Week 1

**总体(population)**

一组数据（数字或其他），对应于整个单位的集合，关于这些单位的信息被寻求。

的信息。

**样本(sample)**

在研究过程中实际收集的总体数据的一个子集。

利用样本sample推断总体population

**参数与统计数据**

参数是一个描述总体的数字。符号通常用希腊字母表示μ。统计量是一个描述样本的数字。样本统计量通常用罗马字母x表示。一个参数是一个固定的数字（通常是未知的）。统计量是一个变量，其值在不同的样本中会有变化。

Graph P10 boxplot，histogram, bar plot, scatterplot

同质性

矢量Vector

具有相同基本数据类型的数据元素的序列。

矩阵 Matrix

在一个有行和列的二维数组中的数据元素的集合

非同质的

列表 list

包含其他对象（可能包括其他列表）的一般的结构

数据框架 data frame

用于存储数据，每一列可以是不同的基本类型

所有列必须有相同的长度

Week 2: Regression and smoothing回归和平滑

**Linear Regression**

Y = β0 + β1X + ε --- yi = b0 + b1xi + ei

X是预测者predictor（特征或自变量）。

Y是响应response（目标或因变量）。

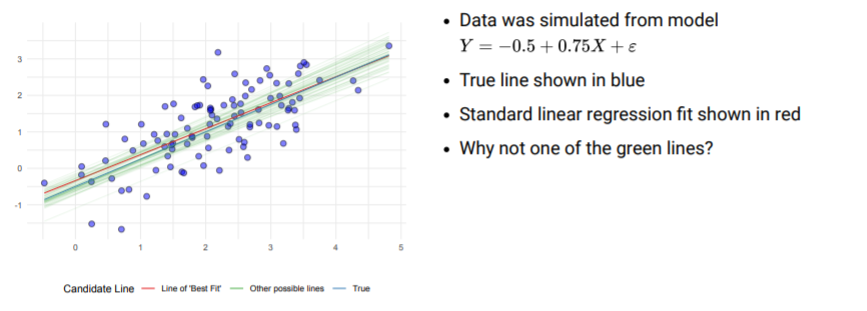
β0是回归线的截距intercept是X=0时y的预期值，

β1是回归线的斜率slope每增加X一个单位，Y平均增加

ε是无法解释unexplained variation的变化或随机误差。

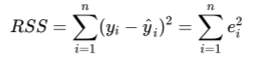
经典的假设是正态分布，均值为零，方差为零

**Performance of linear regression**

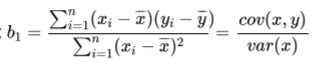


The best estimate of linear regression is least squares criterion 最小二乘法标准

Minimize the residual sum of squares (RSS)



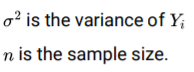
回归（斜率）系数Regression (slope) coefficient:



Intercept: 

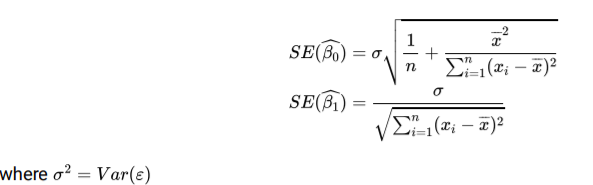
The estimated regression line 

population平均数的标准误差standard error 



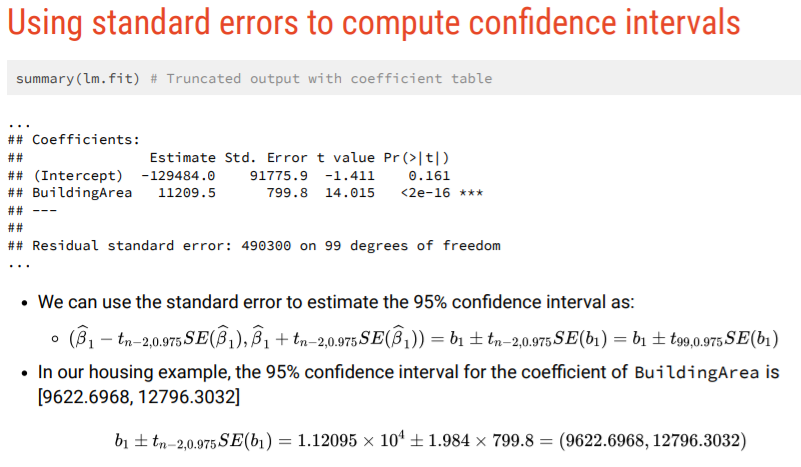
Var(X) = E(X2) − E(X)2

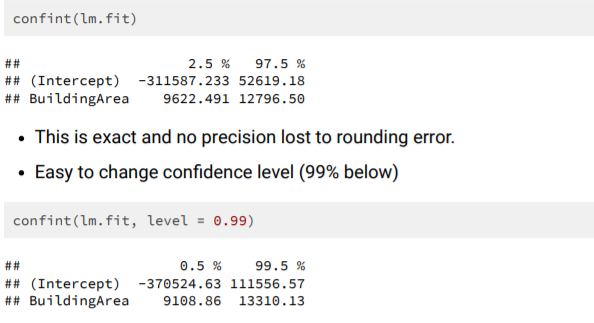
Standard error of regression coefficient estimates



如果这些参数更加分散，标准误差会更小。估计参数的杠杆作用更大

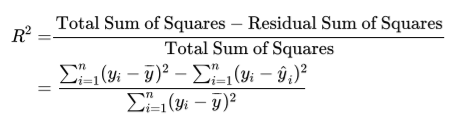
**Confidence intervals**

 95% ==0.975(1-(1-0.95)/2

直接 level = 99%

判定是否predictor 看最后p-value 小于0.05 说明有关联

**Goodness of t statistic拟合度统计**coefficient of determination or R2



R2 的值在0-1中间， 衡量响应Y的变化比例，由对X的线性回归解释。值为0表示Y的方差没有一个可以用X来线性解释。值为1表示Y的所有方差都可以用来线性解释

预测值 by confidence 和prediction

**置信区间估计(confidence interval estimate)：**利用估计的回归方程，对于自变量 x 的一个给定值 x0 ，求出因变量 y 的平均值的估计区间。置信区间是以样本去估算总体

**预测区间估计(prediction interval estimate)：**利用估计的回归方程，对于自变量 x 的一个给定值 x0 ，求出因变量 y 的一个个别值的估计区间。预测区间是指 个体的区间

**预测区间**的范围总是要比**置信区间**的范围要大的。个别值更容易受一些外界因素影响而有差异性，而平均值则相对稳定些。（p52）

Fit improvement提高拟合度 --- 除去异常值outliers

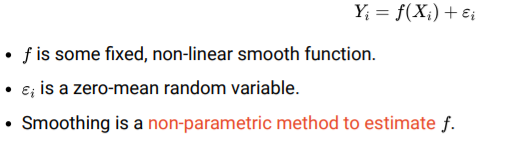
Nonparametric regression or Smoothing

parametric参数化方法包括选择一个统计模型（如线性回归模型），并利用训练数据确定模型的参数（如斜率、截距）。

Nonparametric非参数方法不需要选择一个严格的模型。数据被允许为自己说话。然而，没有容易解释的参数。它们通常是为了描述。而不是正式推断（例如，K-近邻平滑KNN）。

**Data smoothing**

对于预测-反应数据，随机反应变量Y被假定为预测变量x的非线性函数。



**Local averaging**

大多数平滑器（平滑函数）依赖于局部平均的概念。相比之下，简单的线性回归试图拟合出最佳的全局线

generic local-averaging smoother:



**KNN**



回归样条Regression splines P65

拟合分片函数，其中每个函数可以是d维的多项式函数.约束函数是平滑和连续的.

