第四章 分类

任浩

材料物理系

renh@upc.edu.cn

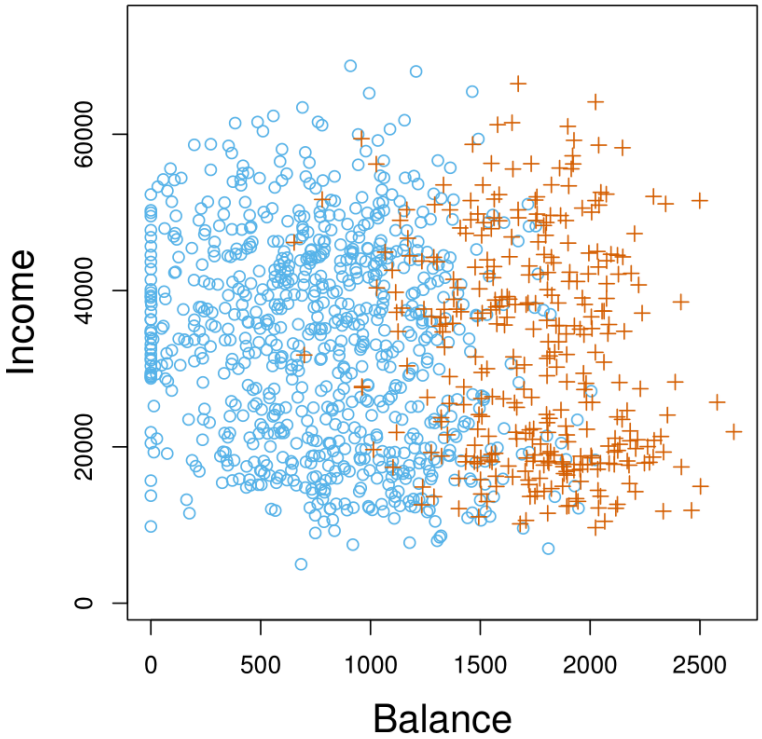
分类（Classification）

- 若需预测的响应值为定性（分类别的）变量➔分类问题
- 分类问题预测或估计某个样本属于某一类的概率，与回归有许多类似之处
- 本章任务:
  - 了解分类问题的基本特征
  - 三种广泛应用的分类方法及其适用范围

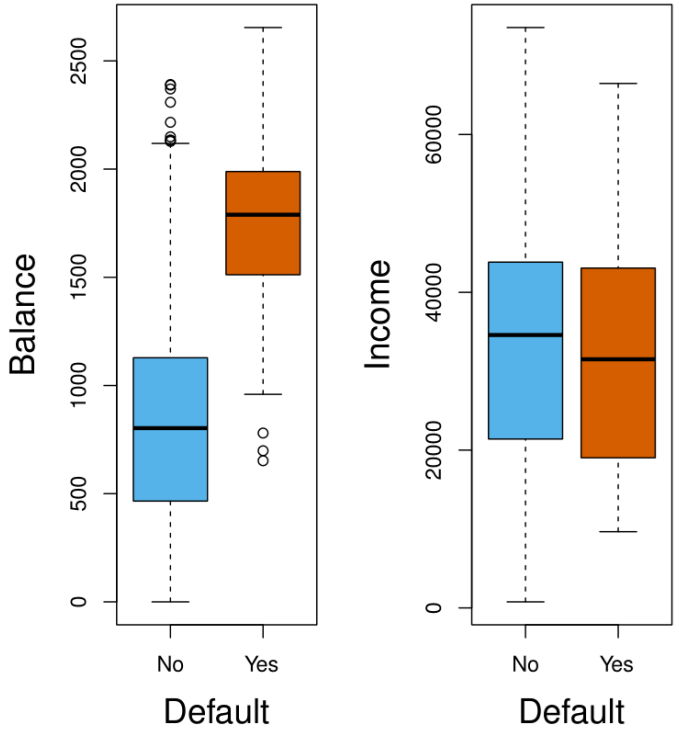
典型分类问题

- 医生根据病人体征及化验数据诊断病情
- 银行的线上服务根据IP地址、收款方等判断交易是否为诈骗
- 邮件系统判断是否为垃圾邮件
- 根据DNA数据判断某生物个体是否有得某种病的风险
- 根据电子结构数据判断某材料是否是高性能催化剂

以信用卡违约数据（Default）为例



Default数据中，信用卡透支（balance）和年收入对违约状态作图



Default关于Balance、Income函数的箱线图

为何不使用线性回归?

假设要通过病人的症状来判断其病情，有三种可能：stroke（中风）、drug overdose（服药过量）和epileptic seizure（癫痫）。使用线性回归有何问题?

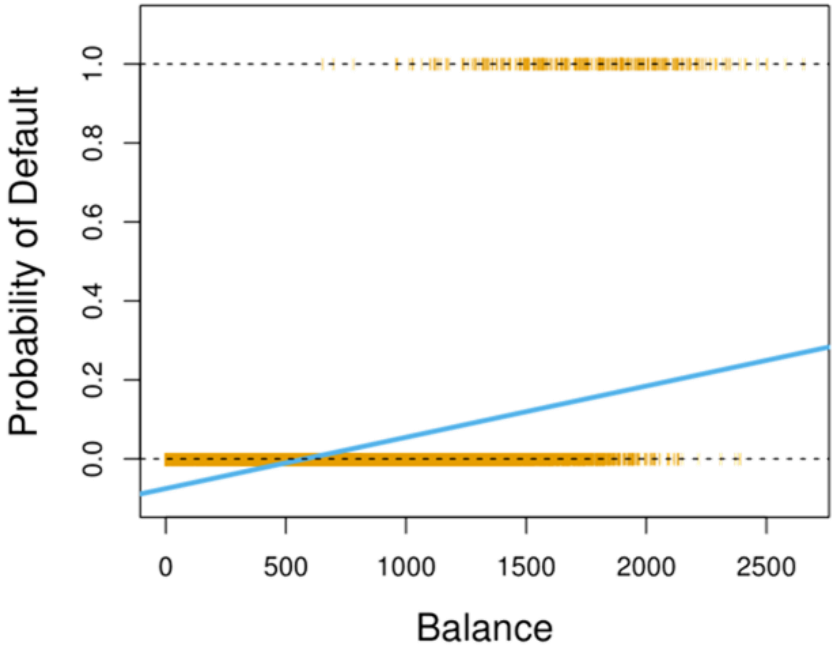
- 需对响应变量Y编码，数值具有定量的意义。

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases} \qquad Y = \begin{cases} 1 & \text{if epileptic seizure;} \\ 2 & \text{if stroke;} \\ 3 & \text{if drug overdose.} \end{cases}$$

