。检测共线性的一个简单方法是看预测变量的相关系数矩阵，但当有多重共线性时（三个或更多变量间存在共线性），更优的检测方法是计算方差膨胀因子（Variance inflation factor，VIF）（不做介绍）。
  - **措施**：1.剔除；2.降维。

3.2 题目答案

1. 电视和广播的低p值表明对于电视和广播的零假设都是错误的。报纸的高p值表明对于报纸的零假设是正确的，即报纸该预测变量不具有统计意义。
2. 差距既是回归和分类任务之间的差距，前者适合连续变量的预测，后者适合离散变量的预测。
3. (a)和(b)略

   (c)错误，不能直接通过回归系数评判交互项的有效性，需要通过该交互项的p值。
4. 一组数据包括单个预测变量和定量响应变量（观测数=100），分别使用线性回归模型和三次项回归模型进行拟合

   (a)假设X和Y满足线性关系，三次项回归模型的训练RSS小于线性回归模型的训练RSS（因为高次模型的误差项更高）；

   (b)条件同(a)，三次项回归模型的测试RSS大于线性回归模型的测试RSS（因为真实的关系为线性）；

   (c)假设X和Y满足非线性关系且具体关系未知，三次项回归模型的训练RSS小于线性回归模型的训练RSS（因为高次模型的误差项更高）；

(d)条件同(c)，则不能判断谁的测试RSS低，因为并不知道真实的关系离线性近还是离三次近。

5. 设$\hat{y}_i = x_i\hat{\beta}$，其中$\hat{\beta} = (\sum_{i=1}^n x_i y_i)/(\sum_{i'=1}^n x_{i'}^2)$，证明：$\hat{y}_i = \sum_{i'=1}^n a_{i'} y_{i'}$

   证：

$$
\begin{aligned}
\hat{y}_i &= x_i\hat{\beta} \\
&= x_i \frac{\sum_{k=1}^n x_k y_k}{\sum_{z'=1}^n x_{z'}^2} \\
&= \sum_{k=1}^n \frac{x_i x_k}{\sum_{z'=1}^n x_{z'}^2} y_k \\
&= \sum_{i'=1}^n a_{i'} y_{i'}
\end{aligned}
$$

