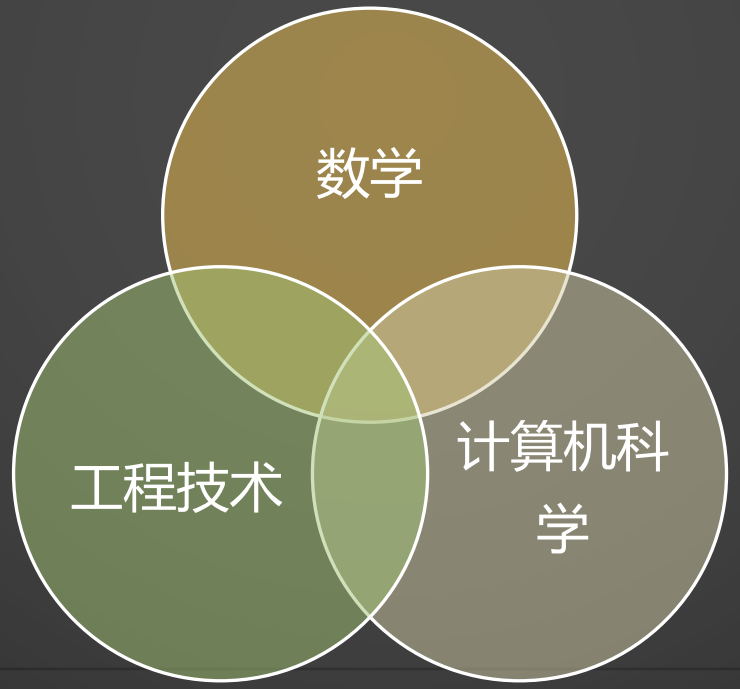
机器学习介绍及问题分类

什么是机器学习?

- 机器学习，Machine Learning，ML
- 机器学习指让计算机具备像人一样的学习能力，从海量的数据中获取新的知识或者技能。



应用

机器学习已经应用到日常生活的方方面面，大致有这么一些方向。

- 推荐系统：例如个人性化新闻推荐、商品推荐、个性化广告等。
- 自然语言处理（NLP）：例如文本分类、语音识别、聊天机器人等。
- 计算机图像识别（CV）：例如人脸识别、车辆识别、物体检测等。
- 数据挖掘（DM）：例如金融机构预测是否逾期、客户分群管理等。

机器学习的两类问题

监督学习

分类问题

- 决策树、逻辑回归、支持向量机等

回归问题

- 线性回归

无监督学习

聚类分析

- 层次聚类法、k均值聚类、DBSCAN

关联分析、主成分分析

监督学习

- 监督学习，supervised learning
- 主要用于解决分类问题，例如金融机构预测一个用户是/否会逾期
- 样本数据包括两部分：特征（X）和类别（y）（也叫标签）

- 监督学习算法根据样本的特征去预测类别（标签）

分类问题

- 给定一封邮件，判断是/否为垃圾邮件，已知的样本数据需要包括电子邮件的特征和类别（标签）。
- 基于医学影像判断肿瘤是/否为良性，已知的样本数据需要包括影像的特征和类别。
- 检测信用卡交易中是/否存在欺诈行为，已知的样本数据需要包括信用卡交易记录的特征和类别。

回归问题

- 线性回归，Linear Regression
- 主要用于预测一个连续数值
- 样本数据包括数值型自变量X和数值型因变量y

- 例如，根据教育水平x1、年龄x2和居住地x3等，预测一个人的年收入y

无监督学习

- 无监督学习，unsupervised learning
- 主要用于聚类、降维，例如K均值聚类法、主成分分析（PCA）等。
- 样本数据只有样本的特征（X），而没有类别（y）（标签）

- 例如，用户分群管理，已知的样本数据只有用户的特征，如消费金额、消费频次等，需要根据用户的特征将用户分为不同的类别（类别事先未知）

机器学习的一般步骤

1. 分析问题，数据探索及预处理
2. 选择合适的模型，利用训练集训练模型
3. 模型评估及使用，利用测试集评估模型

我的第一个机器学习模型

- 鸢尾花是一种植物，如下图所示。
- 有4个特征：
  - 花萼(sepal)长度、
  - 花萼宽度
  - 花瓣(petal)长度
  - 花瓣宽度
- 三个类别：Setosa，Versicolour，Virginica
- 任务：根据这四个特征对鸢尾花进行分类。

