统计机器学习方法

李毅

山西财经大学统计学院

本课学习提示及要求

了解：

- 机器学习与统计学的区别与联系。

理解：

- 机器学习的基本步骤。
- 算法的类型及选择方法。

掌握：

- 面向机器学习的数据划分及准备方法。
- 机器学习中对模型的评估方法。

前期基础

- 已修过数理统计与概率、线性代数、统计学或计量经济学。
- 具备R语言或Python的基础知识。
- 适用于大二、大三经济管理相关专业同学。

课程安排

问卷设计和调查方案

11月11日	中国传媒大学	王小宁	市场调查研究选题
11月11日	中国传媒大学	张　喆	调查问卷的设计
11月18日	中国人民大学	蒋　妍	抽样调查的误差分析和质量控制
11月18日	首都医科大学附属北京地坛医院	郝一炜	问卷调查中的抽样设计和样本代表性

数据收集

11月25日	益　普　索	老大卫	市场研究企业经验分享
11月25日	中国人民大学	胡以松	线下调查实战--以CGSS为例

调查数据分析与建模

12月02日	中国人民大学	唐丽娜	调查资料的整理和清洗
12月02日	中国人民大学	吴翌琳	描述性统计分析与可视化
12月09日	腾　　讯	樊中一	从答题者角度出发，提升问卷数据质量
12月09日	青岛大学	李莉莉	概率调查样本的统计推断

数据

特征（自变量，X）

标记（因变量，Y）

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

训练集

1	青绿	蜷缩	沉闷	?

测试集

提纲

一、市调全局视觉（宏观看）：

① 参数模型与算法模型————研究框架；

② 无监督学习与有监督学习————调查目的：探索？还是证明观点？

③ 分类和回归——看数据结构，即,Y是离散还是连续？

二、建模视觉（微观看）：

① 机器学习模型的性能度量；

② 机器学习模型的评估方法；

市场调研

市场调研是人类获取环境信息、理解生存条件、判断未来趋势，甚至于满足好奇心的一种天性。

（周庭锐，2012）

Hindsight，Insight，Foresight

宝洁中国的市场研究部（Consumer and Market Knowledge，CMK）

CRISP-DM 模型

CRoss-Industry Standard Process for Data Mining

CRISP-DM，即跨行业数据挖掘标准流程，此KDD过程模型于1999年欧盟机构联合起草。

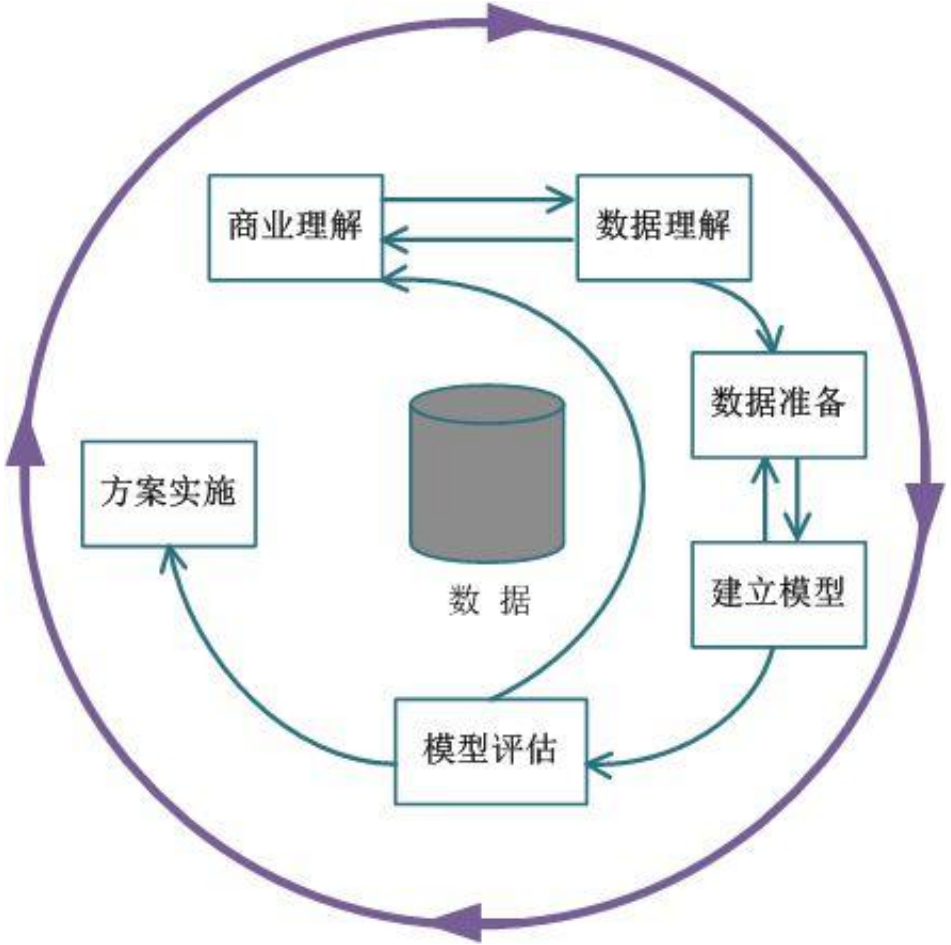


图 1 展示了 CRISP-DM 中定义的数据挖掘生命周期中的六个阶段。

- **商业理解**：了解进行数据挖掘的**业务原因**和数据挖掘的**目标**；
- **数据理解**：深入了解可用于挖掘的数据
- **数据准备**：对待挖掘数据进行合并，汇总，排序，样本选取等操作
- **建立模型**：根据前期准备的数据选取合适的模型
- **模型评估**：使用在商业理解阶段设立的业务成功标准对模型进行评估
- **结果部署**：使用挖掘后的结果提升业务的过程

模型是如何来的?

- 首先，根据数据状况(探索性数据分析)主观确定出一个参数模型类型或者一个算法；

- 其中典型参数模型例子为线性回归模型等，而算法模型例子为决策树，随机森林等

- 其次，根据已知数据训练/学习出出参数模型的参数或者算法模型的程序

模型是根据数据训练出来的!