6. 在简单线性回归中，最小二乘线通过点$(\bar{x}, \bar{y})$.
7. 证明简单线性回归中的R^2统计量等于X和Y之间的相关系数的平方。

第四章 分类

4.1 相关概念

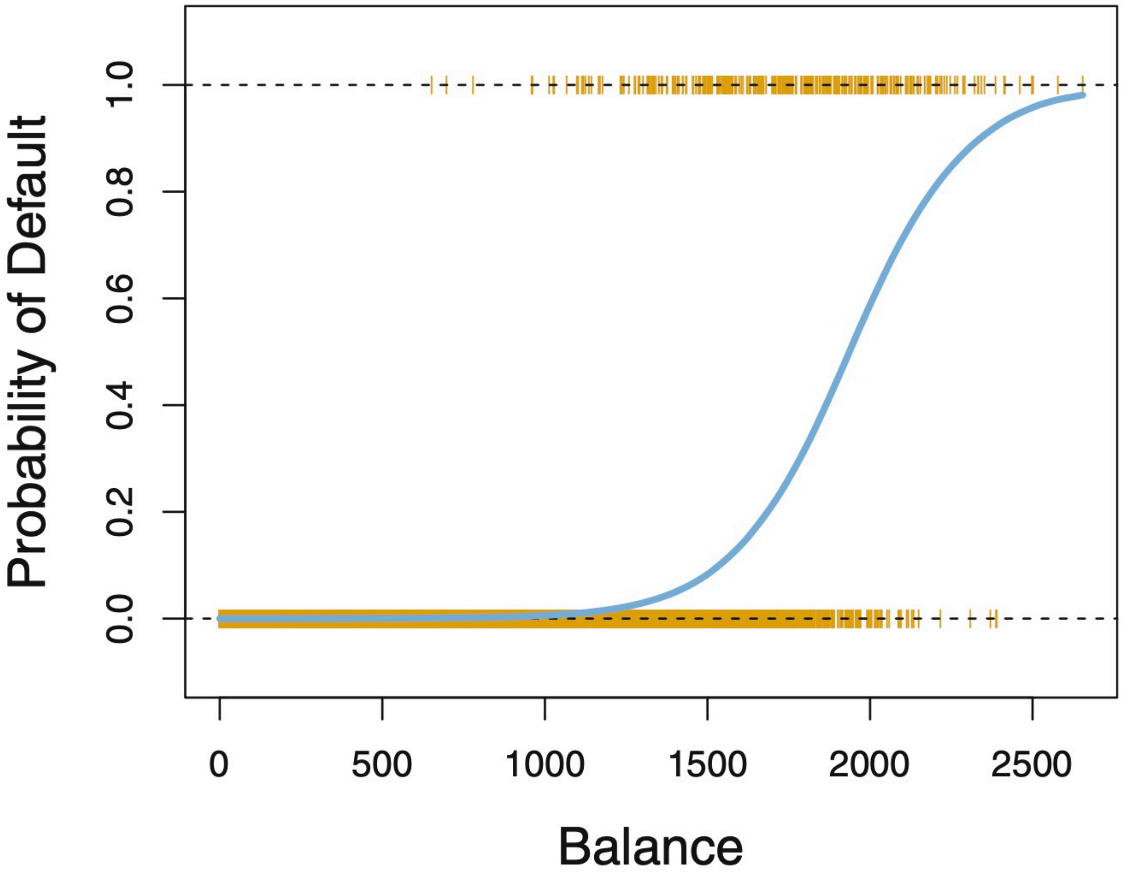
4.1.1 分类问题概述

1. 分类问题是针对定性变量的，大部分基于不同类别的概率，将分类问题作为**概率估计**的一个结果。
2. 当类别数较多（大于2类），则线性回归不具有意义，因为类别数无法被定量表达，例如：1代表红色，2代表绿色，3代表蓝色，则使这些类型具有可度量性，与实际不符，但对于2分类（0-1分类）来说，线性回归具有一定的意义，但预测结果很容易超过0-1范围。

4.1.2 逻辑斯蒂回归

1. **逻辑斯蒂回归（Logistic regression）**可以看成是线性回归的推广（广义线性回归），针对2分类问题，是神经网络中重要部分（激活函数，Sigmoid）。其形式为：

$$
p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{4.1}
$$



当概率超过阈值时，则为正类，小于阈值时为负类。

2. 4.1式可整理为$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$，其中$\frac{p(X)}{1-p(X)}$称为**几率（odd）**，取值范围为0到$\infty$，对其两边取对数可得$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X$，因此，逻辑斯蒂回归可以视为分对数变换下关于X的线性回归模型，逻辑斯蒂模型也较为**对数几率回归**。（参考《机器学习》—周志华）
3. 逻辑斯蒂回归**对Y属于某一类的概率建模**，而不直接对响应变量Y建模。
4. 估计回归系数可使用极大似然估计，似然函数为：

$$L(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

> 推导：

> $$\begin{aligned}
> \text{似然函数：} L(w) &= \prod_{i=1}^{N}[p(x_i)]^{y_i}[1-p(x_i)]^{1-y_i} \\
> \text{对数似然函数：} \mathbb{L} &= \sum_{i=1}^{N}[y_i log\, p(x_i) + (1-y_i)log(1-p(x_i))] \\
> &= \sum_{i=1}^{N}[y_i log\, \frac{p(x_i)}{1-p(x_i)} + log(1-p(x_i))] \\
> &= \sum_{i=1}^{N}[y_i(w\cdot x_i) - log(1+e^{(w\cdot x_i)})]
> \end{aligned}$$
>
> 使用梯度下降算法或牛顿法求解对数似然函数的极大值

