数据建模回顾

特征输入 → 拟合算法 → 相应输出

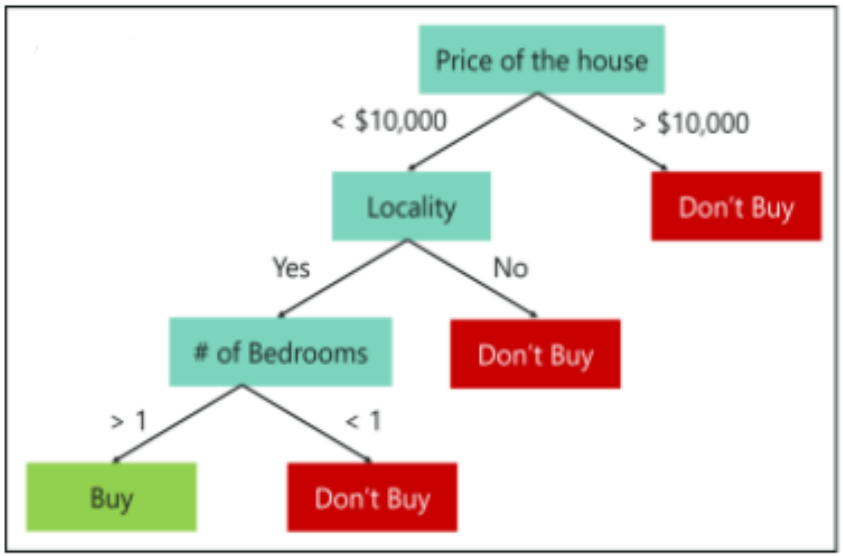
线性回归

$$Y=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$$

逻辑回归

$$\log(P/(1-P))=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$$

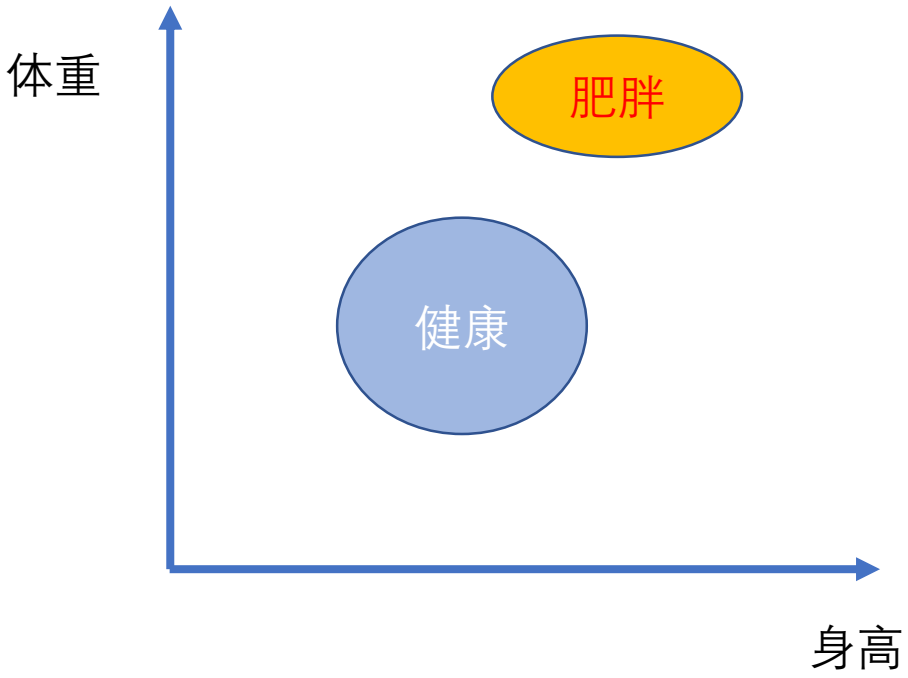
随机森林