立体花键拟合立体多项式函数，约束条件是。每个结点都是连续的，第一次导数的连续性和第二次导数的连续性

立体花键cubic spline的优点是曲线看起来很平滑，而且可以用来处理几乎所有的函数。

**Loess (local regression)**

一种局部加权的散点图平滑方法,是一种广泛使用的方法，具有良好的稳健性。

它本质上是一种加权流水线平滑法，只是每个局部的流水线是用一种稳健的方法而不是最小二乘法来计算的。因此，该平滑器是非线性的。loess是计算密集型的，需要密集的采样数据。



非参数平滑法与线性回归的比较

非参数平滑法的优点

可以对非线性函数进行建模（例如，splines, loess）。

对数据的函数形式不做任何假设

线性回归的优点

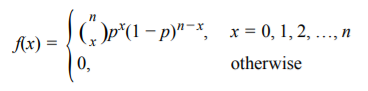
计算效率高，甚至对多变量线性回归也是如此

模型是可解释的，即人们可以知道估计的斜率参数的统计意义。

Week 3 : Density Estimation 密度估计

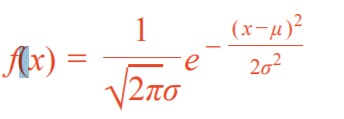
概率密度的总体形状被称为**概率分布 (probability distribution)**，常见的概率分布有均匀分布、正态分布、指数分布等名称。对随机变量特定结果的概率计算是通过**概率密度函数**来完成的，简称为**PDF (Probability Dense Function)**

Discrete distributions离散分布 P74 f(x) = P(X = x), F(x) = P(X ≤ x) (dnorm)

Binomial distribution 二项分布 P75  (dbinom)

Continuous distributions 连续分布

Normal(Gaussian) distribution 正态高斯分布



**Density estimation**

在探索性数据分析中，密度函数的估计可用于评估多模态、偏斜、尾部行为等。

在决策、分类和总结贝叶斯后验中

作为一个有用的可视化工具（一个分布的简单总结）

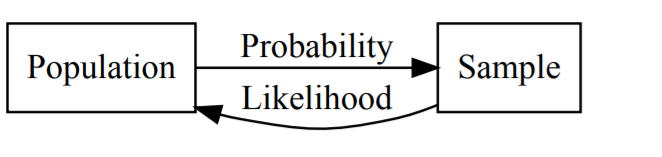
假设随机变量 X1 ，X2， Xn已经被观察到，并假定从密度为f的分布中独立取样

**Parametric density estimation参数密度估计**

密度估计的参数化方法假定了一个参数化模型。

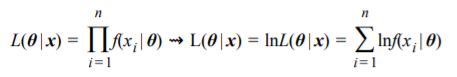
通常，参数θ是用最大似然法maximum likelihood估计的。是获得观察样本的概率最大的一个。

**Likelihood(likelihood 是用 sample 算population) P81**



Probability 已知μ和σ 求概率 用pnorm， likelihood 已知求值 用dnorm

Maximum likelihood



**Non-parametric density estimation**

参数化方法的错误描述的危险

如果假定的fθ 是不正确的。

有推断错误的严重危险。

密度估计的非参数方法

对f的结构假设很少

使用局部信息来估计x点的f

柱状图是非参数密度估计的一种类型 非参数密度估计的一种类型

平行常数密度估计器

由大多数软件包自动生成

**Kernel functions**

核心是一种特殊类型的概率密度函数（PDF），具有以下特性。

**Kernel density estimate核心密度估计**

核心密度估计是一种估计密度的非参数方法

不需要对f的结构进行了解

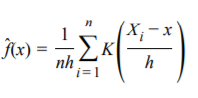
基本上，在每个数据点上，都会以该点为中心创建一个核函数。

通过将所有这些核函数相加并除以数据的数量来估计PDF。

确保它能满足PDF的每个可能值都是非负的和PDF在其支持集上的定积分等于1

Overall density estimate 等于全部weight相加

**Kernel density estimator (KDE) P91 选择bandwidth 选择不同的k**



如果bandwidth太小，密度估计器将倾向于在观察到的数据附近分配太多概率密度。 会造成一个具有许多错误模式的摇摆不定的估计密度函数。

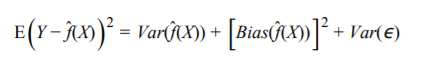
如果bandwidth太大，密度估计器会使概率密度贡献分布得过于分散使得f的重要特征变得平滑

小的bandwidth会产生高的变异

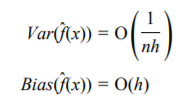
较大的bandwidth会产生较高的偏差

**平均平方误差MSE、偏差Bias和方差Variance**

我们可以将平均平方误差（MSE）分解为三个量的总和。方差variance、偏差square bias和误差的方差variance of the error。



Kernel density estimation



Week 4: High dimensional visualization and analytics

**Clustering**

根据预先设定的标准对相似的观察结果进行分组。

需要一个相似性或不相似性的衡量标准

聚类的目标goal。

我们希望聚类是紧凑compact的。

同一个聚类内的观测值之间的距离要小

不同聚类之间的观测值之间的距离大

算法。

-分层聚类Hierarchical clustering

-均值聚类 k-means clustering

高斯混合模型 Gaussian clustering

**Method**

**1.分区Partitioning**

预先指定数量的互斥和穷举组。

迭代直到满足标准。

**2. Hierarchical分层的 methods.**

聚合式Agglomerative。自下而上，更受欢迎

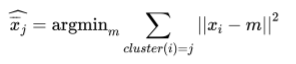
分裂性Divisive。自上而下，不太流行

用树状图显示结果

**k-means approach**

将每个观测值随机地初始化为一个簇。迭代以下内容，直到收敛。

1. 找到群集的平均值cluster means，群集成员关系固定



2. 在集群平均值cluster means固定的情况下，寻找集群成员cluster memberships



**K-means properties**

需要指定群集clusters K的数量。

局部解决方案，不一定是全局解决方案。

取决于起始值（随机起始值）。

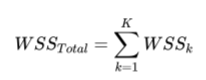
对紧凑的、球形的聚类最好。

当集群大小不同时，效果不好

**Choose K**

方差之和 

这是Euclidean distances的总和，除以聚类中观测值的总数。

WSS为其它点到中心点的距离

K值的选择 看Elbow plot, 下降点越明显然后平缓（拐弯处）的则为最适合的k值

**分层Hierarchical聚类**

开始时，每个观测值都代表一个聚类。在每次迭代中，将两个最接近的聚类合并为一个聚类。需要对两个聚类之间的相似性/不相似性进行衡量。这些措施被称为联系Linkages。

联系Linkages - 衡量两组对象之间的不相似性。

衡量两组对象之间的差异，决定两组对象如何

Single linkage. Complete linkage. Average Linkage

**降维**

多维数据是指特征feature多余观察值observation

**降维**

降维可以是一个预处理步骤，在应用聚类、分类和/或回归之前进行。维数少的数据更容易被可视化和绘制。降维可以是一个有用的探索性数据分析exploratory data analysis工具

**降维策略**

消除或剔除特征： 需要决定哪些特征要被消除？保留那些高变异的特征？

选择特征： 例如，套索和岭回归Lasso and ridge regression。

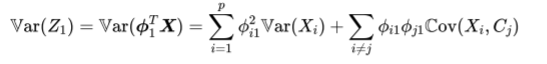
从现有的特征中建立或构建新的特征：用一个特征取代许多现有特征。PCA和t-SNE

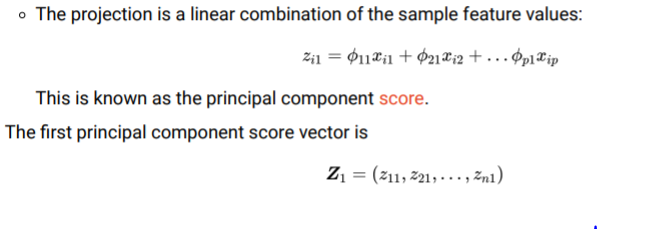
**Principal Components Analysis（PCA）**

第一个主成分是特征的归一化线性组合，使新成分的方差达到最大。



元素ji1被称为第一个主成分的载荷。





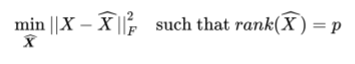
**Scaling the variables（标准化）**

在PCA分析中，通常通过去除平均值来对变量进行中心化处理。你也可以将数据标准化，使所有变量的标准差为1。然而，如果所有的变量都有相同的单位，那么标准化可能就没有必要了

**PCA的属性**

Unique and Global solution!

