机器学习应用分享

主讲人：孟现超

个人介绍

姓名：孟现超

现就职某咨询管理公司，任分析经理

项目经历：

(1) 顾客分群

对顾客进行聚类划分，分别制定不同的市场营销策略，提高营销沟通效率，促成更多销售转化

(2) 中签用户付款概率预测

品牌会不定期进行尖货抽签抢购的活动，为了提高中签人群的付款比例，需要建模筛选出高分人群

(3) 个性化产品推荐

在品牌自有官网及小程序中，使用已有的CRM数据建模，为每一位用户提供个性化的产品推荐清单

目录

学习心得

机器学习回顾

学习建议

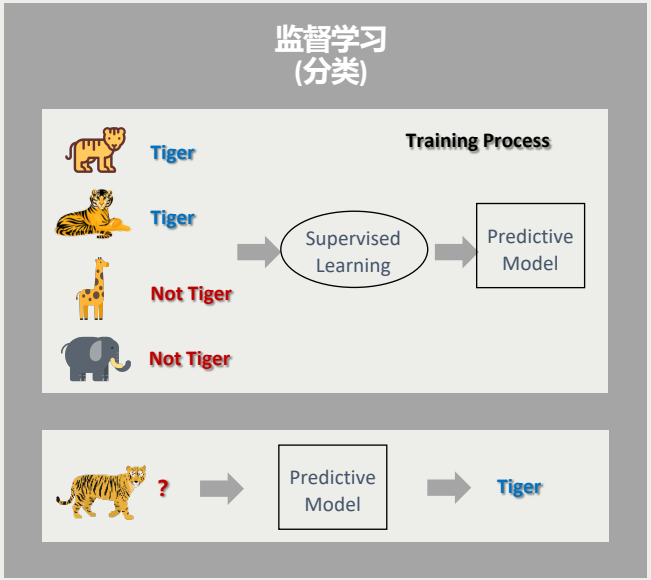
机器学习模型分类

监督学习: 基于有标签的训练集，训练模型来预测新样本的标签

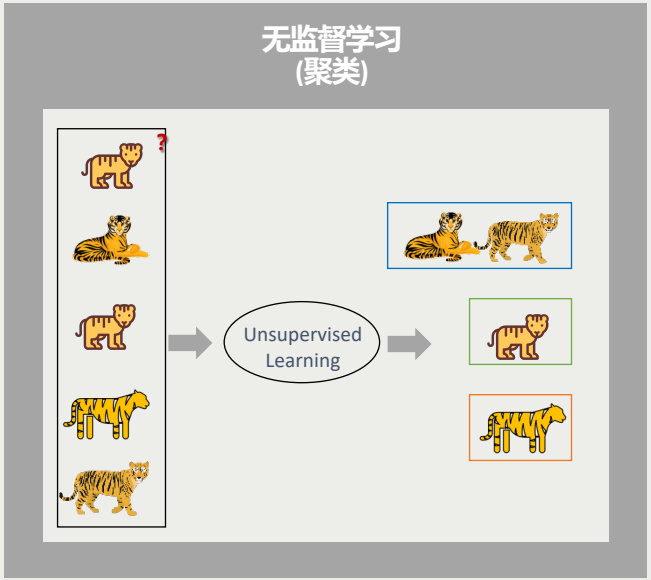
训练学习老虎的特征 → 分辨一只动物是否为老虎

无监督学习: 基于无标签的训练集，训练模型以识别数据内在的结构

根据动物特征将其分群 → 赋予每个聚类适当的定义



VS.