参数模型

- 经典统计---假定背景分布下的假设检验及点估计、线性回归、多元分析等等；模型形式是根据经验及数学的可算性假定的可以写出公式的模型，模型参数由数据来估计
- 经典统计判断模型好坏的方法：拟合优度检验、t检验、F检验，p值、残差分析等等。只有一个训练集（自己对自己投票）没有测试集
- 误差基本上应该是数据和模型的差距+模型本身解的精确性，但只能核对后者

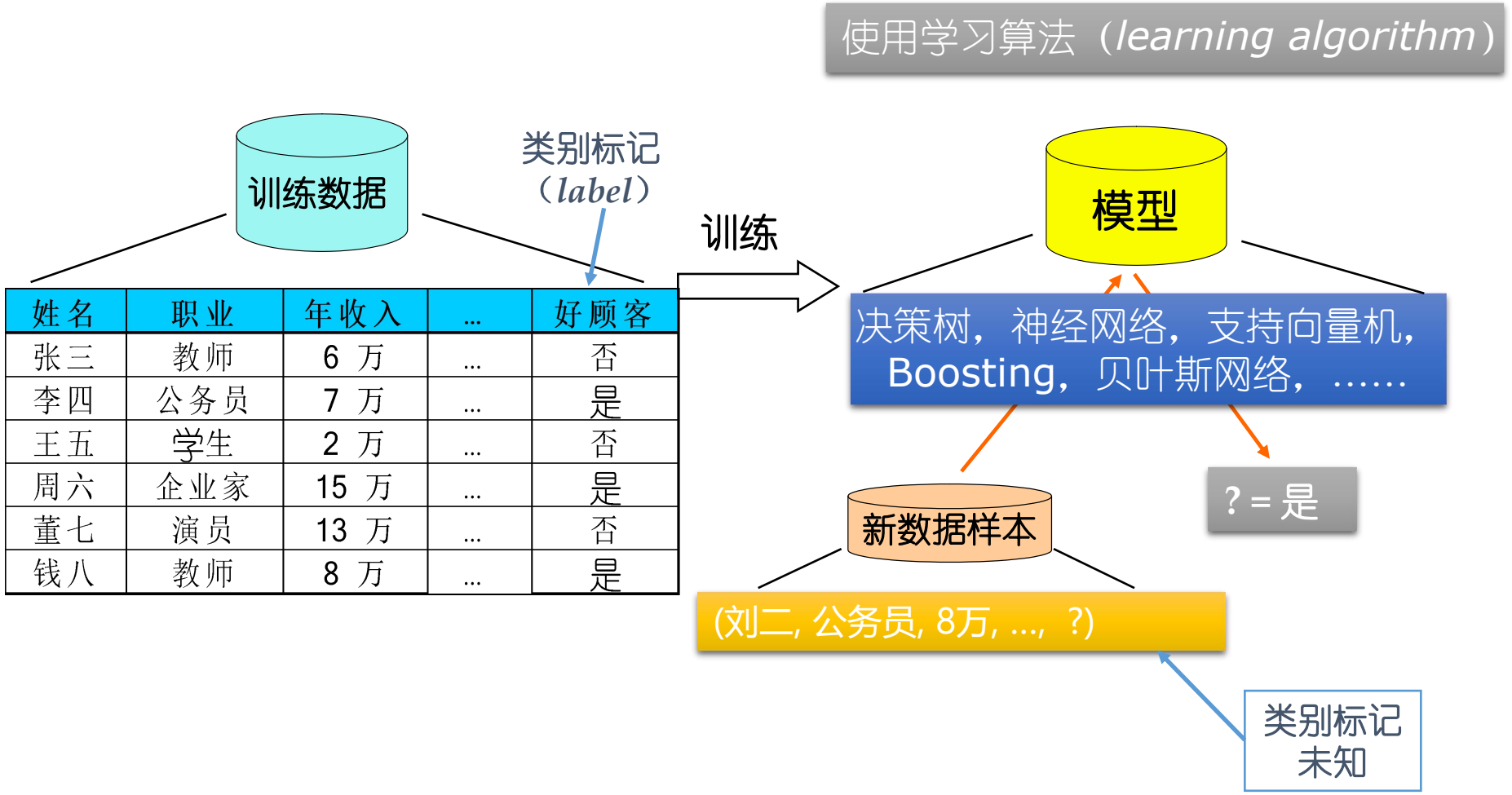
参数模型是用写得出来的公式描述的数学模型，有了形式之后才训练以得到参数估计

算法模型

- 算法建模（或称为机器学习），特点：根据数据训练一个算法模型（模型是一个计算机程序，而不是一个数学公式）
- 判断模型的好坏：用一部分不参加建模的数据(测试集)检验另一部分数据(训练集)建立的模型，即客观公正的交叉验证
- 泛化误差

算法模型是计算机程序,以不同的方法来区分,由数据来训练

典型的机器学习过程


使用学习算法（learning algorithm）
训练数据
类别标记
（label）
训练
模型
姓名 职业 年收入 ... 好顾客
张三 教师 6 万 ... 否
李四 公务员 7 万 ... 是
王五 学生 2 万 ... 否
周六 企业家 15 万 ... 是
董七 演员 13 万 ... 否
钱八 教师 8 万 ... 是
决策树，神经网络，支持向量机，
Boosting，贝叶斯网络，……
新数据样本
? = 是
(刘二, 公务员, 8万, ..., ?)
类别标记
未知

小结

- 参数模型：需要事先对处理对象做出假定（如正态分布等）
- 算法模型：不需要做事先假定。

统计学两个文化的差异是什么？

朱宏图 青牛帮 2021-12-13 01:54

——“Breiman访谈录｜《统计建模：两种文化》”的读后感（二）

非常荣幸，上个月被统计之都邀请写个读后感。我有感而发，仓促地完成短文“统计核心是什么？”。因为有大佬的背书，特别是统计之都、小罗同学、和张总的支持，虽然我人言轻微，但它造成的影响力远超过我的预计。这里要感谢所有帮我转贴的朋友们，现在总阅读量已经超过 12 万啦。我有幸发过许多顶刊论文，但是还没有一篇文章有这么大的影响力，说明有许多同仁有共鸣，都在思考统计的未来和我们需要干点啥。这文章的许多地方措辞上可以更准确一些，欢迎大家多提宝贵建议，希望在后面的短文中来不断改进。今天我重读了 Breiman 的访谈录，并且经历最近几件大事的冲击，我参悟了一些新东西，便有了以下新的读后感。