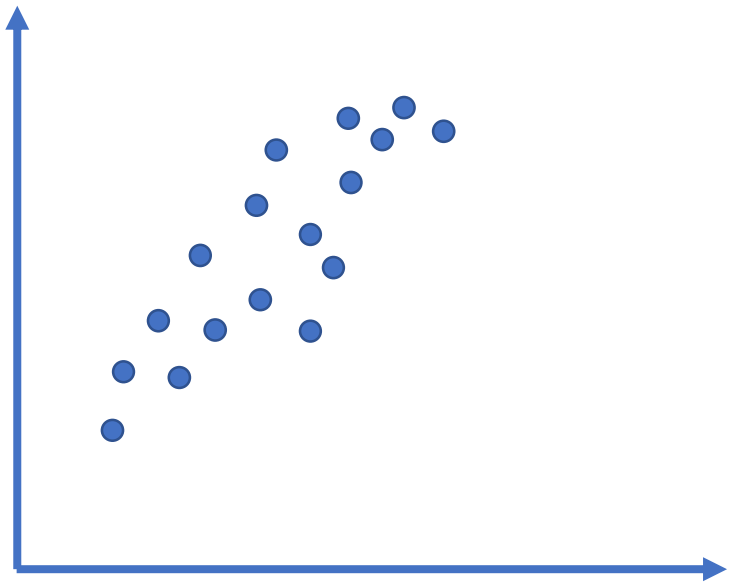


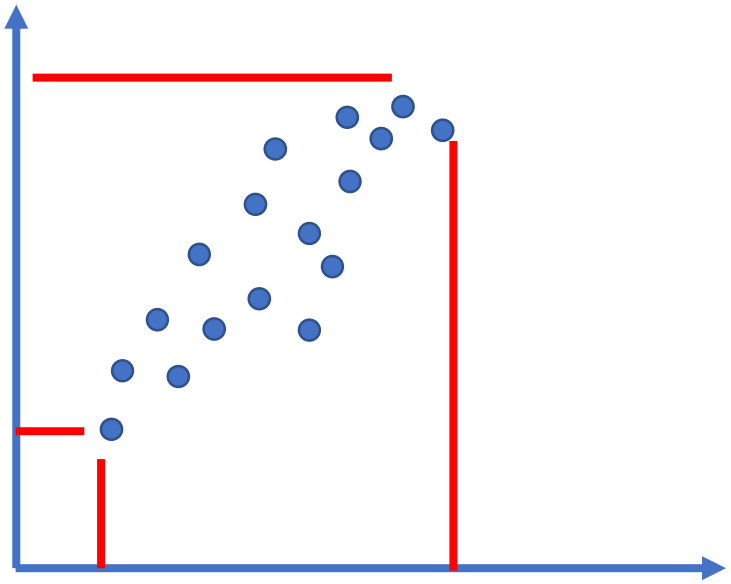
数据建模回顾

算法	线性回归	逻辑回归	随机森林
功能	回归	分类	分类&回归
建模（训练）	lm()	glm(), multinom()	randomForest()
特征选择	step() #AIC指标	step() #AIC指标	rf$importance #基尼系数
预测	predict.lm()	predict()	predict()