Ordered components

Best low rank approximation to the data 

可能的最佳线性降维

对非线性关系不是最好的

**带有K-means的PCA**

处理高维数据的非常普遍的方法

使用前M个主成分分数作为Kmeans算法的输入(M<p)

如果数据中的信号能被几个主成分所捕捉，则有助于改善聚类模型组成部分

PCA与回归

在线性回归模型中使用前M个主成分的分数作为预测因子

我们假设少量的主成分可以解释数据中的大部分变异性variability和响应response。

当数据中的变量高度相关highly correlated时，PCA很有用。

**t-SNE**

为实现高维数据集的可视化而开发的非线性技术.利用数据中的局部结构，形成低维表示法

**在T-SNE中的三个步骤**

1. 在高维物体对上构建一个概率分布，使相似的物体被选中的概率很高，而不相似的点被选中的概率极小。物体有很高的概率被选中，而不相似的点有极小的概率被选中。

2. 在低维地图中的点上定义一个类似的概率分布。

3.使两个分布之间的Kullback-Leibler分歧最小化，即关于地图中各点的位置

**Multidimensional Scaling (MDS)（欧式距离）**

视觉上表示低维空间中物体之间的接近性（相似性或距离）。(通常是2或3维空间)

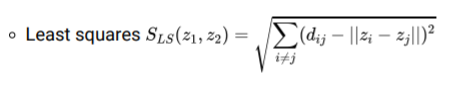
MDS的目的是取一个相似性或不相似性的矩阵，以及投影，其中是所需的低维。

通过优化一个压力函数来保留距离。

不需要完整的数据

这些函数试图迫使低维投影保留原始数据中的距离。

常见的应力函数



不需要完整的数据，只需要它的距离或异质性矩阵

需要选择K

可以作为非线性数据的可视化技术来使用

解释MDS地图。

可以旋转（坐标轴和方向有点随意）。

只有相对位置重要。

通常在MDS地图中寻找接近的物体

**PCA VS T-SNE**

T-SNE是一种概率方法--每次运行它都会给你一个不同的表示。PCA是由一个数学公式定义的

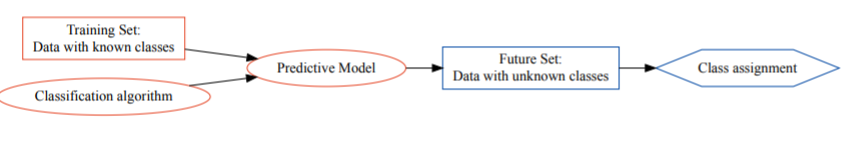
T-SNE主要是一种可视化的方法。来自PCA的PC可以被解释，而T-SNE表示不能用于推理。

T-SNE的计算量比PCA要大。PCA是一种线性方法，所以只能捕捉线性关系，而T-SNE可以发现更复杂的非线性关系。

PCA是一种线性降维技术。它试图保留数据的全局结构Global Structure。与 t-SNE 相比，它poor effect。它不涉及超参数Hyperparameters。它受到异常值outlier的影响很大。PCA 是一种确定性算法Deterministic algorithms。它的工作原理是旋转向量以保持方差Rotate vectors to maintain variance。我们可以找到决定使用特征值保留多少方差variance to retain using feature

t-SNE是一种非线性降维技术。它试图保留数据的本地结构local（集群）。它是最好的降维技术之一。它涉及超参数。它可以处理异常值。它是一种非确定性或随机化算法Non-deterministic or randomized algorithms。它通过最小化 Gaussian 中点之间的距离来工作。我们不能保留方差，而是可以使用超参数来保留距离。

Week 5 classification techniques



聚类：类是未知的，想从数据中发现它们（无监督的unsupervised）。

分类：类是预先设定好的，想用一个（训练或学习）标记的对象集来形成一个分类器来对未来的观察进行分类（有监督(supervised）。

**聚类是无监督， 分类是有监督**

回归：没有类别定义，响应变量是一个连续值。对解释变量和响应变量之间的关系进行建模 解释变量和响应变量之间的关系。

分类：样本被预先定义为来自一个特定的类别。分类模型产生一个连续值的预测，它通常是以概率的形式出现的（即任何一个样本的类属预测值都介于类成员资格的预测值在0和1之间，总和为1）预测的类别是为了做出决定，需要一个预测的类别。

**Classification algorithms**

Logistic Regression

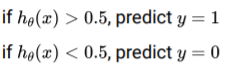
Linear discriminant analysis (LDA)

-nearest neighbors

Supper vector machines (SVM)

**Binary classification**

0: "negative class" 1: "positive class"



线性回归不能使用因为不是0或1

**参数机器学习算法**包括: 逻辑回归，线性成分分析（PCA）

优点：

简洁：理论容易理解和解释结果， 快速：参数模型学习和训练的速度都很快

数据更少：通常不需要大量的数据，在对数据的拟合不很好时表现也不错

局限性：

约束：以选定函数形式的方式来学习本身就限制了模型，有限的复杂度：通常只能应对简单的问题，拟合度小：实际中通常无法和潜在的目标函数吻合

**非参数机器学习算法**decision tree，Bayes，SVM

优势：

可变性：可以拟合许多不同的函数形式。模型强大：对于目标函数不作假设或者作微小的假设, 表现良好：对于预测表现可以非常好。

局限性：

需要更多数据：对于拟合目标函数需要更多的训练数据,速度慢：因为需要训练更多的参数，训练过程通常比较慢。过拟合Overfitting：有更高的风险发生过拟合，对于预测也比较难以解释

**Logistic regression function(需要数值型变量)**

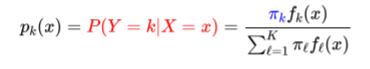


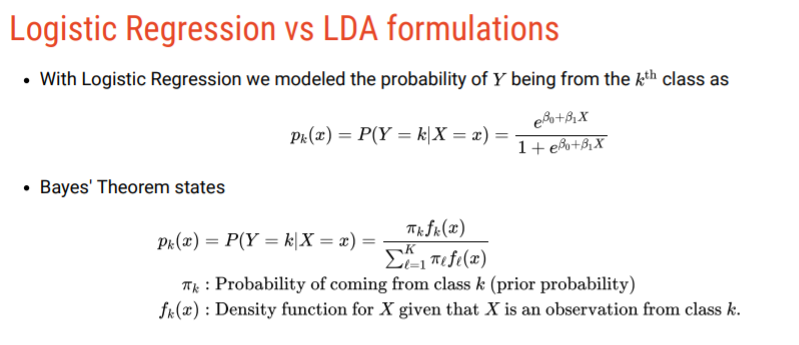
Odds ratio: p/ 1-p

Log-odds or logit: log(p/1-p)

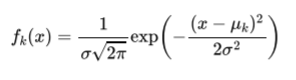
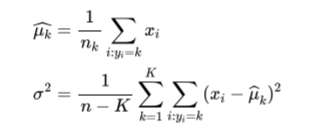
**Linear Discriminant Analysis (LDA)**

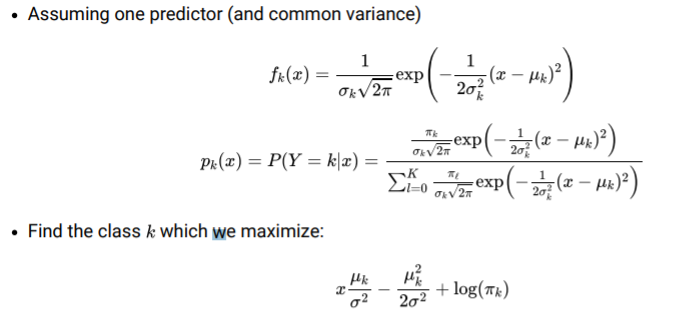
Bayes’ Theorem



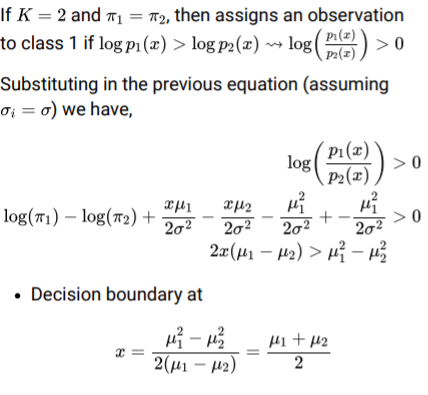


Normal Density (LDA)