今天讲讲统计学的两个文化到底是什么？

首先，我在这里摘取 Breiman 教授的几句话：

在从数据到结论的过程中，有两种统计建模文化。第一种是数据模型，假设数据是通过给定的随机数据模型生成的。另一种是算法模型，将数据生成机制视为未知。一直以来，统计界几乎完全使用数据模型。这种情况造就了无关紧要的理论、有问题的结论，并使统计学家无法研究广阔、有趣的现实问题。算法建模，都在统计学之外的领域飞速发展。它既可以用于大型复杂的数据集，也可以用于小型数据集。

以下是我对这段话的一些浅显的认知：

这两个不同的文化的本质是什么？

这里面的数据模型到底是什么？其实没有一个明确的定义数据模型到底是什么？统计学里面有许多模型，包含线性模型、广义线性模型、生存分析模型、时间序列模型、潜变量模型、非参数模型、和半参数模型等等。所有这些模型都是为了某些特定的应用而被提出来，并随着在相关应用的广泛使用它们而得到进一步发展，包含相关的理论和可能应用的场景。可以说线性模型是所有这些模型的核心，许多线性模型的相关理论都被推广到其它模型，由此统计学里面许多理论结果都是从线性模型开始。某种程度上（我个人的理解），数据模型可能指的就是以线性模型为核心（或者跟线

提纲

一、市调全局（宏观看）：

① 参数模型与算法模型————研究框架；

② 无监督学习与有监督学习————调查目的：探索？还是证明观点？

③ 分类和回归——看数据结构，即,Y是离散还是连续？

二、建模（微观看）：

① 机器学习模型的性能度量；

② 机器学习模型的评估方法；

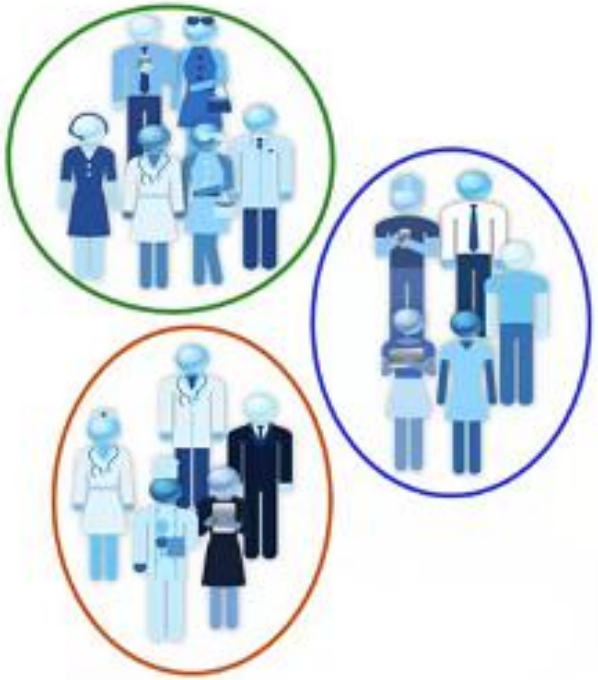
有监督学习（X，Y）—关联?

图 1.3 一般监督学习模型

图 2 使用监督学习模型进行预测

Customer Number	Age	Credit Score	Home Status	Vehicle Type	Outcome
1	52	420	Own	Sedan	Default
2	52	460	Own	Sedan	Default
3	64	480	Rent	Sports	Repaid
4	31	580	Rent	Sedan	Default
5	36	620	Own	Sports	Repaid
6	29	690	Rent	Pickup	Repaid
7	23	730	Rent	Sedan	Repaid
8	27	760	Rent	Pickup	Repaid
9	43	790	Own	Pickup	Repaid

无监督学习（X）—探索?