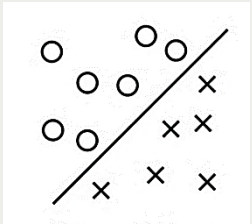


监督学习的两种应用

分类预测

当数据的标签属于**离散型**变量时，我们使用**分类模型**进行预测

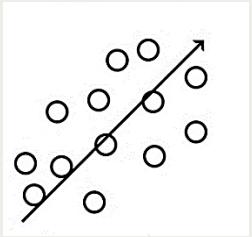
场景举例：判断贷款申请人是否会违约



回归预测

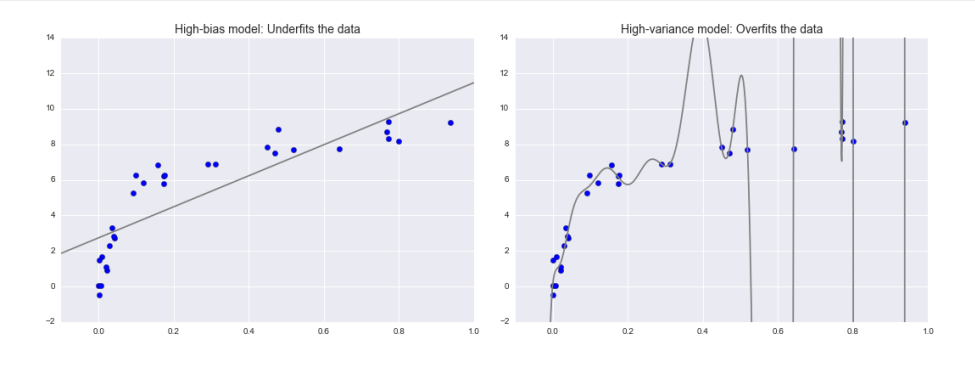
当数据的标签属于**连续型**变量时，应当使用**回归模型**进行预测

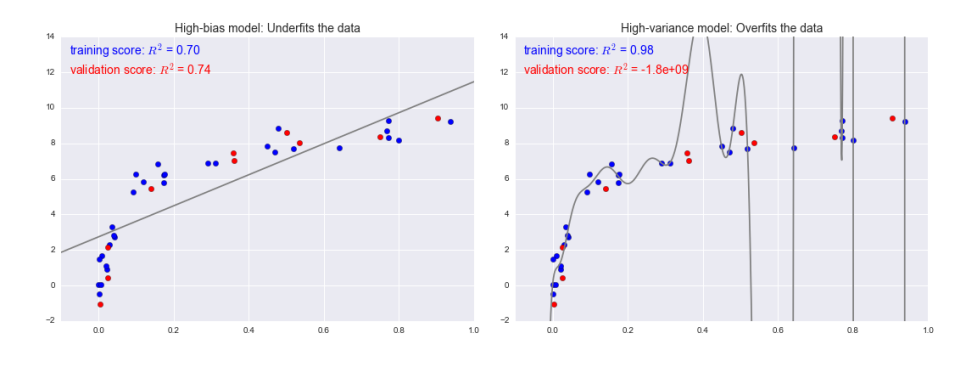
场景举例：预测某地区房价或某产品的销量

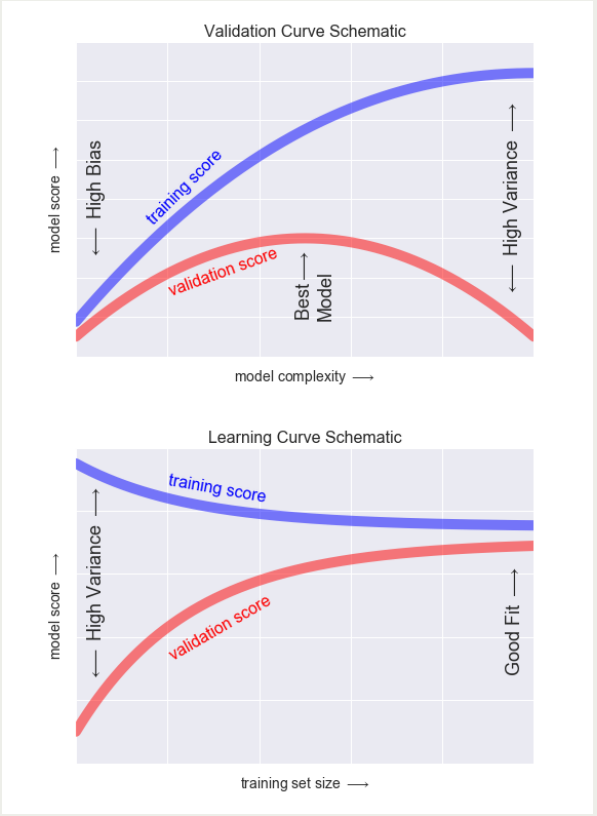


模型的验证与选择

模型需要兼顾准确性和泛化能力：模型复杂度越高验证集的准确性会先升后降，增加训练集样本数可以提升泛化能力





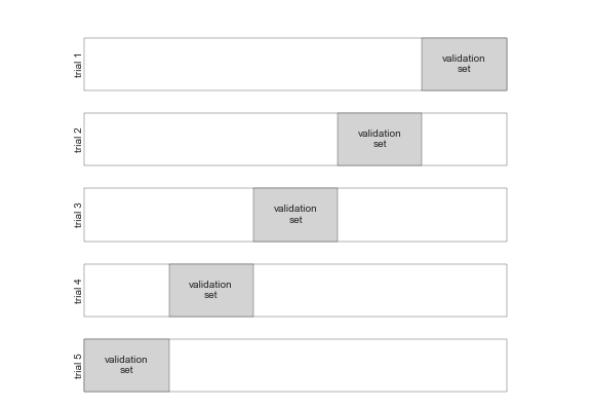


模型选择：交叉验证

为了选择好的模型，可以采用交叉验证方法，其基本想法是重复地使用数据

简单交叉验证

S折交叉验证



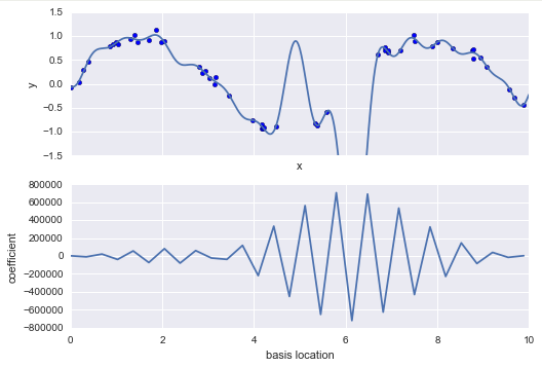
留一交叉验证

S折交叉验证的特殊情形是S = N，称为留一交叉验证（leave-one-out cross validation），往往在数据缺乏的情况下使用。这里，N是给定数据集的容量。

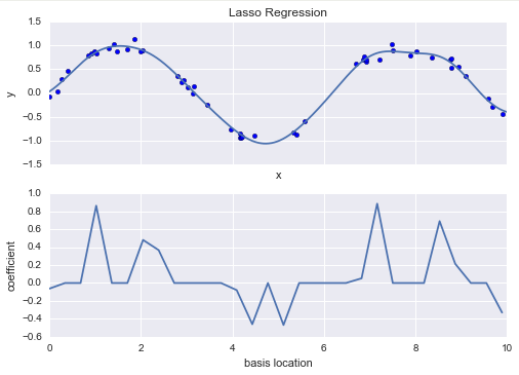
模型选择：正则化

正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项（regularizer）或罚项(penalty term)