> 注：极大似然估计可参考另一个笔记《概率论与数理统计笔记》https://github.com/QianXzhen/Statistics-note

4.1.3 线性判别分析和二次判别分析（LDA和QDA）

> 注：本书中是以统计（贝叶斯决策理论）角度来阐释的，从线性空间角度可以参考《机器学习》—周志华

1. **贝叶斯定理（Bayes theorem）**可以表述为

$$p_k(X) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

其中，$p_k(X) = P(Y=k|X=x)$表示$X=x$的观测属于第k类的后验概率，$f_k(X) = P(X=x|Y=k)$表示第k类观测的X的密度函数，π_k为一个随机选择的观测来自k类的先验概率。**贝叶斯分类起将观测分到$p_k(X)$最大的一类中。**

2. **单预测变量线性判别分析**，假设$f_k(x)$是**正态的或高斯的**，一维情况其密度函数为

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

其中，μ_k和σ_k^2是第k类的平均值和方差，且假设所有**方差是相等**的，则

$$p_k(x) = \frac{\pi_k\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)}$$

化简得到等价式

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

> 证明：
>
> $$\begin{cases} p_k(x) = \dfrac{\pi_k \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)} \\ \delta_k(x) = x\dfrac{\mu_k}{\sigma^2} - \dfrac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \end{cases}$$
>
> $$\text{令} c = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}x^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)}$$
>
> $$\therefore p_k(x) = C\pi_k \exp(-\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k))$$
>
> $$\therefore log(p_k(x)) = log(C) + log(\pi_k) + (-\frac{1}{2\sigma^2}(\mu_k^2 - 2x\mu_k))$$
>
> $$\therefore log(p_k(x)) = (\frac{2x\mu_k}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}) + log(\pi_k) + log(C)$$
>
> $$\because C\text{不随着}k\text{的变化而改变，其值是一个定值}$$
>
> $$\therefore \text{令}\delta_k(x) = (\frac{2x\mu_k}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}) + log(\pi_k)\text{，最大化}p_k(x)\text{等价最大化}\delta_k(x)$$

对于贝叶斯分类只要$\delta_k(x)$达到最大即可。与之相同，因为并不知道原始的参数，需要通过样本估计μ_k、σ和π_k，其中

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k}\sum_{i:y_i=k} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)^2 \\ \hat{\pi}_k &= \frac{n_k}{n} \end{aligned}$$

将这些估计值代入$\delta_k(x)$中，得到$\hat{\delta}_k(x)$，找到令其值最大的类别k作为观测的预测类别。

3. **多预测变量线性判别分析**，其原理和但预测变量相同，但是基于多元预测变量，所以均值和方差都变成了**均值向量**、**协方差矩阵**，多元高斯分布密度函数形式为

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$\delta_k(x)$形式为

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

4. **二次判别分析**与LDA不同之处在于不假设各样本的协方差矩阵（方差）相等，每类观测都有自己的协方差矩阵，形式为

$$\delta_k(x) = -\frac{1}{2}x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1} \mu_k + \log \pi_k$$

QDA和LDA的关系如同非线性回归和线性回归的关系，归属于方差和偏差权衡的问题。

4.1.4 ROC曲线

> 参考知乎回答：https://www.zhihu.com/question/39840928/answer/241440370（作者：无涯）

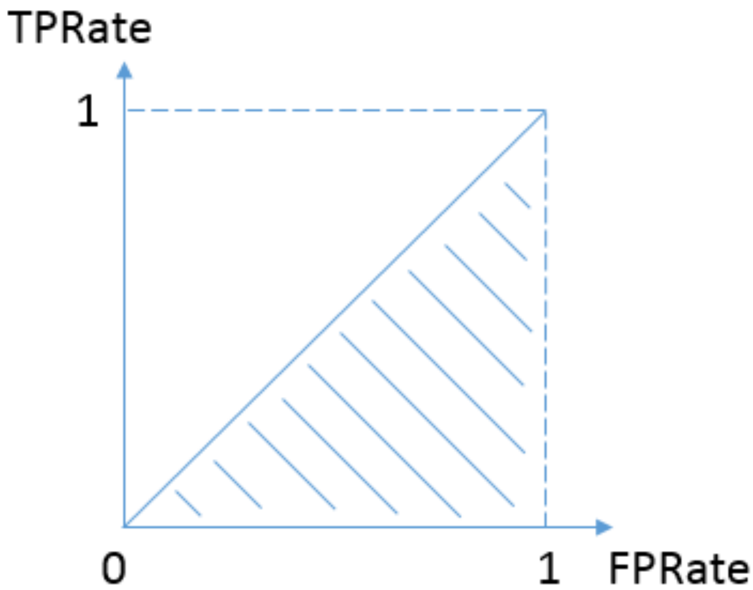
1. 混淆矩阵中有着Positive、Negative、True、False的概念，其意义如下：
    - 称预测类别为1的为Positive（阳性），预测类别为0的为Negative（阴性）。
    - 预测正确的为True（真），预测错误的为False（伪）。