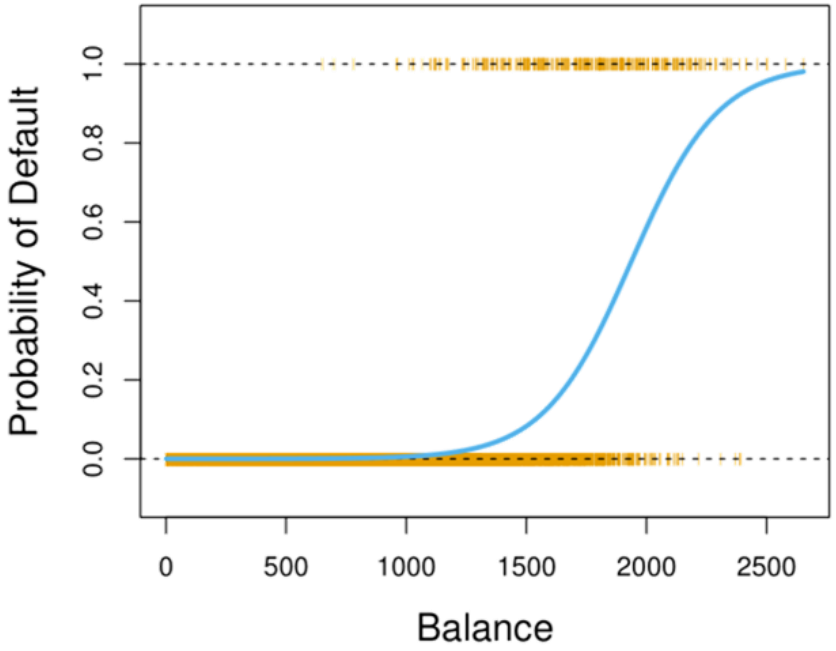
- 对一个二元定性响应变量，该问题影响较小：

$$Y = \begin{cases} 0 & \text{if stroke;} \\ 1 & \text{if drug overdose.} \end{cases}$$

为何不使用线性回归?



线性回归估计的违约概率



逻辑回归预测的违约概率

逻辑回归（Logistic Regression）

Default数据集中，响应变量Y只有两个取值Yes（违约）或No（不违约）。逻辑回归通过预测 Y=Yes 和 Y=No 的概率对样本进行分类。

例如，给定balance时，可以记为：

$$\Pr(\texttt{default} = \texttt{Yes} | \texttt{balance})$$

Pr(default = Yes | balance)是一个条件概率，简记为p(balance)，取值范围在0到1之间。

- p(balance) > 0.5：判定为具有违约风险
- p(balance) < 0.5：无风险

逻辑斯蒂模型（The Logistic Model）

为避免线性回归模型输出的概率结果在0和1之间以外，在逻辑斯谛回归中，使用逻辑蒂斯函数（logistic function）

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

可定义发生比（odds，取值范围$[0,+\infty)$）：

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

上述函数两边同时取对数，得到对数发生比（log odds or logit，Joseph Berkson, 1944， **log**istic un**it**）

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X$$

求解（估计）逻辑回归系数

- β_0和β_1均未知，需基于有效数据进行训练
- 可采用极大似然法估计系数，似然函数（likelihood function）为：

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

- 即求$\hat{\beta}_0$和$\hat{\beta}_1$，使得违约个体对应p(X) 接近1，未违约个体对应p(X) 接近0
- 以上条件等价于将似然函数最大化

Default数据的逻辑回归模型

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

基于估计的回归系数，对任意给定的透支额度可计算违约概率。例如：当某人的透支额度为1000美元时，预测其违约概率为

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

Default数据的逻辑回归模型：基于学生身份

对于离散特征，可构建哑变量或one-hot编码：

- 是学生：1
- 不是学生：0

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

多元逻辑回归（Multiple Logistic Regression）

响应变量受多个因素影响的情况：类似于简单线性回归推广至多元线性回归，将逻辑回归推广至多元逻辑回归

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

此处$X=(X_1,\ldots,X_p)$是p个特征变量，上述方程可以写为

$$p(X)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}$$

利用极大然方法估计$\beta_0,\beta_1,\cdots,\beta_p$ 。

多元逻辑回归（Multiple Logistic Regression）

为什么表1中学生身份会增加违约概率而表2学生身份会降低违约概率？

表1建立用学生身份预测违约概率的逻辑回归模型的系数估计

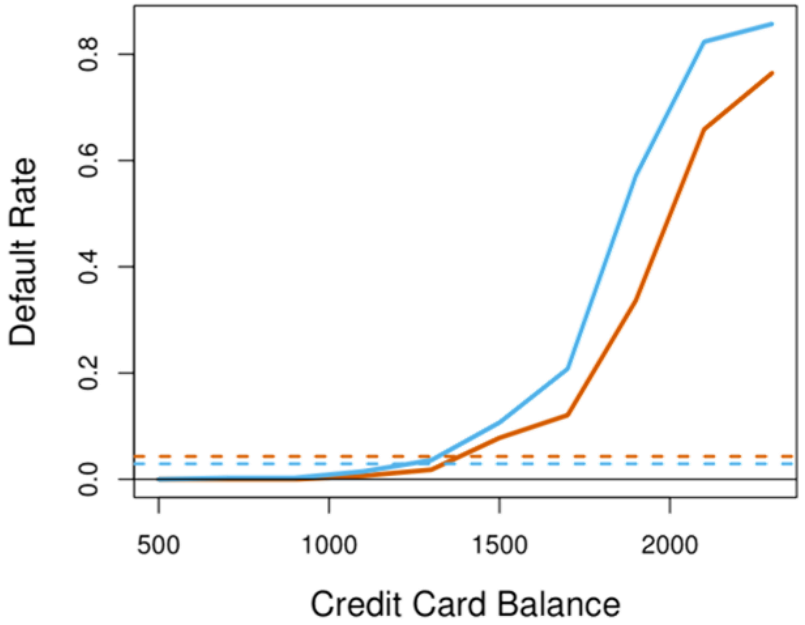
	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

表2结合balance, income, student建立的预测违约概率的系数估计

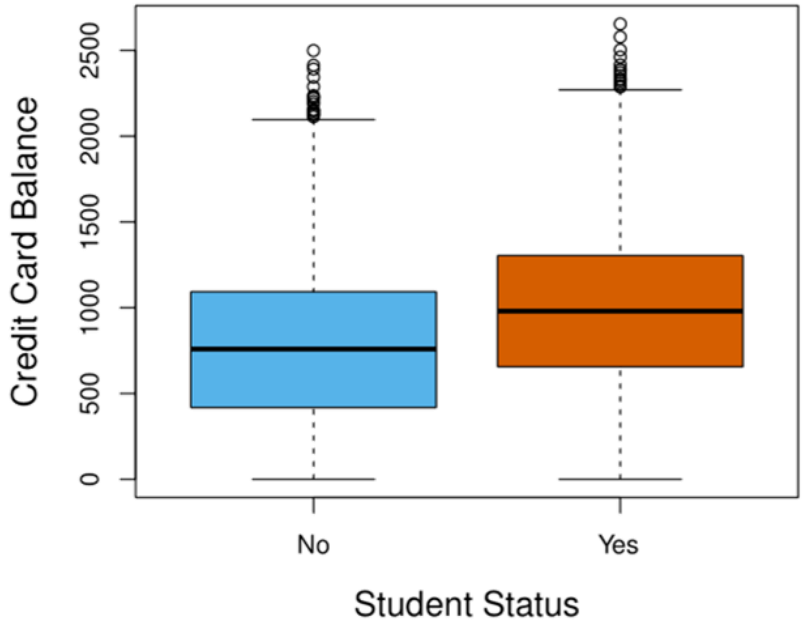
	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