基于无监督学习在杂货店策略性地放置物品—关联规则

基于无监督学习的“物以类聚，人以群分”—聚类

提纲

一、市调全局视觉（宏观看）：

① 参数模型与算法模型————研究框架；

② 无监督学习与有监督学习————调查目的：探索？还是证明观点？

③ 相似性、分类和回归——看数据结构，即，Y是离散还是连续？

二、建模视觉（微观看）：

① 机器学习模型的性能度量；

② 机器学习模型的评估方法；

数据

特征（自变量，X）

标记（因变量，Y）

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

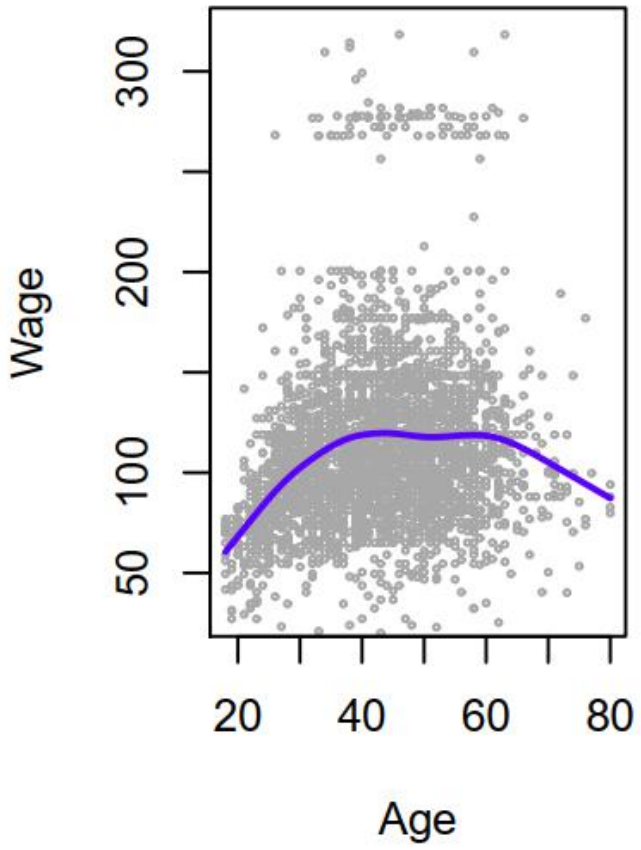
训练集

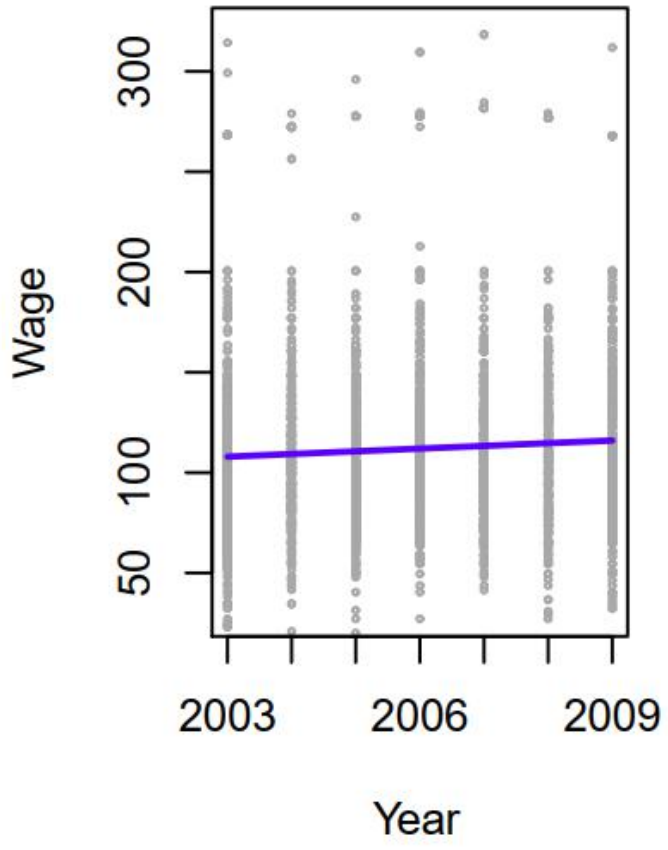
1	青绿	蜷缩	沉闷	?