		真实类别	
		1	0
预测类别	1 Positive（阳）	True Positive 真阳	False Positive 伪阳
	0 Negative（阴）	False Negative 伪阴	True Negative 真阴

然后，由此引出True Positive Rate（真阳率）、False Positive（伪阳率）两个概念：

- $TP = \frac{TP}{TP+FN}$，指所有真实类别为1的样本中，预测类别为1的比例；
- $FP = \frac{FP}{FP+TN}$，指所有真实类别为0的样本中，预测类别为1的比例。

2. ROC曲线的横轴是FPRate，纵轴是TPRate，当二者相等时，即y=x，如下图



一个理想的ROC曲线会紧贴左上角，即期望真阳率为1，假阳率为0.

3. AUC（area under the ROC）是ROC曲线下的面积，其最小值为0.5，即上图的面积，一个理想的ROC曲线的AUC为1，AUC的优势是**AUC的计算方法同时考虑了分类器对于正例和负例的分类能力，在样本不平衡的情况下，依然能够对分类器作出合理的评价**。

   例子：

   例如在反欺诈场景，设欺诈类样本为正例，正例占比很少（假设0.1%），如果使用准确率评估，把所有的样本预测为负例，便可以获得**99.9%的准确率**。

   但是如果使用AUC，把所有样本预测为负例，TPRate和FPRate同时为0（没有Positive），与(0,0)(1,1)连接，得出**AUC仅为0.5**，成功规避了样本不均匀带来的问题。

4.2 题目答案

1. 证明逻辑斯蒂函数表达式和分对数表达式等价

$$\begin{cases} P(X) = \dfrac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \\ \dfrac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \end{cases}$$

下证：

$$\frac{p(X)}{1-p(X)} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$$

$$= \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1+e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$$

$$= \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1}{1+e^{\beta_0+\beta_1 X}}}$$

$$= e^{\beta_0+\beta_1 X}$$

2. 贝叶斯将观测分入最大概率类别中$p_k(x)$和$\delta_k(x)$等价，已证。
3. 设一类观测服从均值向量不同、协方差矩阵不等的正态分布，考虑只有一元变量，有K类观测，证明该种情况下，贝叶斯分类起不是线性的，是二次的。

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2)}$$

$$\text{令} C' = \frac{\frac{1}{\sqrt{2\pi}}}{\sum \pi_l \frac{1}{\sqrt{2\pi}\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_l)^2)}$$

$$\therefore p_k(x) = C' \frac{\pi_k}{\sigma_k} \exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

$$\therefore log(p_k(x)) = log(C') + log(\pi_k) - log(\sigma_k) + (-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

所以$log(p_k(x))$是关于x的二次函数。

4. 当变量维数p很大时，只用测试观测附近的观测去做预测的局部方法效果都不理想，这种现象称为**维数灾难**（curse of dimensionality），即当p很大时，非参数模型效果很差。
5. LDA v.s. QDA
   - 如果**贝叶斯决策边界是线性**的，则训练集上QDA比LDA的效果好，测试集上LDA比QDA的效果好；
   - 如果**贝叶斯决策边界是非线性**的，则训练集和测试集上QDA比LDA的效果好；
   - 在一般情况下，随着样本量n增大，相比于LDA的测试预测率，QDA的预测率将变得更好，因为较大的样本量可以抵消方差，避免过拟合；

    - 如果贝叶斯边界是线性的，应该使用LDA，不能因为QDA的光滑度高而选用。

6-9. 略