混淆现象

学生身份因素对违约概率的影响



学生身份（橙色）与非学生身份（蓝色）的违约率



学生身份（橙色）与非学生身份（蓝色）balance的箱线图

多类别逻辑回归（multinomial logistic regression）

- 响应变量多于2个类别（逻辑回归中2各类别：是或否）
- 如病情判别：中风，药物过量或癫痫（三类）
- 将2类别的逻辑回归扩展至K类别（K > 2）
- 将某一个类别选为参考（或基线，baseline），如将第K类选为基线：

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

多类别逻辑回归（multinomial logistic regression）

- 相对于第K类，第k类的log odds为

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)=\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p$$

- 隐含：$\beta_{K0}=\beta_{K1}=\cdots=\beta_{Kp}=0$
- 任意两类之间的log odds关于特征变量X是线性的
- 可任选一个类别作为基线，对用同样的X分布，两个类别之间的概率比例不会改变

softmax

- 多类别分类的另一种编码方式
- 任意类别之间的log odds不发生改变
- 常用于现代机器学习实践中
- 不选择某一类别作为基线，同等处理所有类别

$$\Pr(Y=k|X=x)=\frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K}e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=k'|X=x)}\right)=(\beta_{k0}-\beta_{k'0})+(\beta_{k1}-\beta_{k'1})x_1+\cdots+(\beta_{kp}-\beta_{k'p})x_p$$

生成模型（Generative Models）

- 逻辑回归：通过给定的预测变量X直接建立响应变量Y的条件分布模型：基于logistic函数模拟条件概率 $\Pr(Y = k|X = x)$
- 另一个思路：利用贝叶斯定理（Bayes' theorem），先计算每个类别中X的概率分布，结合响应变量的概率分布，计算以上条件概率。
- 当每个类别中X接近正态分布时，两种方法类似。

为何要采用新的方法（线性决策分析，linear discriminant analysis，LDA）？

1. 类别之间差别较大时，逻辑回归模型不够稳定；新方法则不存在此问题
2. 如果样本量n较小，且每个类别中特征变量X接近正态分布，则新的方法更为精确
3. 可自然扩展至多分类情形，且使用更加普遍。

基于贝叶斯统计的生成模型

若Y分为K($K \geq 2$)个不同类别，记π_k为某样本来自第k类的先验(prior)概率，$f_k(X) = \Pr(X = x|Y = k)$表示第k类中X的概率密度分布，则贝叶斯定理可以表示为：

$$p_k(x) \equiv P_r(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}$$

故将π_k和$f_k(x)$代入上式。即可求得$p_k(x)$。称$p_k(x)$为X=x时，Y=k的后验(posterior)概率。

- π_k：可从Y中随机取样，分别得到第k类样本占总样本的比例
- $f_k(x)$：较难获取，通常采用某种近似

单特征线性决策分析（p=1）

- p = 1(只有一个特征变量)，需要知道该特征的概率分布$f_k(x)$
- 通常假设$f_k(x)$具有正态(Normal)或高斯(Gaussian)分布：

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

其中μ_k和σ_k^2是第k类中特征变量的平均值和方差。

通常进一步假定$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$时，即K个类别的特征变量方差相同，简记为$\sigma^2$，并代入贝叶斯定理式中得：

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^2\right)}$$

单特征线性决策分析

对上式取对数，可得：

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

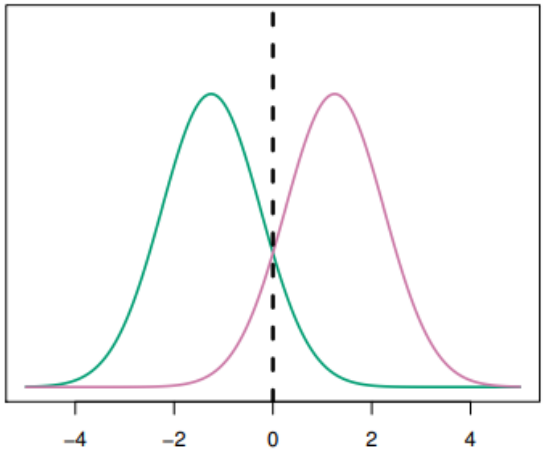
贝叶斯分类器则将样本（X=x）分类至使上式最大的第k组。

例：若K=2，且$\pi_1 = \pi_2$，当$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$时，贝叶斯分类器将观测分入第一类，否则分入第二类；
此时贝叶斯决策边界上的点满足：

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

单特征线性决策分析

例：下图两个一维正态密度函数$f_1(x)$、$f_2(x)$分别为两个不同类别的特征分布，均值和方差分别为$\mu_1 = -1.25$, $\mu_2 = 1.25$，$\sigma_1^2 = \sigma_2^2 = 1$。由于两个分布函数重叠，对于一给定的X=x，对其分类具有不确定性。贝叶斯决策边界上的样本来自两个类别的概率相等，即$\pi_1 = \pi_2$ =0.5，则决策边界为x=0。



实际中通常难以确定每个类别中特征的概率分布，此时可基于已有样本估计参数$\mu_1, \ldots, \mu_k$, $\pi_1, \ldots, \pi_k$, σ^2

单特征线性决策分析

可用训练数据进行参数估计

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2 \qquad \hat{\pi}_k = n_k/n.$$

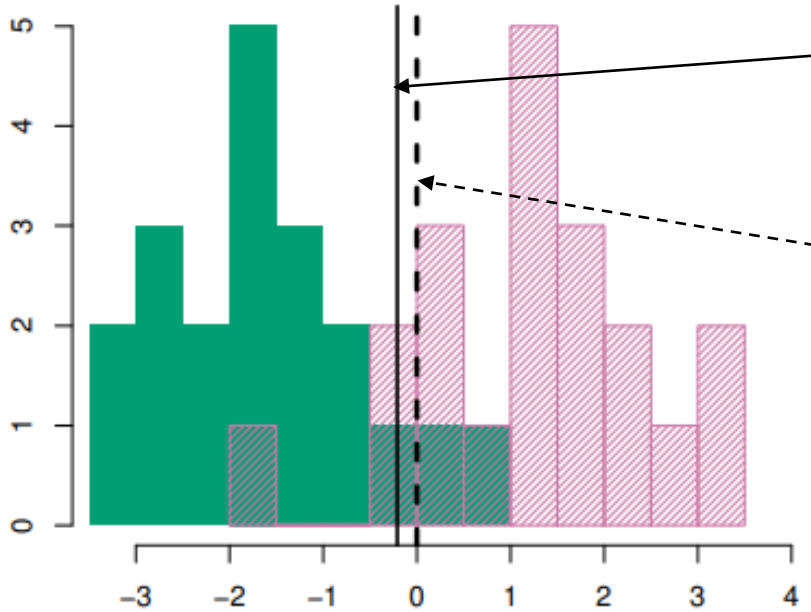
其中n为样本数，n_k为属于第k类的样本数，μ_k的估计即为第k个类别对应的X的均值，$\hat{\sigma}^2$为其标准偏差。将以上估计值代入决策函数，

$$\hat{\delta}_k(x) = x\cdot\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