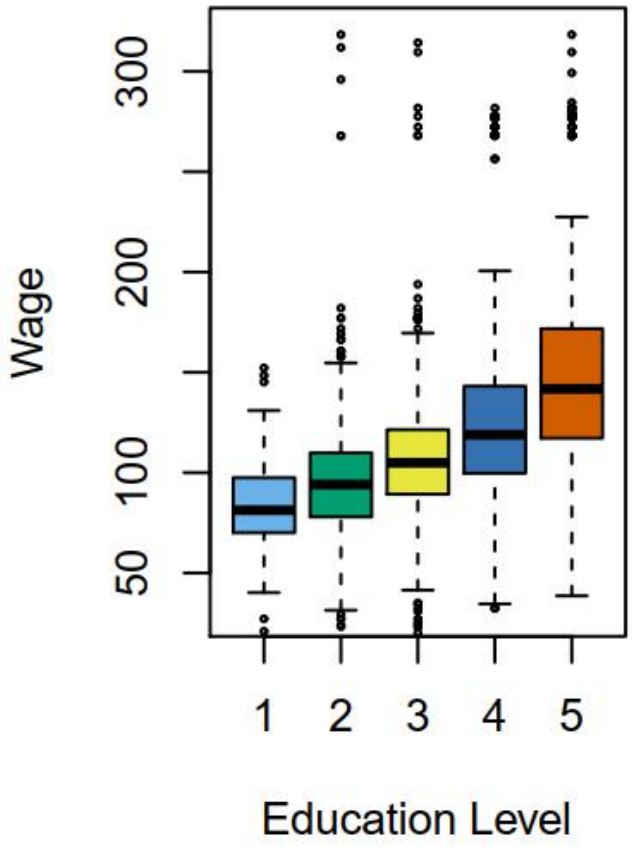
测试集

回归：（X，Y 且 Y是连续）

实例1： 工资数据：考察与某地区男性收入相关的几个因素。

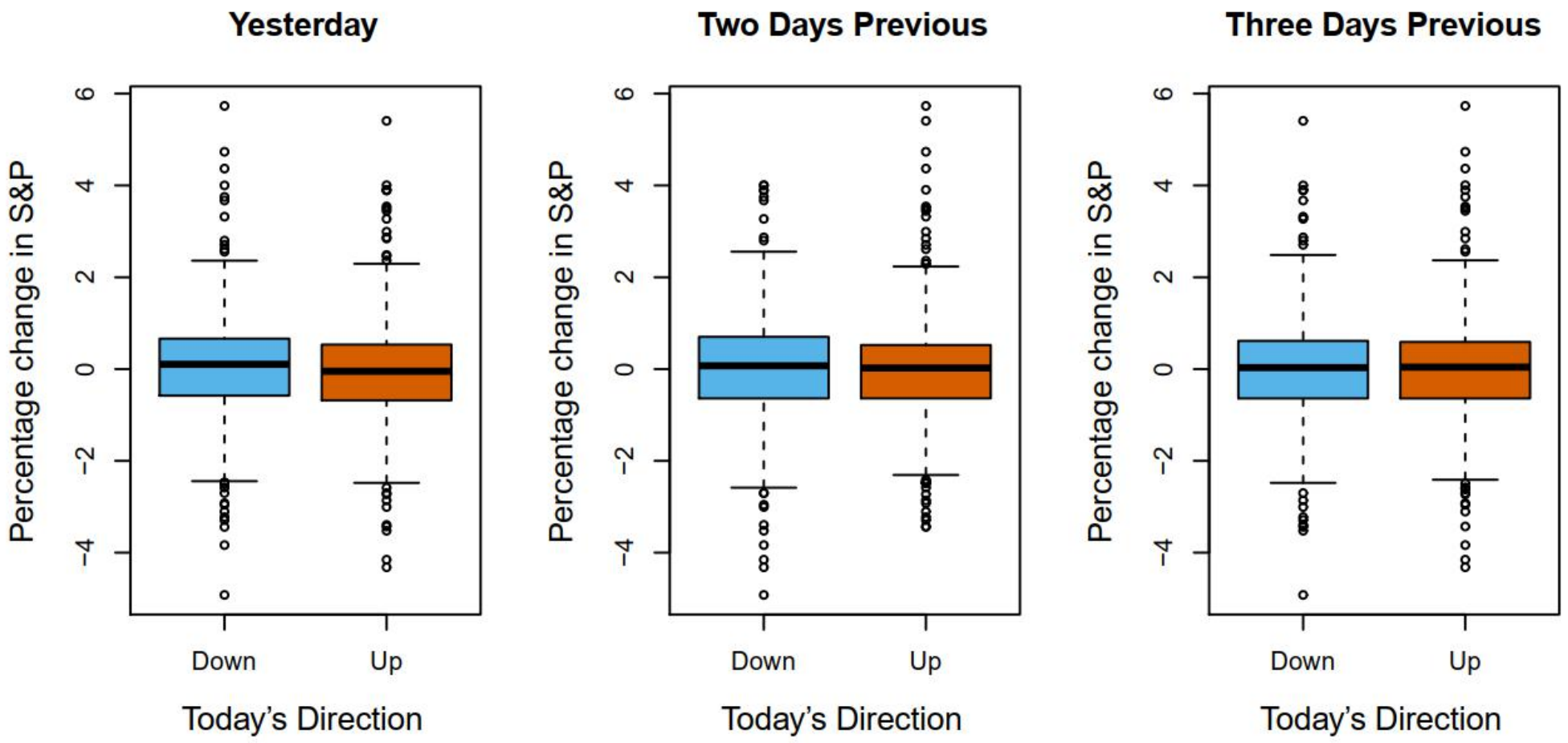






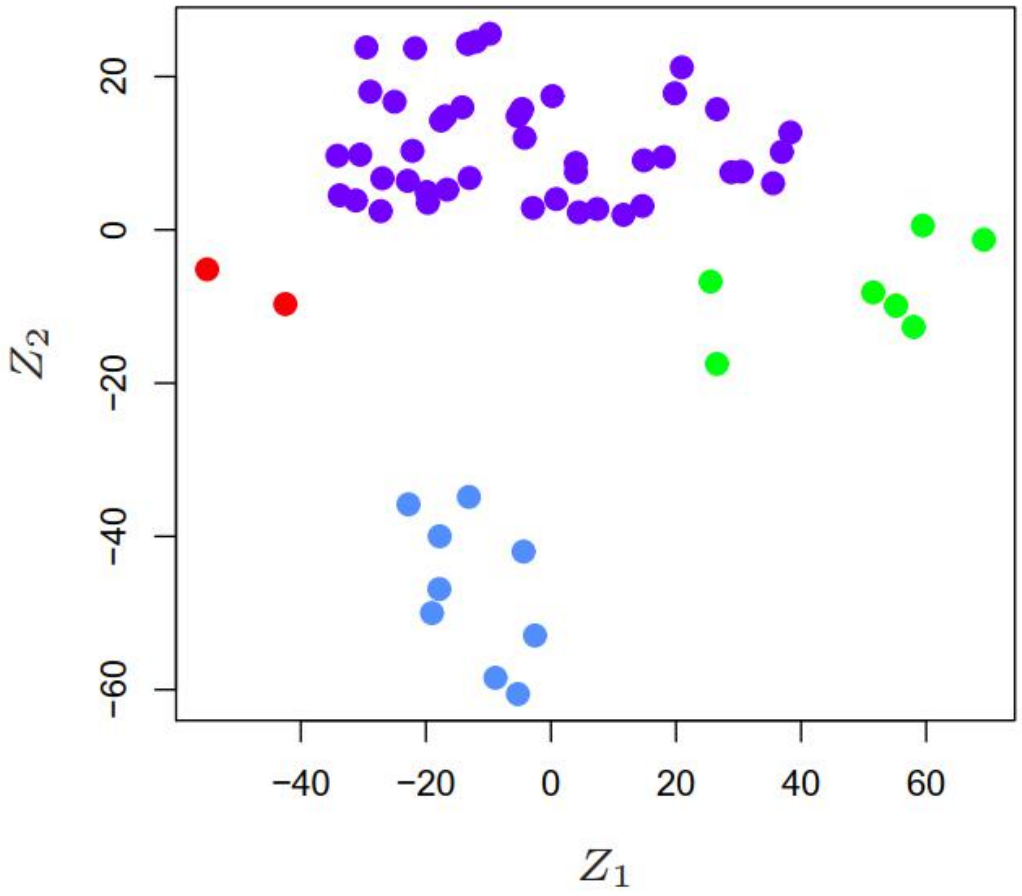
分类（X，Y 且 Y是离散的）

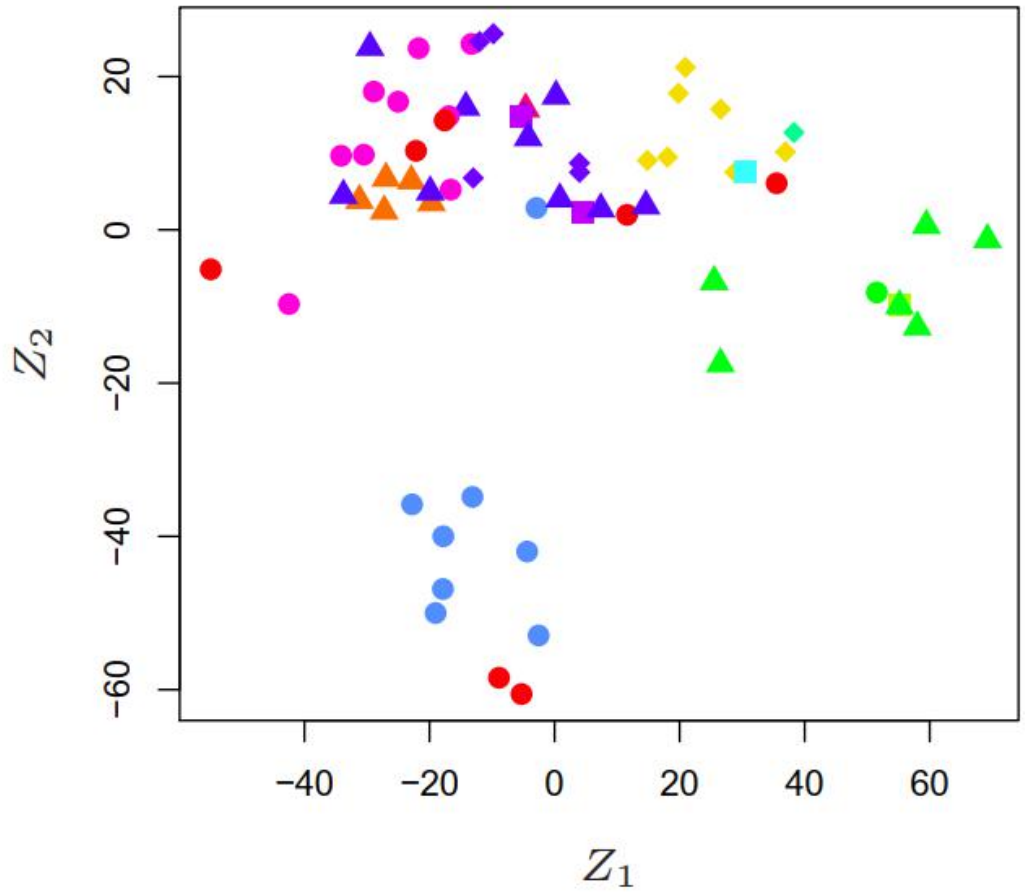
实例2： 金融市场数据：2001-2005年标准普尔500股票指数（S&P）



相似性（X，无Y）