线性回归模型

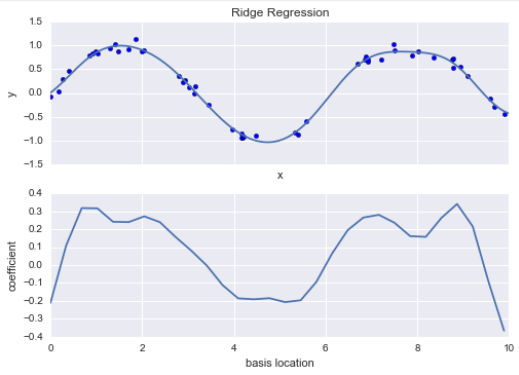


L1-正则



$$P = \alpha \sum_{n=1}^{N} |\theta_n|$$

L2-正则



$$P = \alpha \sum_{n=1}^{N} \theta_n^2$$

特征工程

数据集本身的质量决定了模型准确性的上限，模型调优只是不断逼近这个上限，特征工程是提升数据质量的重要方式

缺失值填充

在现实的数据训练集中，各种缺失值的存在会影响建模过程中的计算，或者导致最终结果的偏差。一般要视场景的差异，对缺失值进行“删除”或者“填补”处理

数据归一化

数据归一化的目的是消除量纲的影响，把数据转换到同一量级，从而使其具有可比性。Min-Max和Z-score是比较常用的两种方法

类别特征

One-hot编码：将类别型变量变为0或1

文本特征

Word2Vec：拆分文本，构造为词向量

TF-IDF：降低文本中高频词的权重

数值特征

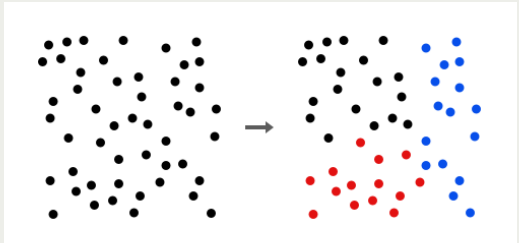
多项式变换：构造高维特征

无监督学习的两种应用

聚类分析

识别出数据集内明显的聚集模式，实现对数据集的群组划分

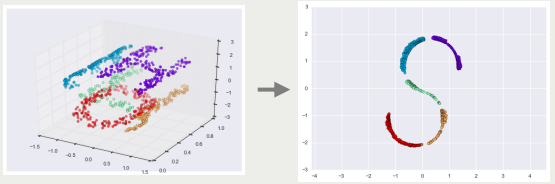
场景举例：聚类客户以针对性营销



数据降维

识别出高维数据在低维空间上的数据结构，实现数据从高维到低维的映射

场景举例：高维数据的可视化



机器学习相关算法

有监督

- 线性回归
- 多项式回归
- KNN
- 朴素贝叶斯
- 支持向量机
- 逻辑回归
- 决策树
- 随机森林
- GBDT

……

无监督

- K-means聚类
- 层次聚类
- 密度聚类
- 主成分分析

……

学习建议

- 熟练使用sklearn等机器学习库提供的工具
- 《数据科学手册》、《机器学习实战》

- 算法原理及公式推导
- 《统计学习方法》、《机器学习》（西瓜书）

- 大数据分析，分布式机器学习工具Pyspark
- 深度学习， tensorflow/pytorch

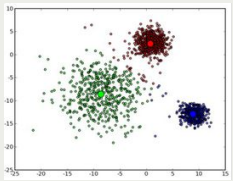