决策函数(discriminant function) $\hat{\delta}_k$(x)是x的线性函数，即为“线性”决策分析名称的来源

单特征线性决策用于实际数据

每个类取20个随机观测样本的特征分布直方图



黑色实线为LDA决策边界$\frac{\hat{\mu}_1+\hat{\mu}_2}{2}$

黑色虚线为贝叶斯决策边界$\frac{\mu_1+\mu_2}{2}=0$

贝叶斯错误率：10.6%

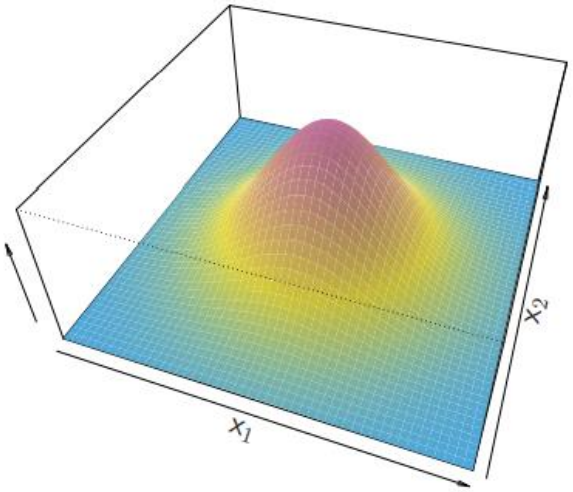
LDA错误率：11.1%

LDA 分类器的错误率只比最小可能的错误率高0.5%

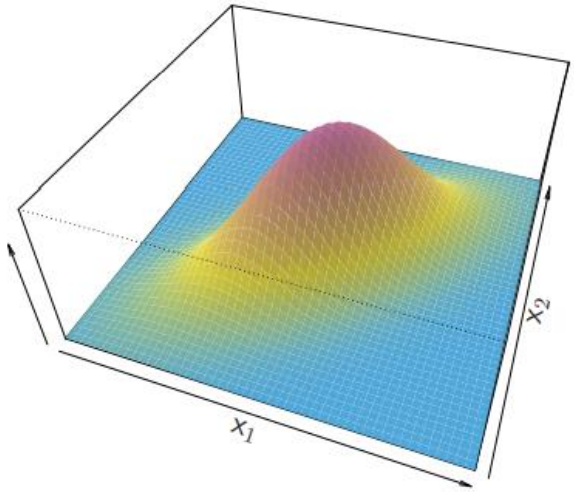
多特征线性决策分析（p>1）

存在多个特征时，通常假设 $\mathrm{X} = (X_1, X_2, \ldots, X_p)$ 服从多元高斯分布（或多元正态分布），每个类别特征具有不同均值；不同特征之间的关联用协方差矩阵描述（$p \times p$），假设不同类别的特征具有共同的协方差矩阵

$\mathrm{p} = 2$的高斯分布



$\mathrm{Cor}(X_1, X_2) = 0$



$\mathrm{Cor}(X_1, X_2) = 0.7$

多特征线性决策分析

若p维随机变量X服从多元高斯分布，则记为$X \sim N(\mu, \sum)$，其中$E(X) = \mu$是X的均值，$Cov(X) = \sum$是X的协方差矩阵，多元高斯分布函数形式上定义为：

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

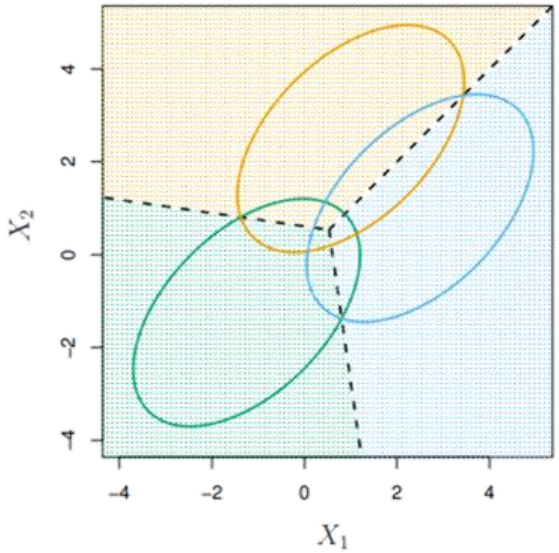
将第k类的密度函数$f_k(X = x)$代入贝叶斯定理，可得贝叶斯分类器将样本$X = x$分入下式值最大的一类：

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

多特征线性决策分析

例：三个类别，样本量相同，均为高斯分布，均值不同，协方差相同。三个椭圆为各自的95%概率边界。虚线是贝叶斯决策边界，满足$\delta_k(x) = \delta_l(x)$，即：

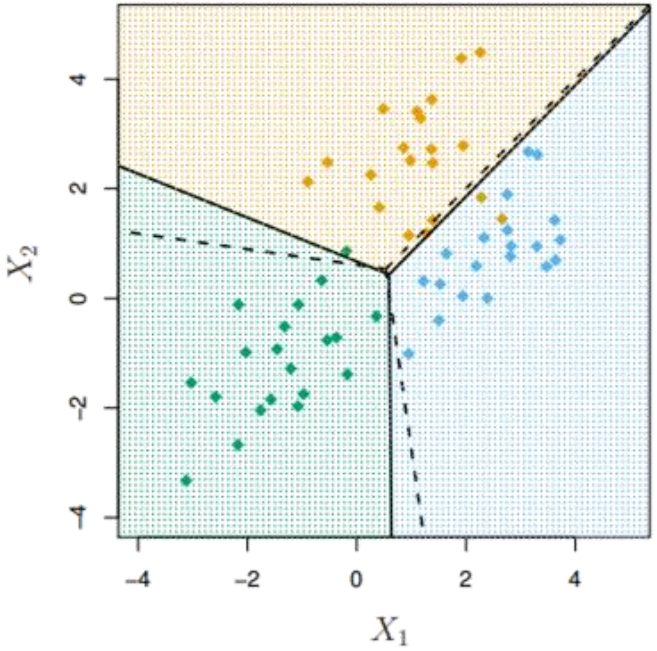
$$x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k = x^T\mathbf{\Sigma}^{-1}\mu_l - \frac{1}{2}\mu_l^T\mathbf{\Sigma}^{-1}\mu_l$$



三条贝叶斯决策边界，是因为三类里有三次的两两比较。三条边界将预测变量空间分为三个区域，贝叶斯分类器会根据观测落入哪个区域来对观测进行分类。

多特征线性决策分析

- 实际应用中，同样需要依据已有样本估计参数$\mu_1,\ldots,\mu_k,\pi_1,\ldots,\pi_k,\sum$
- $\hat{\delta}_k$是一个关于x的线性函数，也就是说LDA决策规则只依赖于x与其他元素的线性组合



- 三类各取20个样本，黑色实线为LDA决策边界，虚线为贝叶斯决策边界。
- 贝叶斯测试错误率和LDA测试错误率分别为0.074 6和0.0770

混淆矩阵（confusion matrix）

将LDA模型应用到前文中Default数据集中，根据样本的透支额度和学生身份预测其是否会违约。训练样本1,000组，训练错误率2.75%

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

对全局：准确率：$Accuracy = \frac{9644+81}{9644+252+23+81} = 97.25\%$

对default Yes类：

$$\begin{cases} \text{精确度：} Precision = \dfrac{81}{23+81} = 77.88\% \\ \text{敏感度：} Sensitivit = \dfrac{81}{252+81} = 24.32\% \end{cases}$$