实例3：基因表达数据：64个癌症细胞系6830个基因表达测量数据（NCI60）。





提纲

一、市调全局（宏观看）：

① 参数模型与算法模型————研究框架；

② 无监督学习与有监督学习————调查目的：探索？还是证明观点？

③ 相似性、分类和回归——看数据结构，即,Y是离散还是连续？

二、建模（微观看）：

① 机器学习模型的性能度量；

② 机器学习模型的评估方法；

数据

特征（自变量，X）

标记（因变量，Y）

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

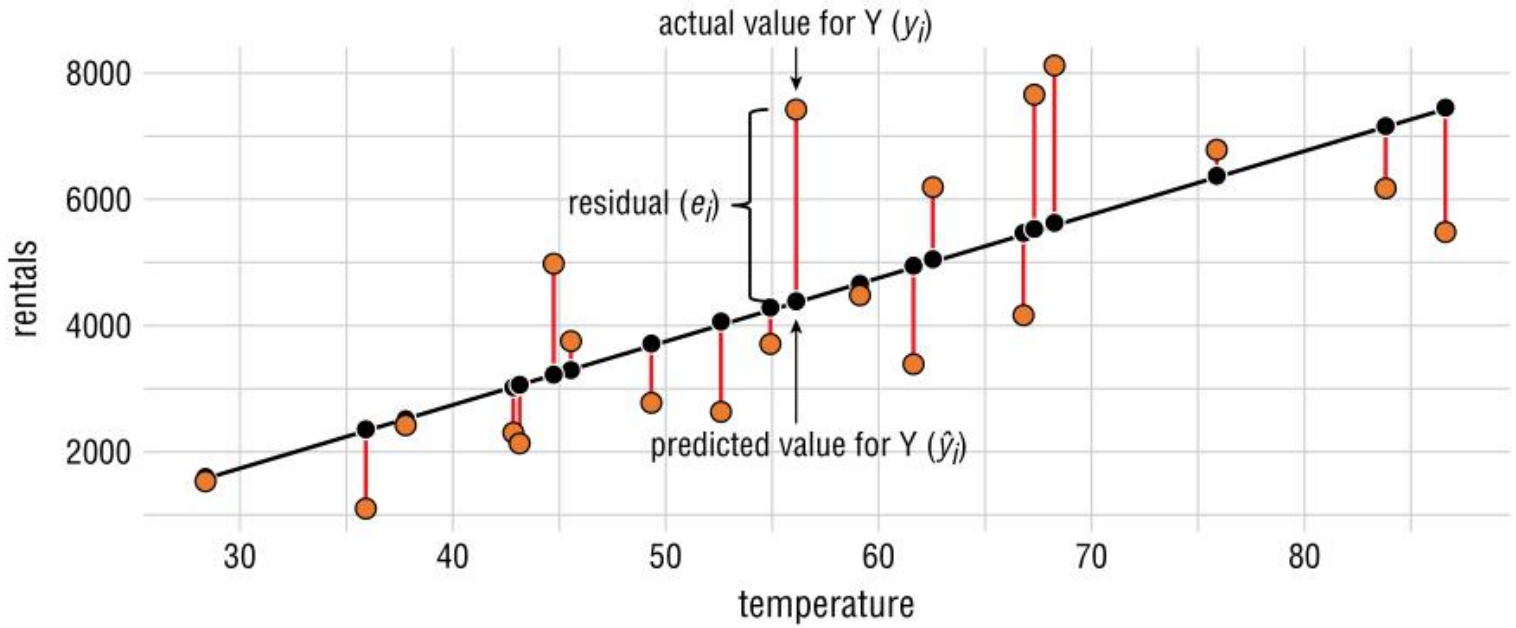
训练集

1	青绿	蜷缩	沉闷	?