**LDA Decision boundary**



**LDA 比 logistic的优势**

在n很小的情况下，预测因子X的分布近似于正态，那么LDA是比Logistic回归更稳定

比Logistic回归更稳定

当我们有两个以上的响应类时，LDA更受欢迎。更加直观地预测类 分配。

当类被很好地分开时，逻辑回归的参数估计是不稳定的。然而，LDA在这种情况下不会出现任何稳定性问题。逻辑回归 VS LDA

相似性。

Logistic回归和LDA都能产生线性边界linear boundaries

不同之处。

LDA假设观测值来自于正态分布normal distribution，每一类都有共同的方差。

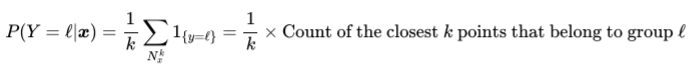
而逻辑回归则没有这个假设assumption。

如果正态性假设成立，LDA会比Logistic Regression做得更好，否则Logistic

回归可能优于LDA

**KNN**

kNN模型是指一个具有特征的观察值属于组的概率取决于与x最近的点的成员资格。



如果k=4 就看与x点最近4个点的比例来判断分类

**kNN vs (LDA and Logistic Regression)**

kNN是完全非参数化parametric的。对决策边界的形状不做任何假设。

kNN的优势。当决策边界高度非线性时，我们可以预期kNN会在LDA和Logistic Regression中占优势。决策边界是高度非线性的

kNN的缺点：kNN不能告诉我们哪些预测因子是重要的（没有协方系数表）

**SVM**

找到一个能在特征空间中分离出各个类别的平面。

如果由于重叠而不可能有一个基本的数学平面

放松完全分离的想法（允许点违反边界）。

丰富和扩大特征空间，使分离成为可能

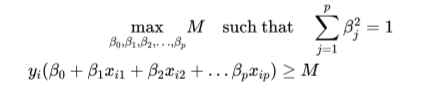
考虑维度扩展

**Hyperplane P180**

在维数为p时，它是一个维数为p-1的仿生子空间 当p=2时，是一条line

Yi = -1 和 yi =1, f(x) > 0 or f(x) < 0

**Maximal Margin Classier**



**Non-perfect separation**

没有一个线性边界（超平面）linear boundary (hyperplane)能完美地分离出各个类别。这通常是指观测值没有一个完美的分离边界的情况。除了在n<p的情况下（特征多于观测值之外

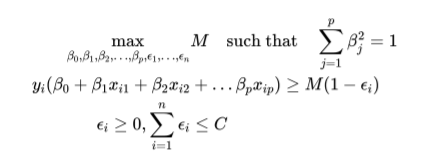
**Effect of noisy data**

数据可能是可分离的，但noisy的不稳定的解决方案的maximal margin classier。

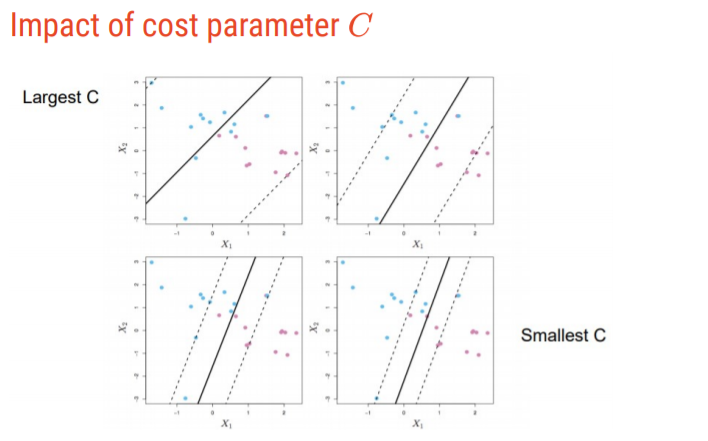
支持向量分类器最大限度地提高了一个软边际soft margin。

放宽了对所有观测值都在余量的正确一侧的要求。

**Support Vector Classier**



如果ϵ>1 则i在错边的超平面，如果0到1，在正确的一侧，但在空白inside margin处，如果等于0，在正确的一侧，并越过边缘。



**Limitation**

单一的线性边界可能是不够insucient的

**Feature space expansion P189**

**非线性和核子**

随着维度的增加，多项式变得非常复杂，并成为一种负担。

更优雅的解决方案是在支持向量分类器中用核子诱导非线性结构non-linear structure in Support vector classier with kernels

其优雅性来自于支持向量分类器中的内积的作用

**SVM with more than two classes p196**

**一对all** 把某个类别的样本归为一类,其他剩余的样本归为另一类，这样k个类别的样本就构造出了**k个SVM，**最终的结果便是这四个值中最大的一个作为分类结果

**优点**：训练k个分类器，个数较少，其分类速度相对较快。  
**缺点**：①每个分类器的训练都是将全部的样本作为训练样本，这样在求解二次规划问题时，训练速度会随着训练样本的数量的增加而急剧减慢；

**一对一：** 在任意两类样本之间设计一个SVM，因此k个类别的样本就要设计**k(k-1)/2个SVM**。

如果K是小的，做一个对一个。否则建议一个对所有。

**LR与KNN， LDA， Bayes， SVM 比较**

Week 6 : Cross validation and bootstrapping

**Training error vs test error**

训练误差是应用于用于训练模型的观测的性能指标。

测试误差是应用模型预测新（测试）观测值的反应时的平均误差。

训练误差的大小通常与测试误差大不相同。

训练误差可以低估测试误差

**Estimate the test error**

调整训练误差以估计测试误差

Cp 选择最小

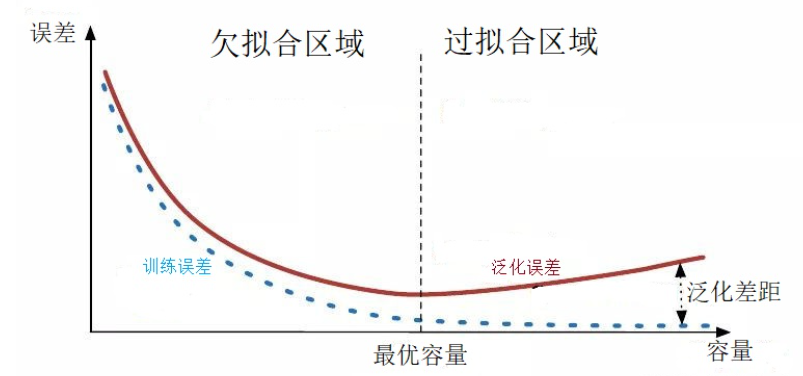
BIC 选择最小

Adjusted R2 选择最大

交叉验证

移除或保留一个观察的子集（测试集），用其余的观察来训练模型。

评估模型在测试集上的表现



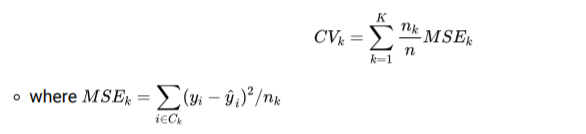
**测试集方法的缺点**

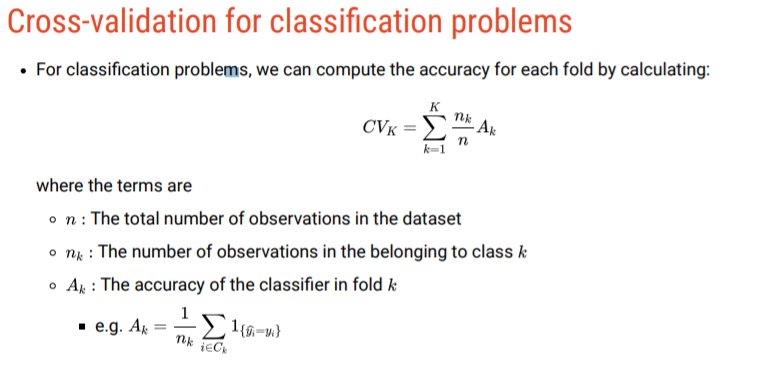
测试误差的估计可能是高度可变的，这恰恰取决于哪些观测值被包括在训练集中，哪些观测值被包括在测试集中。