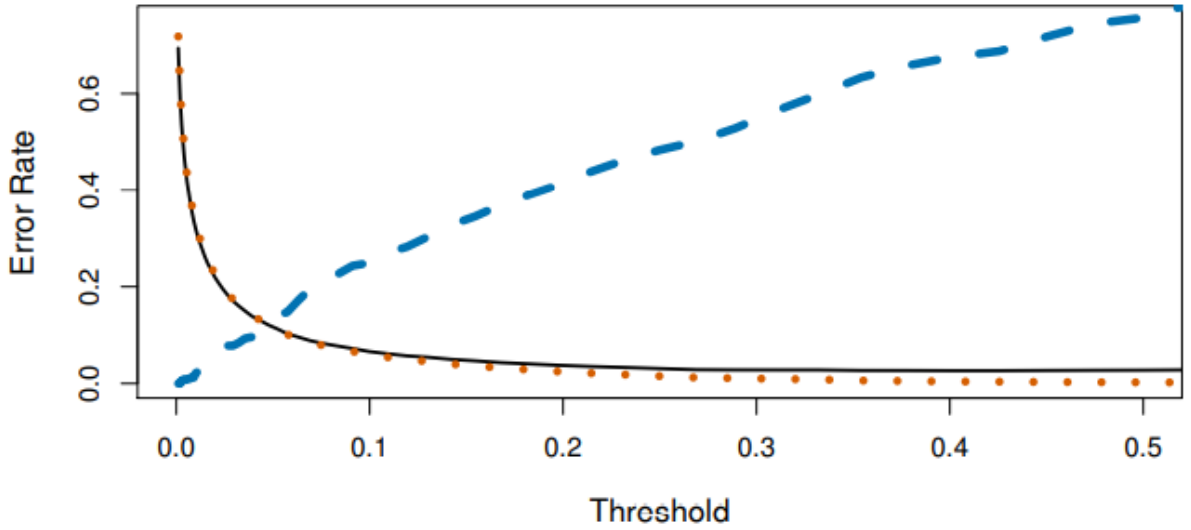
线性决策分析

为什么LDA分类效果差?

LDA以Bayes分类器为标准，降低总错误率。

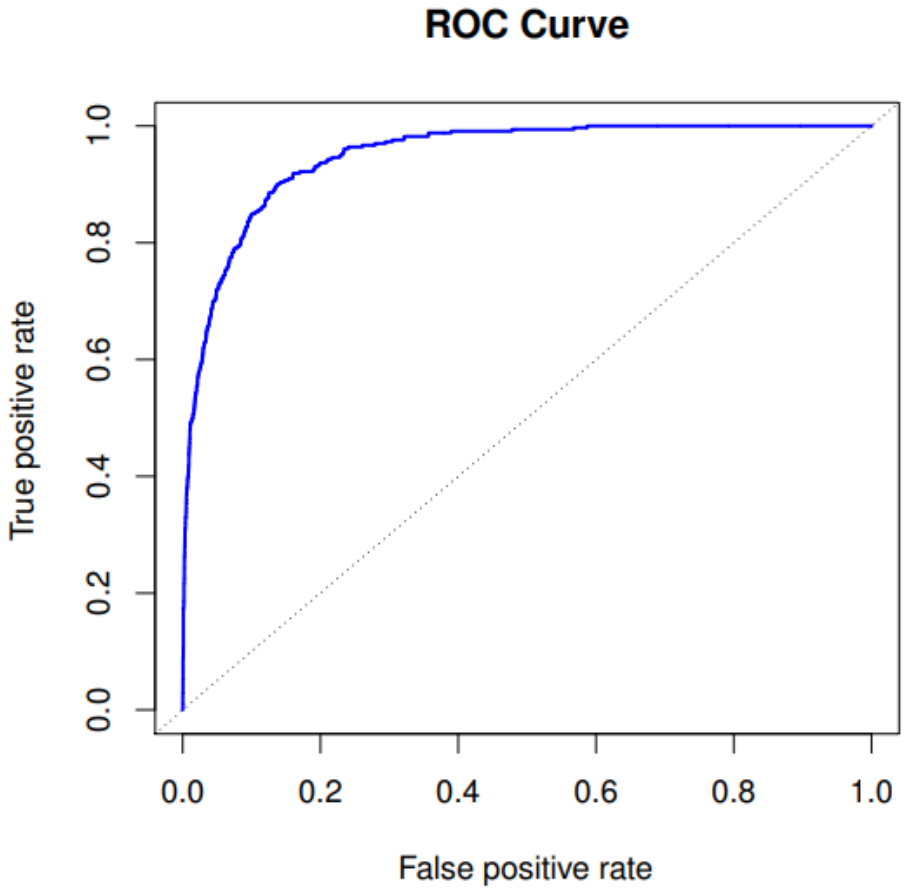
		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

黑线：总错误率
橙色点：假阳性
蓝色虚线：假阴性



ROC曲线

- ROC曲线(Receiver Operating Characteristic)可以同时表达出所有可能阈值出现的两类错误率。
- 右图为训练数据上LDA分类器的ROC曲线。
- ROC曲线下面的面积（Area Under the ROC Curve，AUC）可表示分类器质量。
- 一个理想的ROC曲线会紧贴左上角
- 该分类器AUC=0.95



线性决策分析

分类结果和分类器性能度量总结：

Confusion Matrix

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N^*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall (召回率)
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion (精确度)
Neg. Pred. value	TN/N^*	