测试集

看误差：$e = y - \hat{y}$

回归

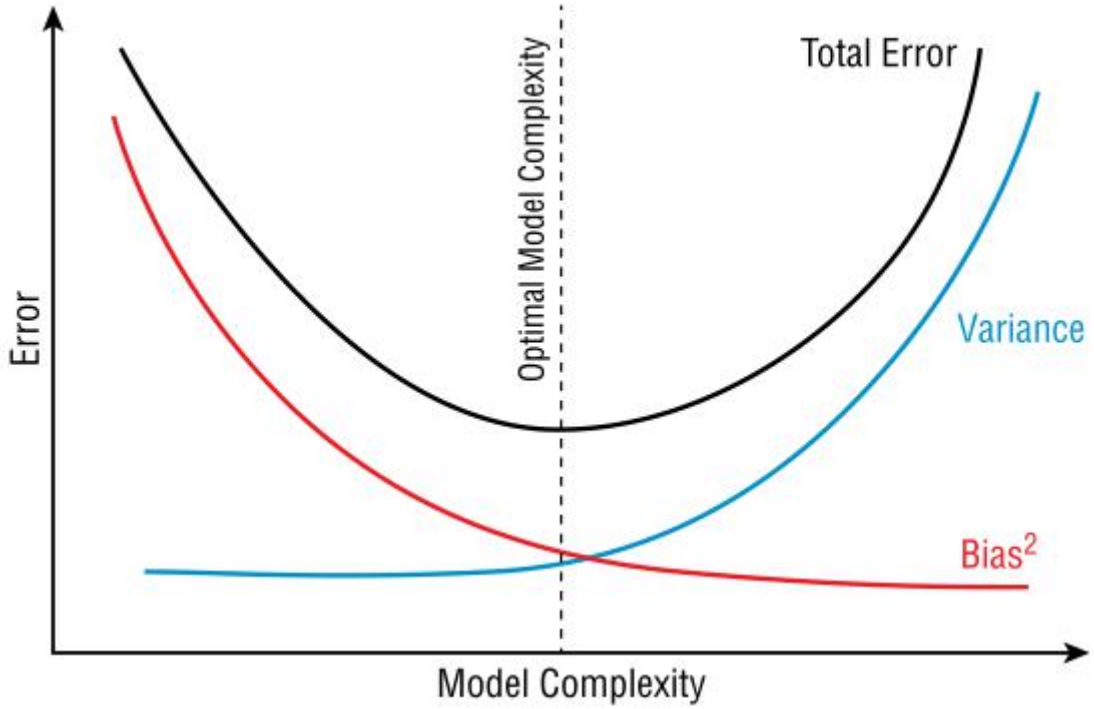


分类：

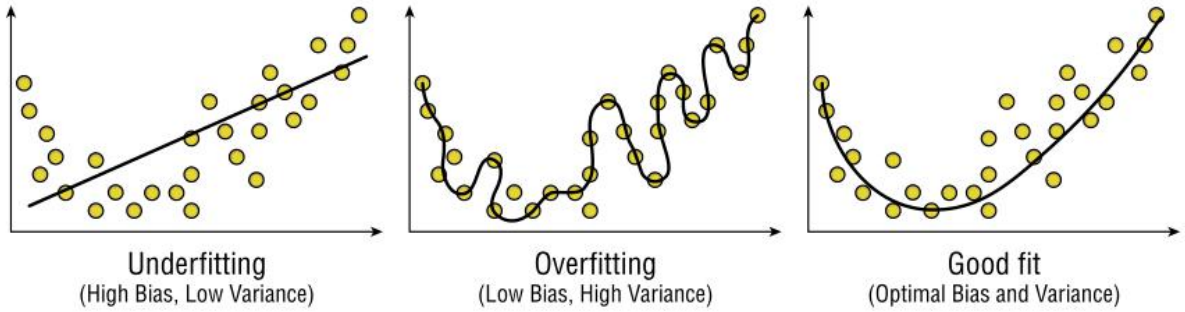
	真实正例	真实反例
预测正例	真正例	假正例 I 型错误
预测反例	假反例 II 性错误	真反例

建立机器学习模型时，该模型将包含某种类型的预测误差。这个错误有三种不同的形式。

- 偏差：模型的期望值与真实值的偏离程度，即刻画了模型本身的拟合能力。
- 方差：不同训练集导致模型性能的变化，即刻画数据扰动所造成的影响。
- 噪声：当前问题上任何模型所能达到的泛化误差的下界。