在测试集方法中，只有观测值的一个子集被用来拟合模型。测试集误差可能倾向于高估整个数据集上模型拟合的测试误差。

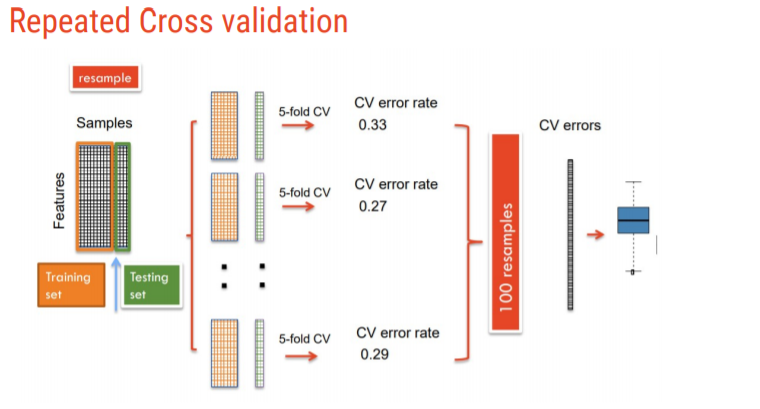
K-fold cross validation

把数据集分成多部份其中一个数据作为测试集，其他作为训练集，重复N次,最后平均

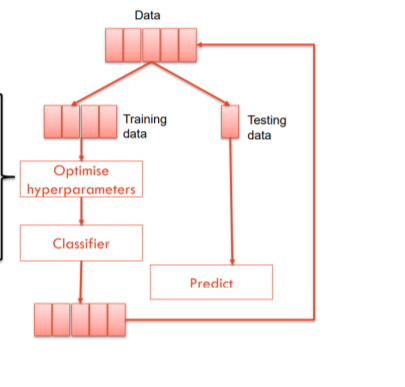




重复多次实施



Nest-cross validation 嵌套交叉验证是通过对基础模型泛化误差的估计来进行超参数的搜索，以得到模型最佳参数



**最终模型的建立**

进行交叉验证的原因是通过估计不同模型的。在未见过的数据上的表现来评估不同的模型

例子。如果你需要在kNN、LDA和逻辑回归以及SVM之间做出选择，那么你可以

用交叉验证法运行这些分类算法，然后挑选出具有最高准确率

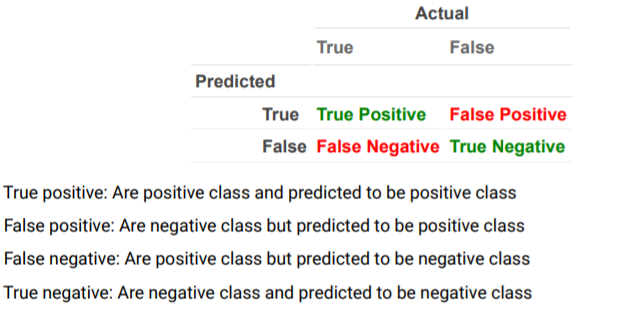
Classification evaluation metrics

**Overall classification accuracy**

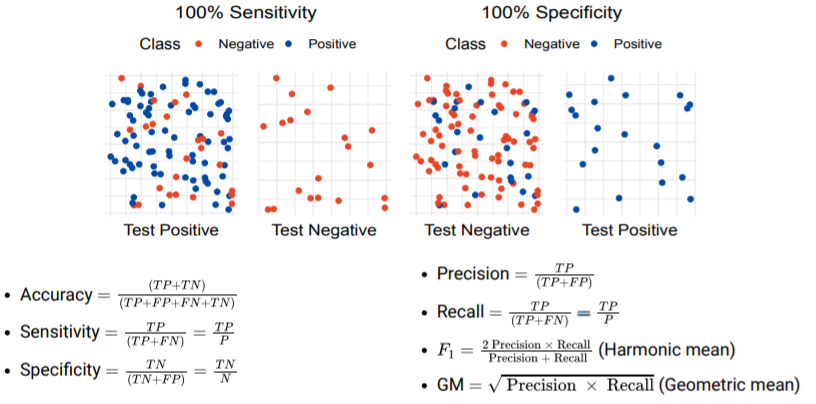
缺点。

对所犯错误的类型不作区分distinction。不考虑每个类别的自然频率e natural frequencies

**Confusion Matrix**



**Sensitivity and Specificity**



Precision是指预测样本正确，实际正确的概率，它显示了模型对阳性案例的准确度。

而sensitivity和recall是指实际正确的样本被预测为正确的概率，它反映了模型的全面性，衡量分类器识别阳性案例的能力

**Accuracy**

binary classification 且正反例不平衡的情况下，模型过于简单导致无法输出错误，而是输出正确的例如真实的癌症患者比例为5%，也就是说100人中，有5个人患有癌症，95个人健康。如果我们建立一个模型，帮助医生建模去做癌症诊断。这个模型很简单，它只会输出‘健康’，而不会输出‘癌症’。将所有病人分类为‘健康’。Accuracy = 95%

**Receiver Operating Characteristics (ROC) curve**

我们可以根据ROC曲线是否在y=x（随机猜测）上方或下方来判别它的性能。若在上面，则性能较好，若在下方，则性能较差。分类器的ROC曲线越靠近(0,1)，则证明该分类器性能越好F1分数是精确性和敏感性的结合和平衡，用来评价分类器的性能。

**AUC**

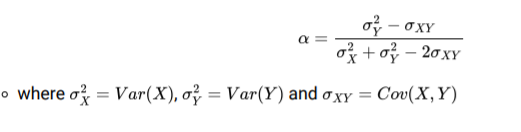
ROC面下面积

<https://www.zhihu.com/question/30643044>

Bootstrap (非参数)

一种灵活而强大的统计工具，可用于量化与给定估计器或统计学习方法相关的不确定性。

例如，它可以提供一个系数的标准误差估计值，或该系数的置信区间



1. 在原有的样本中通过重抽样抽取一定数量（比如100）的新样本，重抽样（Re-sample）的意思就是有放回的抽取，即一个数据有可以被重复抽取超过一次。

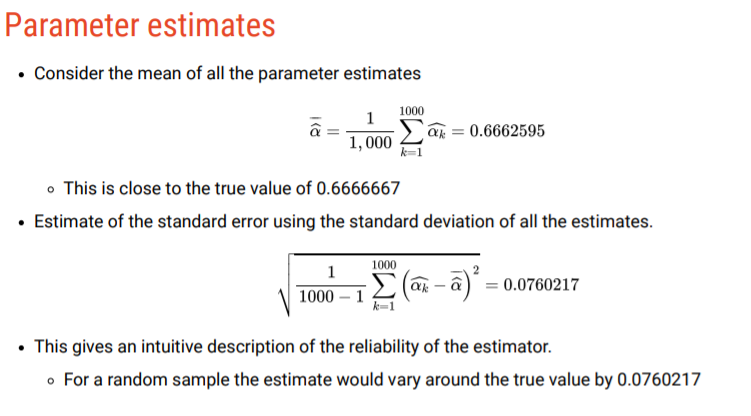
2. 基于产生的新样本，计算我们需要估计的统计量。

在这例子中，我们需要估计的统计量是a，那么我们就需要基于新样本的计算样本方差、协方差的值作为以及，然后通过上面公式算出一个a1

3. 重复上述步骤n次（一般是n>1000次）。

在这个例子中，通过n次（假设n=1000），我们就可以得到1000个ai。

4. 最后，我们可以计算被估计量的均值和方差



Week 7 : Missing data and class imbalance

**Mechanism for missing data**

Missing Completely At Random (MCAR)

例如：假设我们做了一个调查，有些人不愿意在问卷中提供他们的年龄。但这与任何其他变量（包括他们的政党偏好）无关。

Missing At Random (MAR)

一个变量中的数据缺失与该变量无关，而是与其他一些变量有关。例如：在一项调查中，如果来自社会经济地位较低的人可能不太愿意提供工资信息

Missing Not At Random (MNAR)