二次决策分析（Quadratic Discriminant Analysis，QDA）

LDA：假设样本特征为高斯分布，每个类别的特征具有各自的均值，但共用协方差矩阵（每个类别的不同特征具有相同的关联）

QDA：每个类别有自己的协方差矩阵，其余与LDA相同

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k\end{aligned}$$

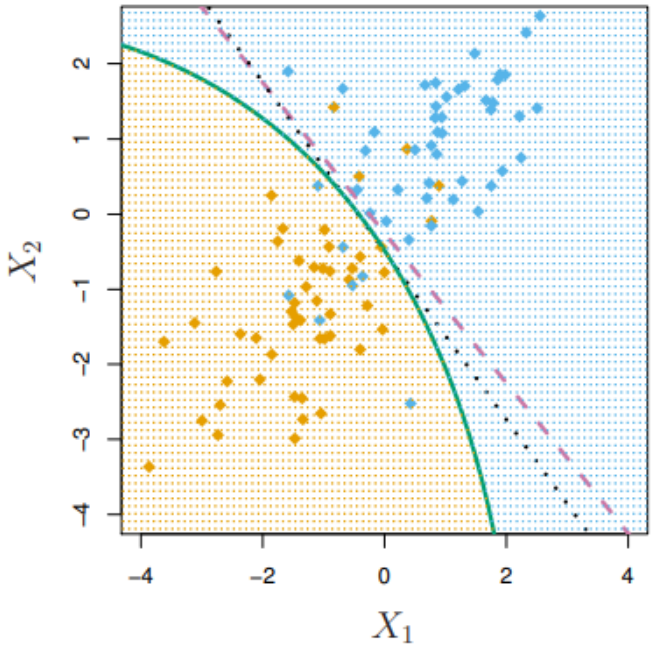
能使上式最大的类别即为QDA的预测结果。其中 δ_k 是关于 x 的二次函数。

参数数量：LDA：Kp

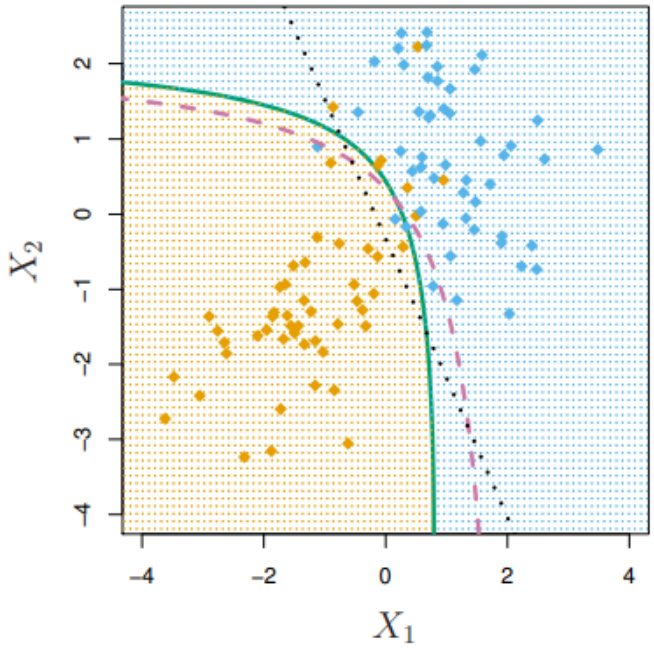
QDA：$Kp(p+1)/2$

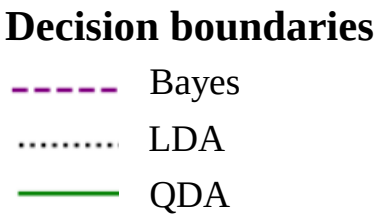
二次决策分析

$\Sigma_1 = \Sigma_2$



$\Sigma_1 \neq \Sigma_2$





LDA参数少，灵活性低，方差（variance）相对较小，偏差（bias）可能较大

QDA更灵活，数据量小时会导致较大方差，同时偏差较小

朴素贝叶斯（Naive Bayes）

假设第k类中，p个预测变量均独立，而不是共同遵循同一特定的分布形式。概率密度函数可以表示为：

$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$$

f_{kj}为第k类中第j个预测变量的概率密度函数。

- 边缘分布：每个变量本身的分布。完全消除变量之间的关联。
- 联合分布：考虑不同预测变量之间的关联的分布，在多元正态(高斯)分布情况下由协方差矩阵的非对角元描述。

代入贝叶斯定理，得到后验概率的表达式为：

$$\Pr(Y = k|X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)}$$

朴素贝叶斯（Naive Bayes）

求一维的f_{kj}：

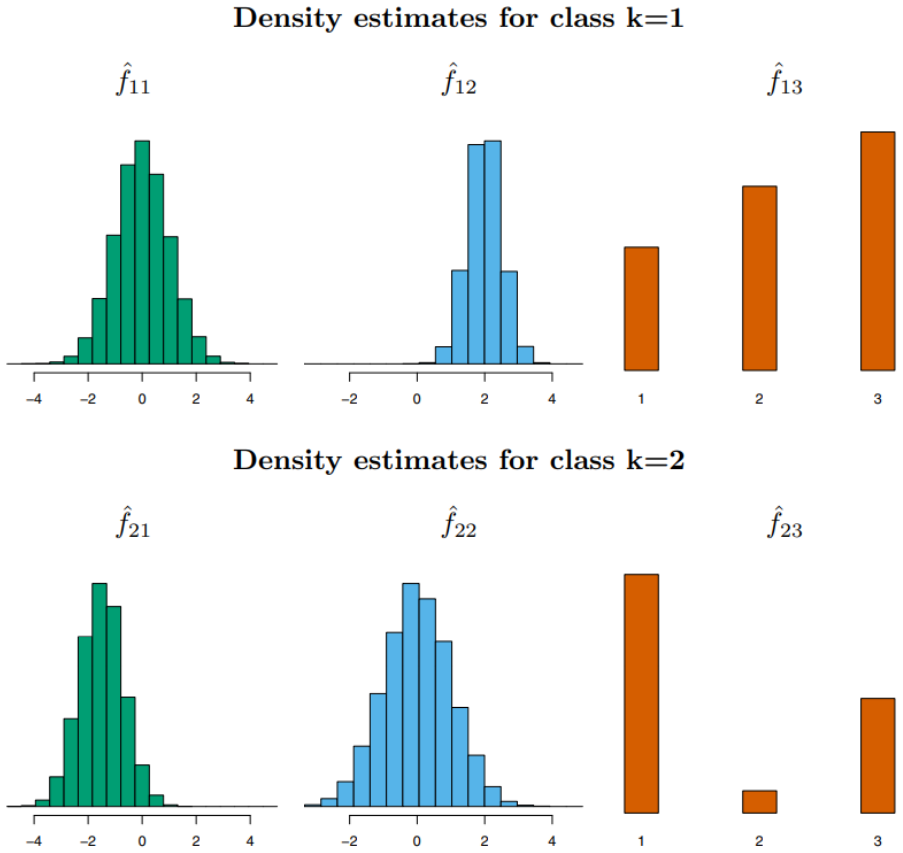
- 当X_j是连续值时，可以假设每个类中的第j个变量来自于正态分布。（与QDA不同的是，各变量之间是独立的，也相当于将QDA协方差矩阵限制为对角矩阵）
- 当X_j是连续值时，也可以进行非参数的估计。例如为每个类的第j个变量绘制直方图(histogram)/进行核密度估计(kernel density estimator)，得到估计的$f_{kj}(x_j)$
- 当 X_j 是离散值时，可以计算每个类中第 j 个变量中各个值的比例。例如，假设$X_j \in \{1,2,3\}$，且在第k类中有100个样本，对应取值为1、2、3的样本数分别为32、55、13，那么：

$$\hat{f}_{kj}(x_j) = \begin{cases} 0.32 & if\ x_j = 1 \\ 0.55 & if\ x_j = 2 \\ 0.13 & if\ x_j = 3 \end{cases}$$

朴素贝叶斯（Naive Bayes）

Exercise:

二分类任务(K=2)，共有三个特征(p=3)，其中前两个为连续型，后一个为离散型特征。假设$\hat{\pi}_1 = \hat{\pi}_2 = 0.5$。使用朴素贝叶斯模型判断$x^* = (0.4, 1.5, 1)^T$属于各类的概率？



解：$\hat{f}_{11}(0.4) = 0.368$，$\hat{f}_{12}(1.5) = 0.484$

$\hat{f}_{13}(1) = 0.226$ ，$\hat{f}_{21}(0.4) = 0.030$

$\hat{f}_{22}(1.5) = 0.130$，$\hat{f}_{23}(1) = 0.616$

$\Pr(Y = 1|X = x^*) = 0.994$

$\Pr(Y = 2|X = x^*) = 0.056$

几个分类模型的比较：解析形式

LDA、QDA、朴素贝叶斯均是将x分到$\Pr(Y=k|X=x)$最大的一类，也等价于下式最大的一类（将第K类作为基线）：

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)$$

- LDA：

$$\begin{aligned}
\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) &= \log\left(\frac{\pi_k f_k(x)}{\pi_K f_K(x)}\right) \\
&= \log\left(\frac{\pi_k \exp\left(-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)\right)}{\pi_K \exp\left(-\frac{1}{2}(x-\mu_K)^T\mathbf{\Sigma}^{-1}(x-\mu_K)\right)}\right) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k) \\
&\quad +\frac{1}{2}(x-\mu_K)^T\mathbf{\Sigma}^{-1}(x-\mu_K) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) \\
&\quad + x^T\mathbf{\Sigma}^{-1}(\mu_k-\mu_K) \\
&= a_k + \sum_{j=1}^{p} b_{kj}x_j
\end{aligned}$$