偏差/方差权衡



欠拟合、过拟合和最佳拟合

小结

$$\text{泛化误差}=\text{偏差}^2+\text{方差}+\text{噪声}$$

提纲

一、市调全局（宏观看）：

① 参数模型与算法模型————研究框架；

② 无监督学习与有监督学习————调查目的：探索？还是证明观点？

③ 相似性、分类和回归——看数据结构，即,Y是离散还是连续？

二、建模（微观看）：

① 如何衡量机器学习算法的有效性；

② 交叉验证如何提高机器学习模型的准确性；

数据

特征（自变量，X）

标记（因变量，Y）

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

训练集

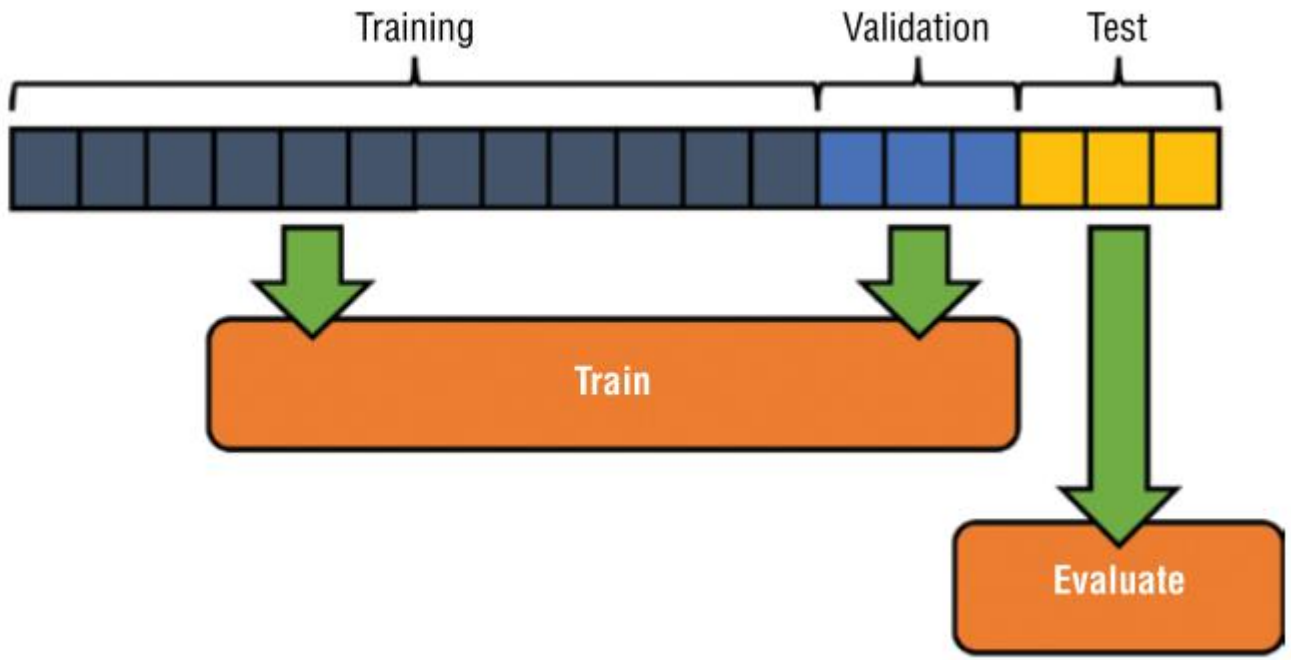
1	青绿	蜷缩	沉闷	?

测试集

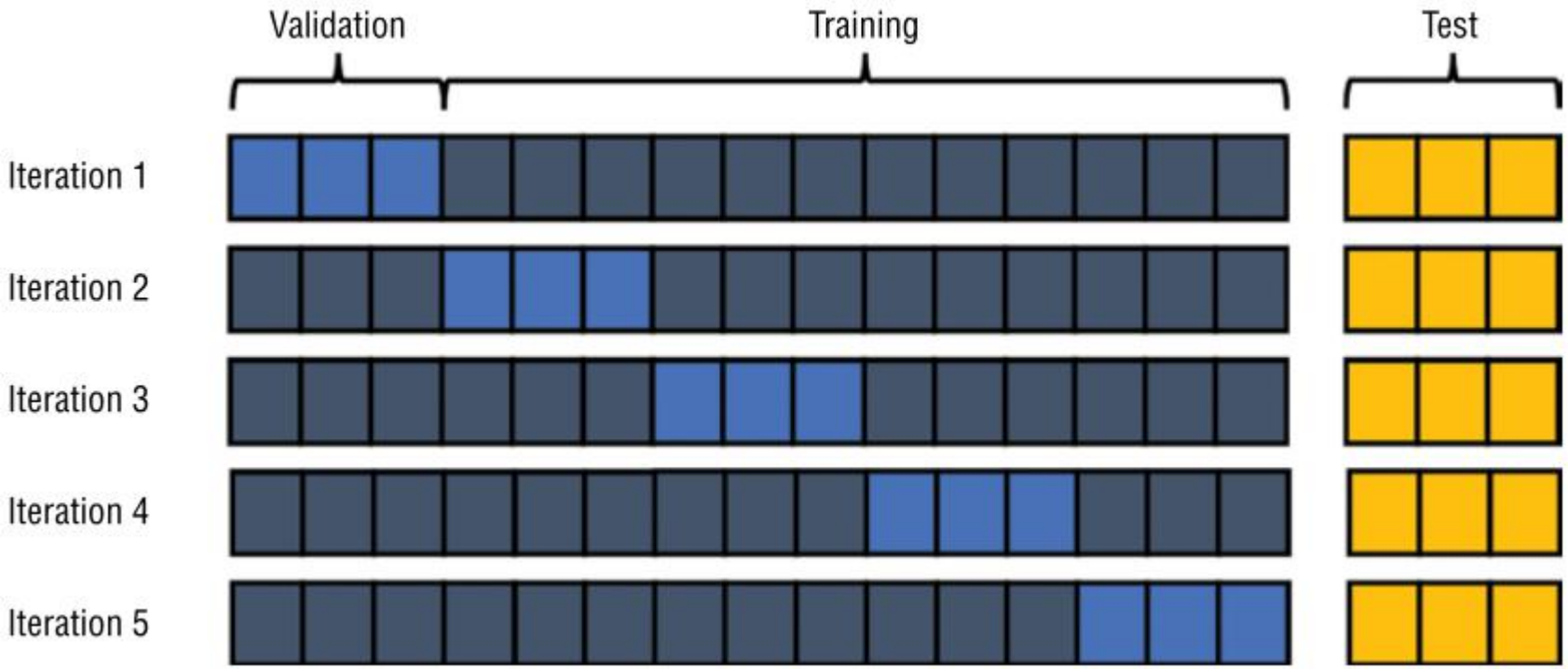
Step 1
Split the data into training, validation, and test partitions.

Step 2
Train and tune a model using the training and validation data.

Step 3
Evaluate the final model using the test data.



留一法


Validation
Training
Test
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

交叉验证法

常见机器学习模型：

- 线性回归
- 基于树的模型：决策树、集成算法（boosting、 bagging、随机森林)
- 神经网络————深度学习
- 支持向量机
- 朴素贝叶斯
- 关联规则
- K-means（聚类)

总结

由于课程时间所限，并没有讲解具体算法的原理及其应用方法。但是，在实际项目中必须掌握常用算法的**原理**、**应用场景**、**数据准备方法**、**结果解读方法及其注意事项**。

市场调研的本质

为学日益，为道日损。损之又损，以至于无为，无为而无不为。

——老子

市场调研的本质

- 在知识层次上，有意义、有结构的信息越多越好，可以增加判断的依据；
- 在智慧层次上，信息量却是越少越好，大多信息反而增添直觉的负担。

- 因此，信息需要过滤、梳理、归纳，才能提供智慧。

有什么想法，问题请使用电子邮件反映给我！！！

Email: liyi@sxufe.edu.cn