当数据不是MCAR或MAR时。数据的缺失是由于变量本身的价值，即使在考虑了其他变量

**Dealing with missing values**

对于分类数据，"缺失 "可以是一个类别。例如，在一个调查投票中，如果有人不愿意透露他们想投给谁。

删除有缺失值的数据。有两个选项。省略有缺失数据的变量。省略有缺失数据的观察值。

缺点是你可能会丢掉有价值的信息，或者不经意间把偏差到数据中

添加数据替代缺失值可以用观察到的该特征的平均值来代替缺失值

**Single imputation**

用单一数值代替缺失的数值。

用一个特征的平均值/中位数替换该特征的缺失值。使用预测方法来填补缺失值，例如回归树、kNN。用该特征的最后观察值替换缺失值在单一归因法中，一旦缺失的数据被添加回来，它就被视为有效的观察数据。因此，缺失值数据的不确定性就消失了。

多重归因法

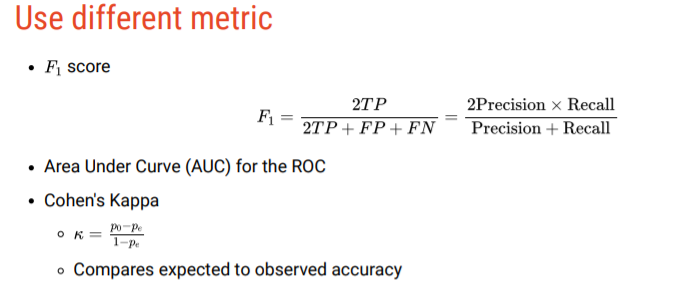
**Multiple imputation**

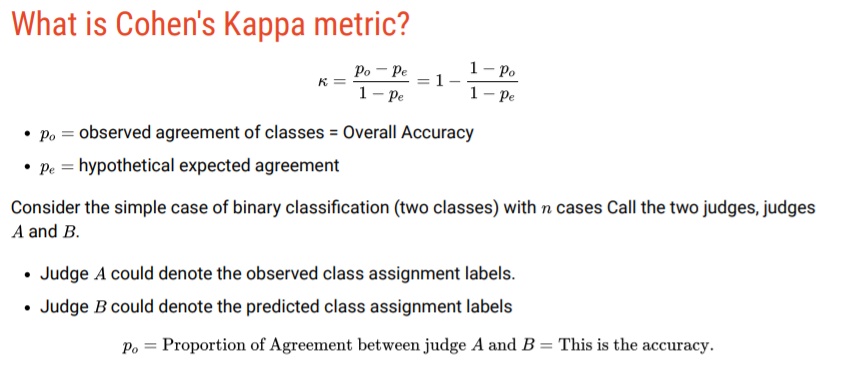
考虑了归因过程中的不确定性。一般来说遵循三个步骤。

对数据进行K次归因（可以使用单一的归因方法进行）。对每一个经过归因的数据集进行分析（如回归）。将K个结果汇集在一起。 链式方程多重归因法（MICE）是一种流行的方法。

Class imbalance

类太少模型没有学习到如何去判别少数类





**Alternatively, modify the input data**

Random up-sampling（增加分类少的observation）

缺点：产生重复的和/或艺术性的实例，给原始数据引入偏见和/或噪音

Down-sampling（删减分类多的observation）

优点是它不会引入重复和/或艺术性的实例劣势。不是所有的数据点都被使用。有可能删除有用的信息。对于类不平衡度非常高的数据来说是更好的选择。

Synthetic Minority Over-sampling Technique (SMOTE)

为少数群体的观察值找到knn。随机选择其中一个最近的knn，然后用它来创建一个类似但随机的新的观察值。在做任过度采样/SMOTE之前，将你的数据分成训练/验证。否则，你会从训练数据集泄露信息到验证数据集。

Week 8 : Feature Selection

**特征选择的方法**

1. Subset selection:

确定一个我们认为与反应或类别有关的预测因子子集。在减少的变量集上拟合一个分类或回归模型。

1. Shrinkage:

主要用于回归模型。拟合一个涉及所有预测因子的模型。一些估计系数被缩减为零。

这种收缩（也被称为正则化）具有减少方差的作用，也可用于特征选择。

1. Dimension reduction

将预测器p投射到M维的子空间, M < p

如果p较大，可以使用best subset selection

**Forward Stepwise selection**

正向逐步选择从一个不包含任何预测因子的模型开始，然后将预测因子逐一添加到模型中。

到模型中，直到所有的预测因子都在模型中。特别是，在每一步中，对拟合有最大额外改善的变量被添加到模型中。

**Backward stepwise selection**

与前向逐步选择一样，后向逐步选择也提供了一种有效的替代方法。最佳子集选择。然而，与前向逐步选择不同的是，后向逐步选择从包含所有预测因子的完整模型开始，逐次迭代地删除最不有用的预测因子。