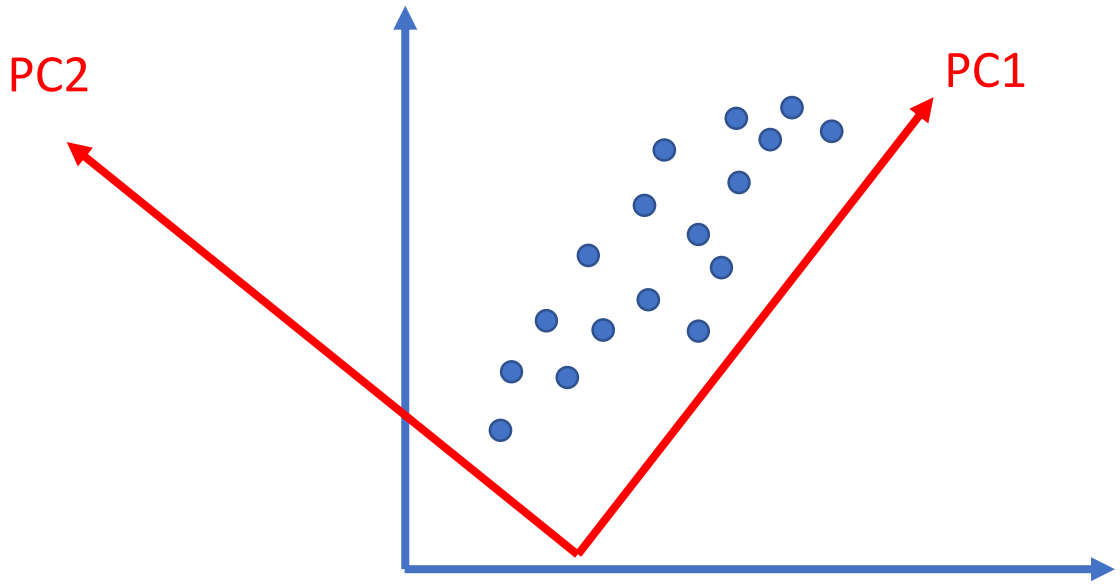
数据可视化


体重
肥胖
健康
身高





主成分分析（Principal Component Analysis, PCA）



主成分分析（Principal Component Analysis, PCA）

PCA, PCoA, PLS-DA, OPLS-DA

PCA

PCoA

PLS-DA

OPLS-DA

PCA

参考：https://zhuanlan.zhihu.com/p/497474588

数据集：鸢尾花数据集

萼片长度　萼片宽度　花瓣长度　花瓣宽度

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

```
> com1$rotation  #载荷矩阵（旋转矩阵）
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

PC1=0.5210659×Sepal.Length-
0.2693474×Sepal.Width+
0.5804131×Petal.Length+
0.5648565×Petal.Width

OPLS-DA

参考：R实战 | OPLS-DA（正交偏最小二乘判别分析)筛选差异变量(VIP)及其可视化 (qq.com)

数据集：液相色谱高分辨质谱法分析的来自183位成人的尿液样品

一些概念：
预测成分：OPLS-DA只有1个，PLS-DA可以有多个
正交成分：OPLS-DA可以有多个，PLS-DA没有这个概念
R2X: 模型对类间差异的揭示程度，越接近1越好
R2Y: 模型对类内差异的解释程度，越接近1越好
Q2Y: 模型的预测能力，越接近1越好
过拟合：模型捕捉了一些“有害”的特征

因子分析

参考：《R语言医学数据分析》第12章

数据集：
样本：220个学生
特征：6个科目成绩，分别是"盖尔语","英语","历史","算术","代数","几何“

	ML1	ML2	h2	u2	com
盖尔语	0.55	0.43	0.49	0.51	1.9
英语	0.57	0.29	0.41	0.59	1.5
历史	0.39	0.45	0.36	0.64	2.0
算术	0.74	-0.27	0.62	0.38	1.3
代数	0.72	-0.21	0.57	0.43	1.2
几何	0.60	-0.13	0.37	0.63	1.1

因子1=0.55×盖尔语+0.57×英语+0.39×历史+0.74×算数+0.72×代数+0.60×几何

因子2=0.43×盖尔语+0.29×英语+0.45×历史-0.27×算数-0.21×代数-0.13×几何

生存分析与Cox回归

参考：《R语言医学数据分析》第9章

场景：研究两种药物对延长寿命的差异。

线性回归 $Y=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$

逻辑回归 $\log(P/(1-P))=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n$

Cox回归 死亡风险$=e^{(\beta_1X_1+\beta_2X_2+\ldots+\beta nXn)}$

生存分析与Cox回归

参考：《R语言医学数据分析》第9章

数据集:

随访时间　患者状态　年龄　疾病残留　治疗方法　患者ECOG评分

	futime	fustat	age	resid.ds	rx	ecog.ps
1	59	1	72.3315	2	1	1
2	115	1	74.4932	2	1	1
3	156	1	66.4658	2	1	2
4	421	0	53.3644	2	2	1
5	431	1	50.3397	2	1	1
6	448	0	56.4301	1	1	2
7	464	1	56.9370	2	2	2
8	475	1	59.8548	2	2	2
9	477	0	64.1753	2	1	1

生存分析与Cox回归

参考：《R语言医学数据分析》第9章