其中，$a_k = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K)$，$b_{kj}$是$\Sigma^{-1}(\mu_k-\mu_K)$的第$j$个分量。

几个分类模型的比较：解析形式

- QDA：

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)=a_k+\sum_{j=1}^{p}b_{kj}x_j+\sum_{j=1}^{p}\sum_{l=1}^{p}c_{kjl}x_jx_l$$

其中，a_k、b_{kj}、c_{kjl}是π_k、π_K、μ_k、μ_K、Σ_k、Σ_K的函数

- naive Bayes：

$$\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)=a_k+\sum_{j=1}^{p}g_{kj}(x_j)$$

其中，$a_k=\log\left(\frac{\pi_k}{\pi_K}\right)$，$g_{kj}=\log(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)})$。

广义可加模型
（generalized additive model）

几个分类模型的比较：解析形式

- LDA 是QDA的特例，当$c_{kjl}=0$时，QDA退化为LDA；OR：将QDA中不同类别的协方差矩阵限制为相等，则得到LDA
- 任意具有线性决策边界的分类模型都是朴素贝叶斯分类器的特例（$g_{kj}(x_j)=b_{kj}x_j$）➔LDA是朴素贝叶斯的特例
- 在朴素贝叶斯模型中，若假设特征遵从一维正态分布，则该分类器等同于协方差为对角阵的LDA
- QDA和朴素贝叶斯互不统属。朴素贝叶斯是可加模型，QDA可包含二次项（特征之间有协同时适用）
- 逻辑回归是一个线性形式的LDA：当特征遵从正态分布时，LDA通常比逻辑回归精度更高

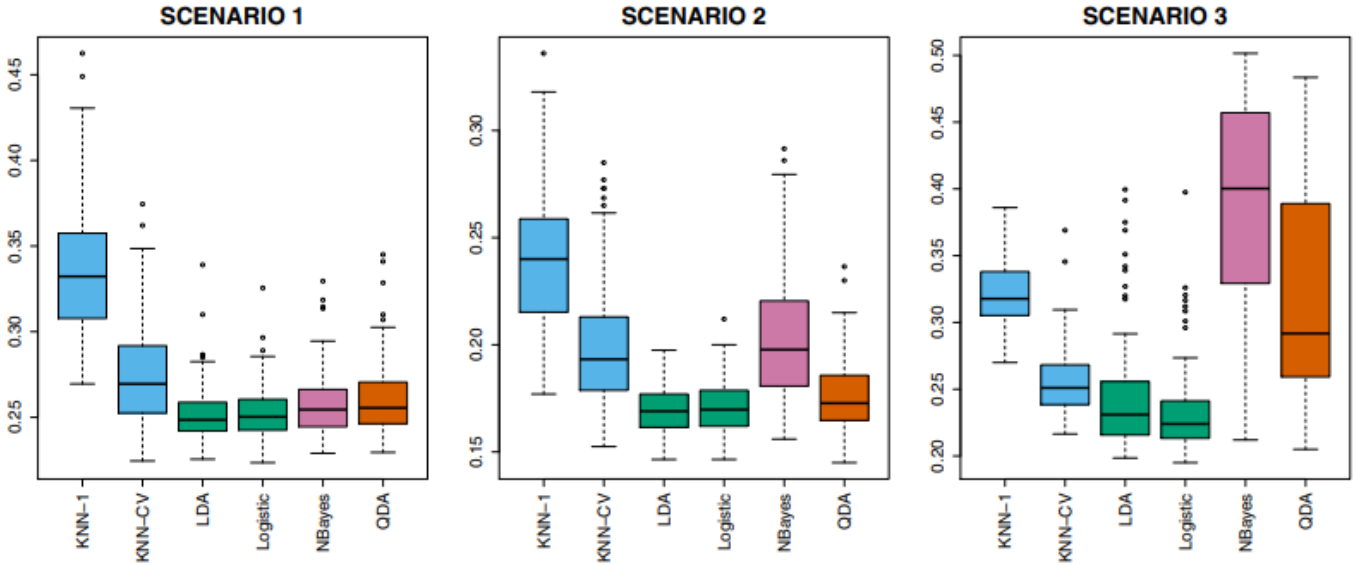
不同分类方法的比较

- KNN 与前述分类器（逻辑回归，LDA、QDA、朴素贝叶斯）的原理完全不同。
- KNN是个纯粹的非参数方法，对决策边界的形状没有做出任何假设：只看样本周围的情况而做判定
- 当决策边界具有高度非线性，且样本数足够，特征数较小时，KNN往往优于 LDA 和逻辑回归。
- 具有非线性决策边界时，若样本数不大，或特征数不小，QDA可能优于KNN
- KNN 无法判断哪些特征重要，不利于推断任务

分类方法的比较：实例

- 6个场景：1-3，贝叶斯决策边界为线性；4-6，非线性
- 均为二分类问题，每个样本有两个特征（p=2）
- 每个场景有100个随机样本用于训练模型，每个训练集对应一个更大的测试集，用于测试模型性能
- KNN：k=1，或利用交叉验证（cross validation）确定最优k值（k=cv）
- 朴素贝叶斯模型：每个类别的特征具有同样方差