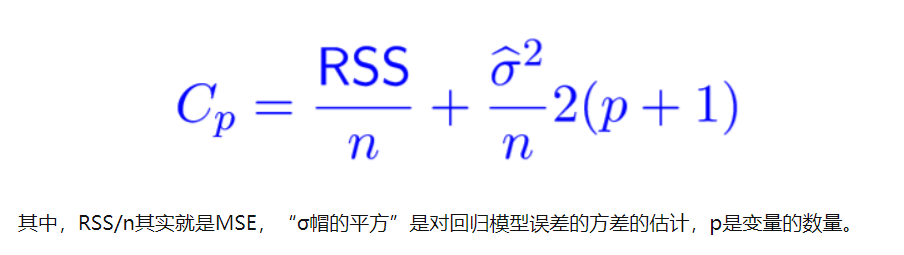
**Estimating test error**

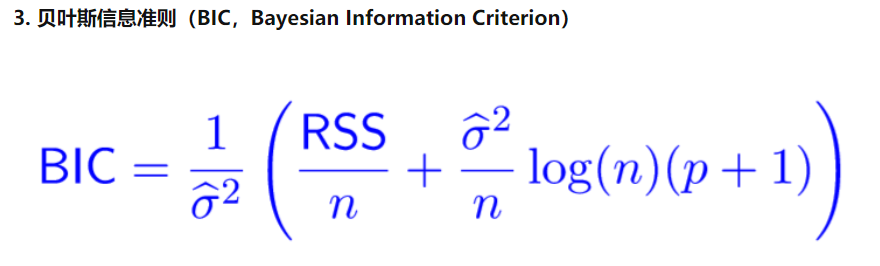
通过对训练误差进行调整来间接估计测试误差。考虑到由于过度拟合而产生的偏差。

直接估计测试误差，使用测试集或交叉验证集的方法

**Indirect approaches**

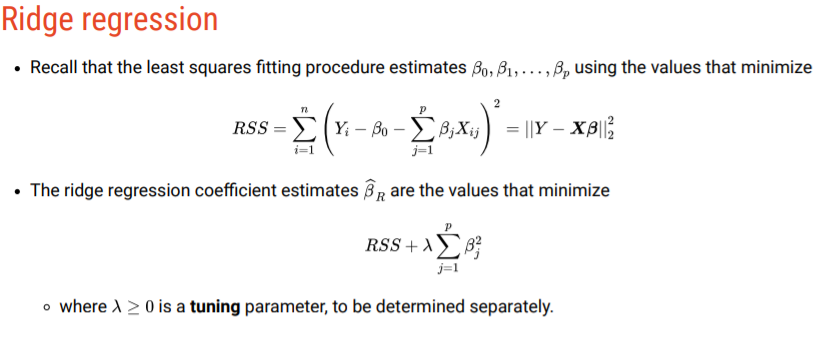
**梅洛 Cp 统计量（Mallow’s Cp statistic）**

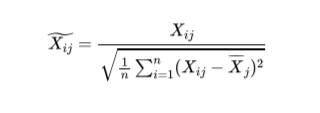




**Shrinkage methods**

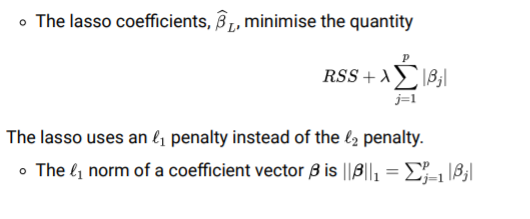
Ridge-regression and Lasso

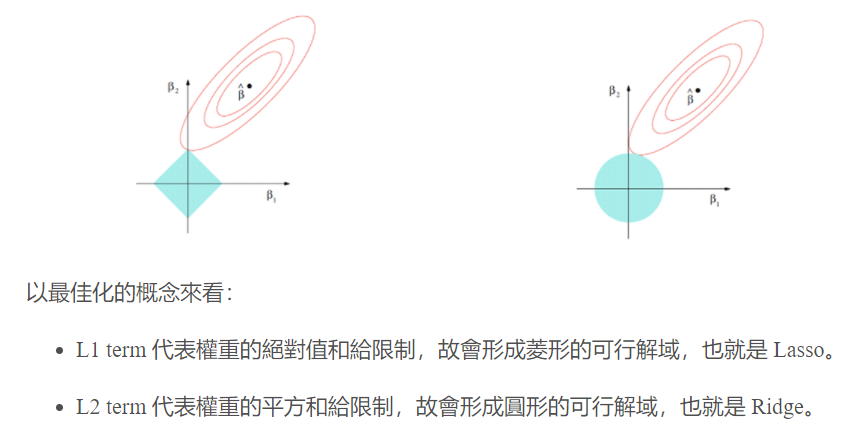




岭回归将在最终模型中包括所有预测因子

**Lasso**





至于子集选择，对于岭回归和套索，我们需要一种方法来确定哪一个的模型是最好的。

也就是说，我们需要一种方法来选择调整参数的值，或者说，选择约束条件的值。

交叉验证法提供了一个解决这个问题的简单方法。我们选择一个值的网格，然后

计算每个值的交叉验证错误率。

然后我们选择交叉验证误差最小的调整参数值。

最后，使用所有可用的观测数据和选定的调谐参数值对模型进行重新测试。

**Elastic net in glmnet**



Week 9 : Tree and Ensemble methods

**Decision tree**

采取一种自上而下的、贪婪的方法，即所谓的递归二元分割法recursive binary splitting。这种方法是自上而下的，因为它从树的顶端开始，然后依次分割预测空间。每次分割都通过树上的两个新分支来表示。它是贪婪的，因为在建树过程的每一步，都会在该特定的步骤中进行最佳分割。

步。而不是向前看，挑选一个在未来某个步骤中会导致更好的树的分裂。

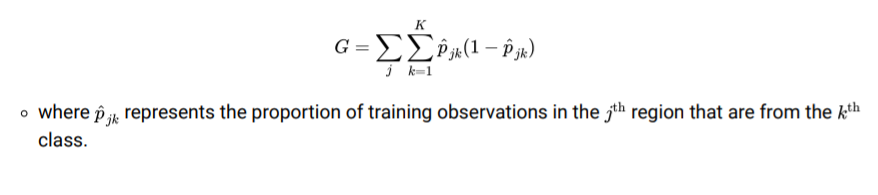
用于分类的决策树与回归树非常相似。

例外的是对**定性**反应的预测，而不是定量反应。

对于一个分类树来说。检查观察结果所属的区域，并预测该区域中最常出现的类别。

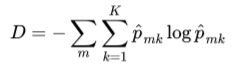
**Gini index**

在分类设置中，RSS不能作为进行二元拆分的标准。如基尼指数被用来代替。



**Cross-entropy**

吉尼指数的一个替代方法是交叉熵，其公式为



**优点。**

树很容易向人们解释。有些人认为，决策树与人类的决策密切相关

树可以用图形显示。可以处理不同的数据类型，不需要缩放

**劣势。**

树的预测准确性不如我们讨论过的其他一些回归和分类方法。

Ensemble methods

**Ensemble of trees**

除了依赖一棵树并希望我们在每一次分裂时都能做出正确的决定之外，还有一个替代方案。

集合方法

将决策树的样本考虑在内

计算在每次分割时使用哪些特征

使用树群的汇总结果建立最终的预测模型。

**Bagging (Bootstrap Aggregation)**

换句话说，对一组观察值进行平均化可以减少方差。由于通常不可能获得多个训练集。我们可以使用引导法来创建多个训练集。使用bootstrapping从一个（单一的）训练数据集中重复取样。在这种方法中，我们产生不同的引导训练数据集。

**Out of bag error estimation P311**

**Random Forest**

随机森林（有些时候）通过一个小的调整，提供了一个比Bagging更好的方法对树木进行装饰。这减少了我们对树进行平均化时的方差。如同袋法，我们在Bootstra**p**的训练样本上建立一些决策树。但在构建这些决策树时，每次考虑树的分裂时，都会随机选择一个从全部p个预测器中选择m个预测器作为分裂候选者。分割时，只允许使用这m个预测器中的一个。

每次拆分都要重新选择预测器，通常我们会选择--即每次分割时考虑的预测器数量大约等于预测器总数的平方根。

**Boosting**

为每个副本建立一个单独的决策树，然后将所有的树结合起来，以创建一个单一的预测模型。Boosting的工作方式与此类似，只是树是按顺序生长的：每棵树的生长都是

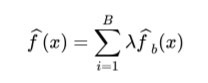
使用来自以前生长的树的信息

与对数据进行单一拟合的大型决策树不同，对数据进行严格的拟合，并有可能过度拟合。而提升方法则是慢慢学习。

考虑到当前的模型，我们将一棵决策树与该模型的残差拟合。然后我们将这个新的决策树添加到拟合函数中，以更新残差residuals.。

这些树中的每一个都可以相当小，只有几个终端节点terminal nodes。

通过将小树拟合到残差中，我们可以慢慢改善它表现不佳的地方。缩减参数shrinkage parameter使这一过程进一步放慢，允许更多不同形状的树攻击残差。



**Parameters to tune in boosting P317**

**AdaBoost**

将一组弱分类器转换成一个强分类器. 在每次迭代中iteration，对数据进行重新加权，将更多的权重放在分类器得到出错的数据点上最后，通过加权组合，将所有的弱分类器结合起来。把更多的权重放在准确率较高的弱分类器

**Stochastic gradient boosting**

在随机提升的每一次迭代中，都要对训练集进行抽样，而不是全部训练集。而不是自举样本（带替换），该算法对训练集的一部分进行抽样。这种引入的随机性可以提高性能

**XGBoost**

支持正则化

可以处理缺失数据

**Bagging vs Boosting**

两种集合方法都是从1个独立建立的学习者中获得N个学习者

**Boosting**中建立的树是弱的学习者（有时只是一个树桩），而随机森林中的树有更高的复杂性，在随机森林中需要调整的参数很少，在Boosting中则更多两者都结合了来自N个树的输出。

**Summary**

决策树是用于回归和分类的简单和可解释的模型。然而，在预测精度方面，它们往往无法与其他方法竞争。Bagging、random forest和Boosting是提高决策树预测精度的好方法。

树。它们的工作原理是在训练数据上生长出许多树，然后结合所产生的树的预测结果。

形成的树的集合。random forest和Boosting--是监督学习的最先进的方法之一。然而，它们的结果可能很难解释

Week 10: Monte Carlo

蒙特卡洛是一类计算方法，可以应用于广泛的问题。蒙特卡洛方法的主要方面是它们依赖于随机抽样，通常提供近似的解决方案而不是精确的解决方案，用于不存在分析解或太难实现的情况下