案例分享

顾客分组
个性化推荐
新品购买意向

顾客分组聚类

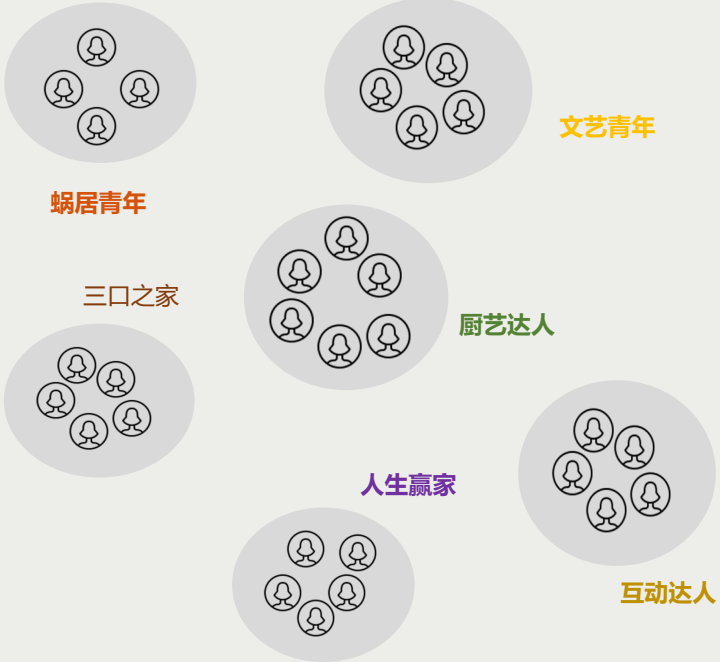
输入

算法



使用K-means聚类模型将目标人群按照彼此之间的相似度（或距离远近）划分为指定的k类群体，根据各群体人群画像所体现出的特征偏好进行更细致的营销活动

输出



个性化推荐

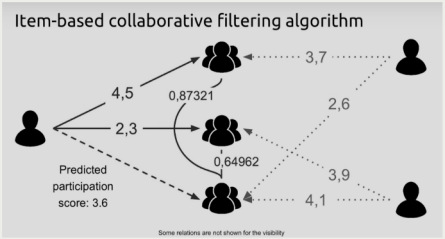
ID	SKU1	SKU2	SKU3	SKU4	…	SKU22	Avg. score
王翠花	10	2	0	0	…	3	4
					…		
					…		
	…	…	…	…	…	…	
ID-2	8	3	1		…		
					…		
					…		
…	…	…	…	…	…	…	
ID-n	0	0	0	0	…	0	
Avg. score	5						

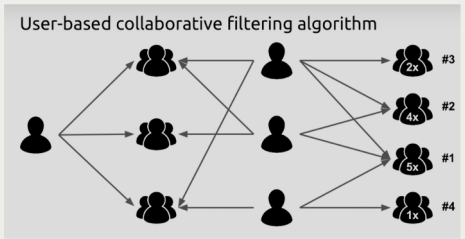
Index-a:
10/4=2.5

综合考虑横纵向指数，按数值从高到低，排出TOP10的SKU

Index-b:
10/5=2

协同过滤模型，包括产品缺失和冷启动





Top 10 SKU

示例: **王翠花**

排名	SKU
1	xxxx酱500g/瓶
2	xxx海鲜xx500g/瓶
3	xxxx爆炒xx500g/瓶
4	xxxx锅酱500g/瓶
5	xxxxx珍酱310g/瓶
6	xx黑胡椒汁310g/瓶
7	xxx印尼风味xxx310g/瓶
8	xxx印度风味孜xx酱310g/瓶
9	xx辣鲜xx448g/瓶
10	xxx酸辣xxx468g/瓶

高

低

新品购买意向

找到购买过和新品相似的N个现有产品的消费者和没有购买的消费者作为基准，通过机器学习模型对新的消费者进行投票，找到有可能购买N个像是产品的人群作为新品推荐的人群。

工作体会

个人的一些见解

人工智能岗位

研究类

大牛博士 + 教授

研究最前沿的技术

发顶会的Paper

应用类

硕士、博士居多

综合落地到产品

计算机视觉

大数据分析平台

业务类

市场需求最大

结合业务逻辑

特征工程

微调模型

Thank You