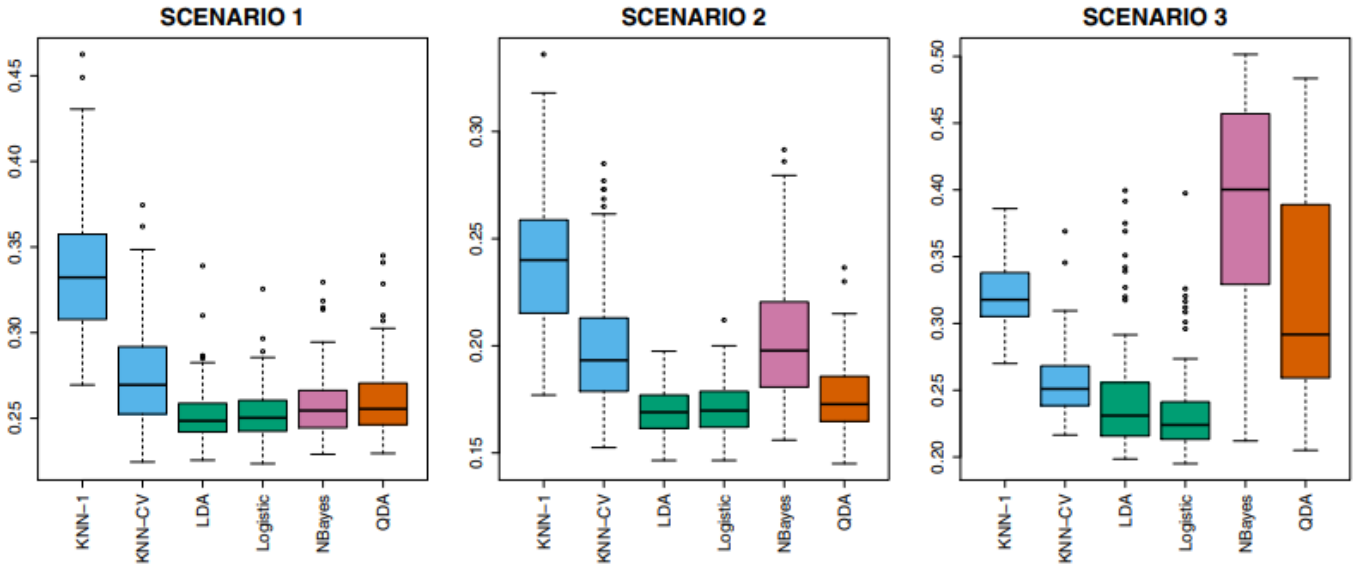
分类方法的比较：实例



Scenario 1：两类各20个样本，每一类样本之间都不相关，呈正态随机分布，且均值不同

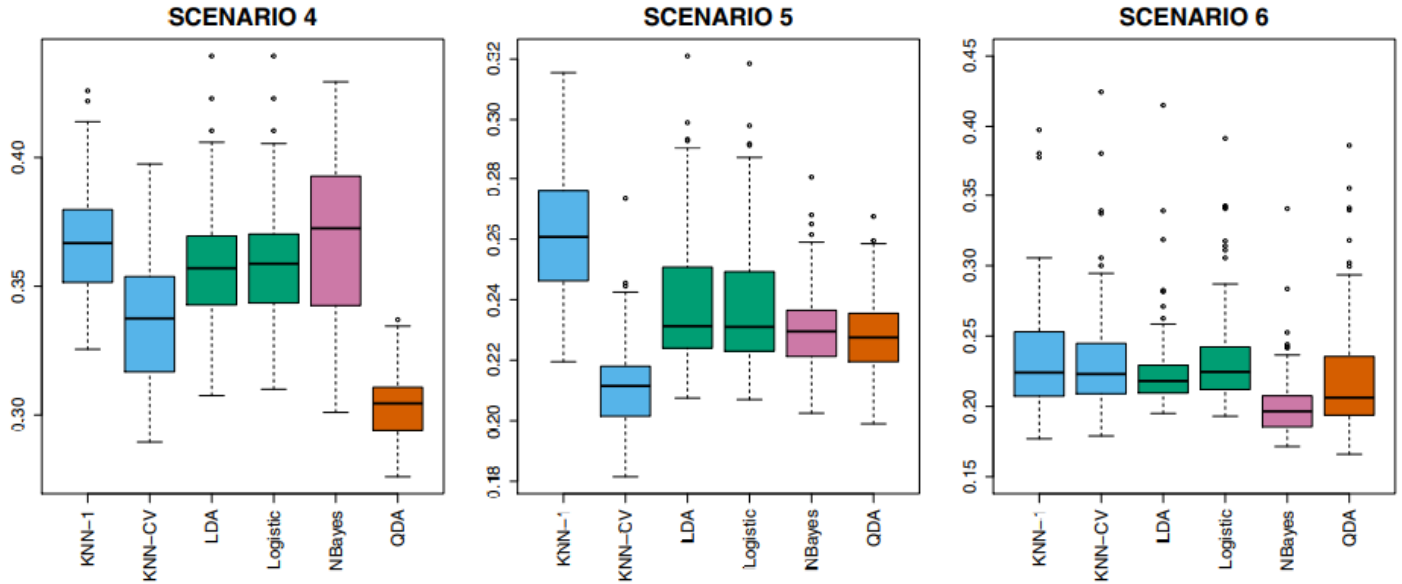
Scenario 2：在Scenario 1 的基础上，每类中的两个变量的相关性为-0.5

分类方法的比较



Scenario 3：从t分布中产生X_1和X_2，且特征间加入负相关；每类50组样本（t分布类似于正态分布很相似，但会产生更多远离均值的点）

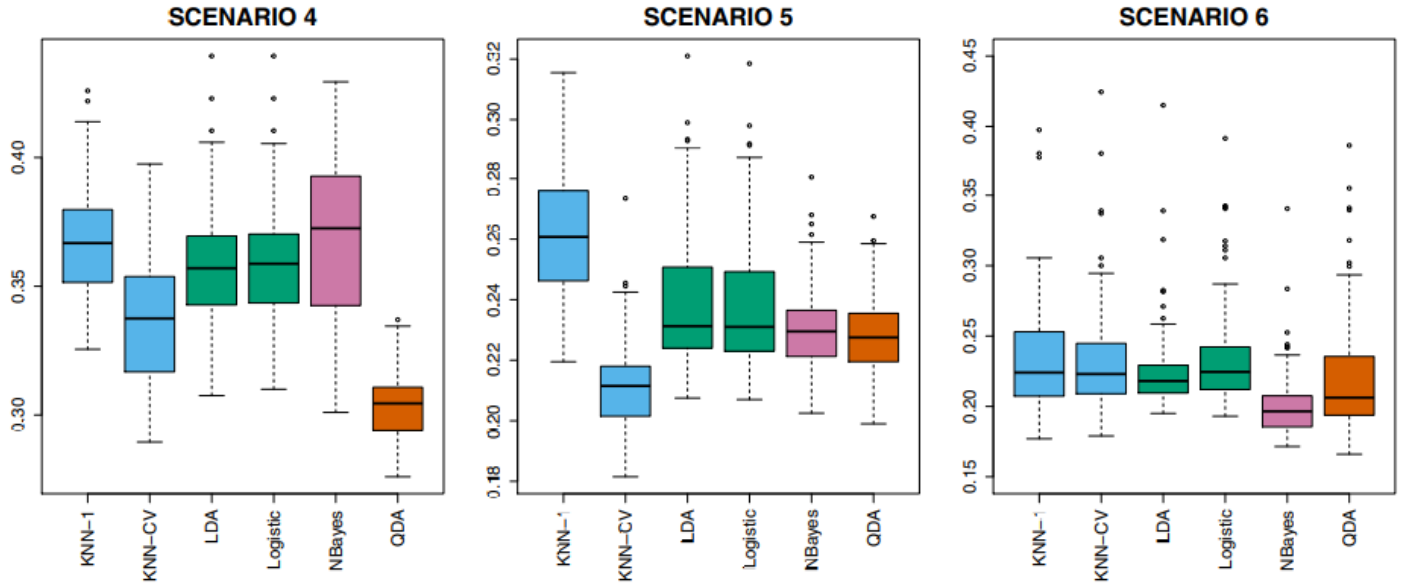
分类方法的比较



Scenario 4：数据都由正态分布产生，第一类中特征变量相关系数为0.5，第二类为-0.5。

Scenario 5：不相关的正态分布特征，响应依据logistic函数产生

分类方法的比较



Scenario 6：观测值由正态分布生成，每个类协方差矩阵具有不同对角元，样本量较小n=6。

分类方法的比较

没有任何一种方法可以在各种情况下都优于其他方法

- 当具有线性真实决策边界时，LDA 和逻辑回归方法较好
- 当边界是一般非线性的时，QDA 会给出较好的结果
- 对更复杂的决策边界，非参数方法如 KNN 可能会更胜一筹，但是应该谨慎选择平滑参数