，蒙特卡洛方法有时被称为随机模拟

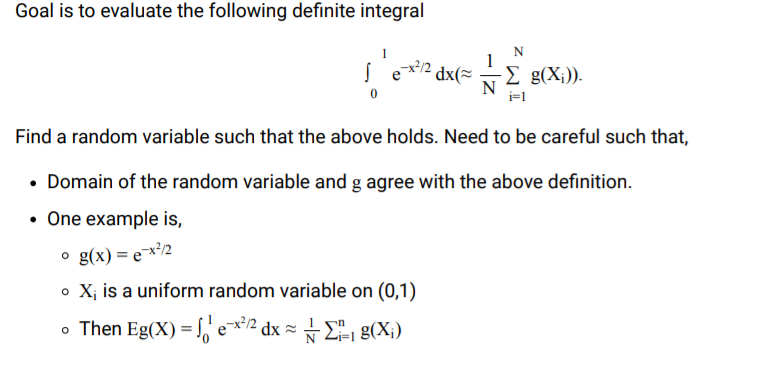
工作原理是通过大量随机样本，去了解一个系统，进而得到所要计算的值。

**Estimation of π**

1. 在单元格内随机抽出N个点。

2. 计算单位圆内的R个点

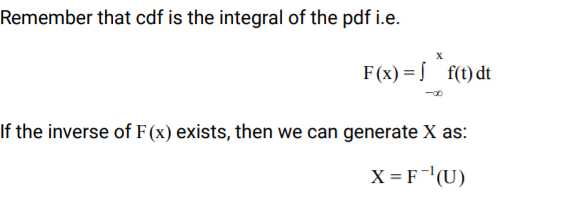
3. 计算R/N的比率并估计为4R/N

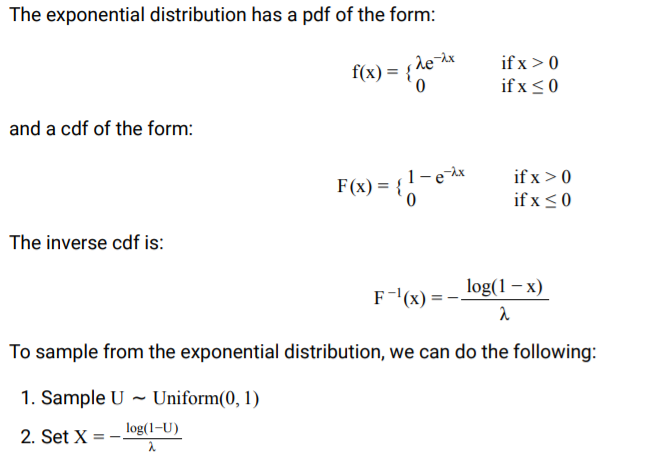


**Simulating random variable**

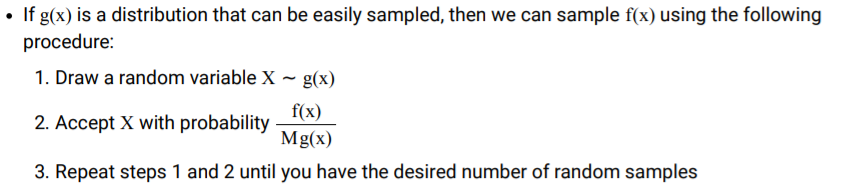
Two methods (others exist):

1. Inverse-transform method





1. Acceptance-rejection method



会抽取大量不需要的样本，因此不是很有效。

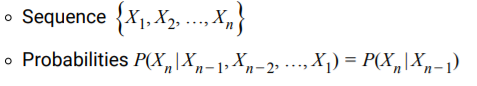
当数据是高维的时候，就会出现很大的问题。

Week 11 : Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC)

是一种用于从任意分布中生成样本的蒙特卡洛抽样技术。区别在于它使用了马尔科夫链。MCMC的两种流行的实现方式是Metropolis-Hastings算法和Gibbs采样器

Markov Chain是一个遵循Markov Chain特性的随机过程. Markov Chain属性意味着过程的未来状态只取决于当前状态.考虑一个依赖性的序列，其中每一个点只取决于刚刚过去的时间。



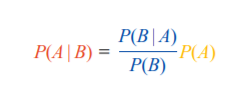
Transition Probability Matrix P363

1. 不可还原. Irreducible --即从每个顶点vertex到另一个顶点都有一条路径

2. 非周期性Aperiodic --即马尔科夫链中没有循环loops。如果这一点没有得到满足，那么系统

将会震荡oscillate

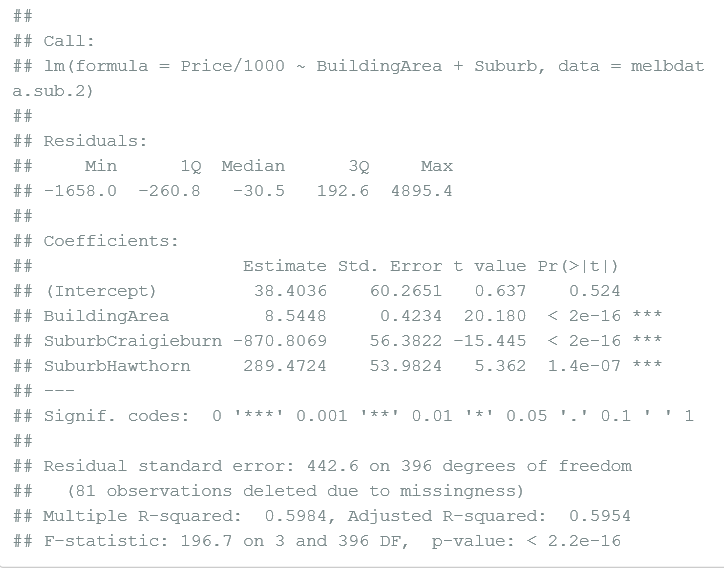
**Metropolis-Hastings**



**Week**

**Week10 week11 要查资料加**

**Linear regression**



**Fit of goodness R-square**

**判别好的预测值 看p-value， 看r-square**

**Logistic regression**

## Estimate Std. Error z value Pr(>|z|)

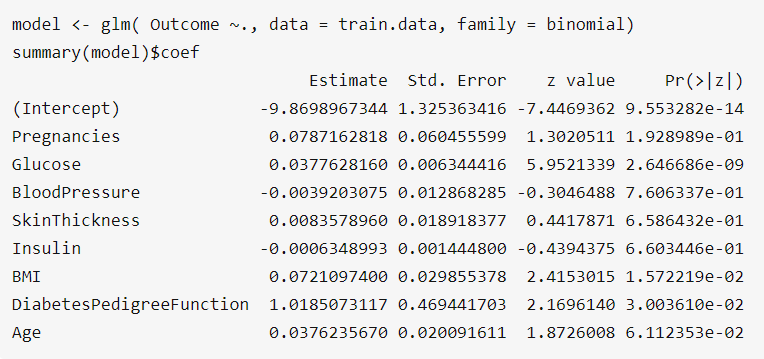
#(Intercept) -6.0593351 0.704588302 -8.599824 7.983874e-18

#Glucose 0.0419662 0.005307411 7.907095 2.634654e-15

上面的输出显示了回归系数β的估计值及其显着性水平。截距（b0）为-6.06，变量血糖的系数为0.042。

逻辑方程可以写成

p = exp(-6.06 + 0.042\*glucose)/ [1 + exp(-6.06 + 0.042\*glucose)]。



Residual deviance 是RSS

Estimate：截距（b0，以上显示为-9.87），即与每个预测变量相关的β系数估计（如Pregnancies的系数为0.078）

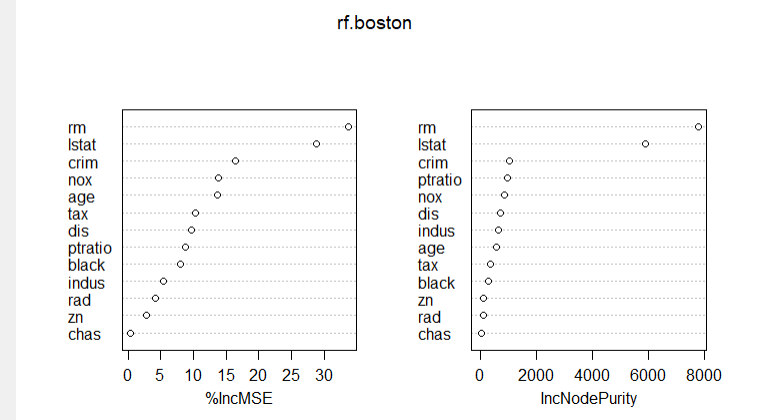
Std.Error：系数估计值的标准误。这代表了系数的准确性。标准误差越大，估计值的准确性越低。

z value：z统计量，即系数估计值（第2列Estimate）除以估算值的标准误（第3列Std.Error）

Pr(>|z|)：对应于z统计量的p值。p值越小，估计值对结果变量来说越重要

log[p/(1-p)] = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn

random forest



“%IncMSE”即increase in mean squared error，通过对每一个预测变量随机赋值，如果该预测变量更为重要，那么其值被随机替换后模型预测的误差会增大。因此，该值越大表示该变量的重要性越大；

“IncNodePurity”即increase in node purity，通过残差平方和来度量，代表了每个变量对分类树每个节点上观测值的异质性的影响，从而比较变量的重要性。该值越大表示该变量的重要性越大。

**Tut 13**

Q4： A