机器学习中的基本术语

- **样本：**数据中的每一行被称为一个样本。例如，鸢尾花分类问题，有150个样本
- **特征：**数据中的每一列，或者说样本的属性被称为特征。例如，鸢尾花分类问题，每个样本有4个特征

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

说明：关于特征的专项课题有特征提取或者特征工程

机器学习中的基本术语

- **类别：**在分类问题中，每一个样本都有一个类别或者标签。
- **二分类问题：**一般常见的分类问题都是二分类问题，有两个类别，正类和反类，也可以叫负类。
- **多分类问题：**多分类问题有三个及三个以上的类别，例如，鸢尾花分类，是一个三分类问题，有三个类别。

```
- class:
        - Iris-Setosa
        - Iris-Versicolour
        - Iris-Virginica
```

- **训练集、测试集：**通常是将收集好的带标签的数据分成两部分，一部分数据用于构建机器学习模型，叫做训练集（train set），其余的数据用来评估模型性能，叫做测试集(test set)。

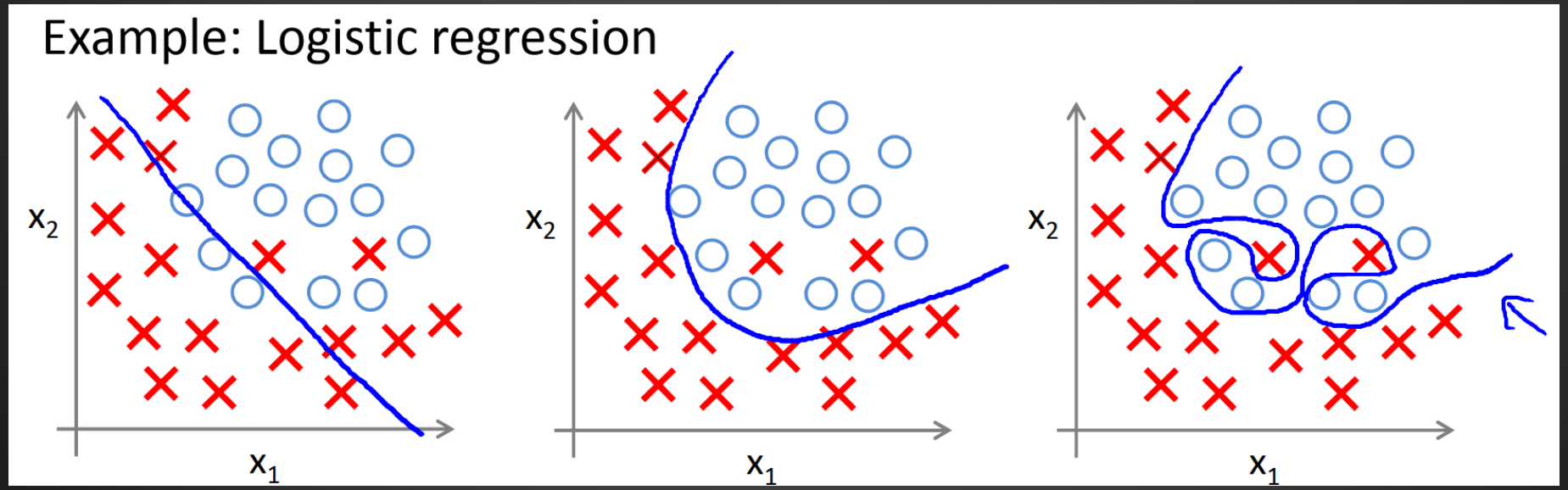
机器学习中的基本术语

- **精度：**预测正确的比例，计算方式：正确分类数/样本总数。
  - 作用：用来衡量模型的优劣
  - 说明：除了精度之外，还有其他的评价指标，例如混淆矩阵、ROC曲线等。

- **泛化：**如果一个模型能够对没见过的数据做出准确预测，就说这个模型能够从训练集泛化到测试集。一般要构建一个泛化精度尽可能高的模型。

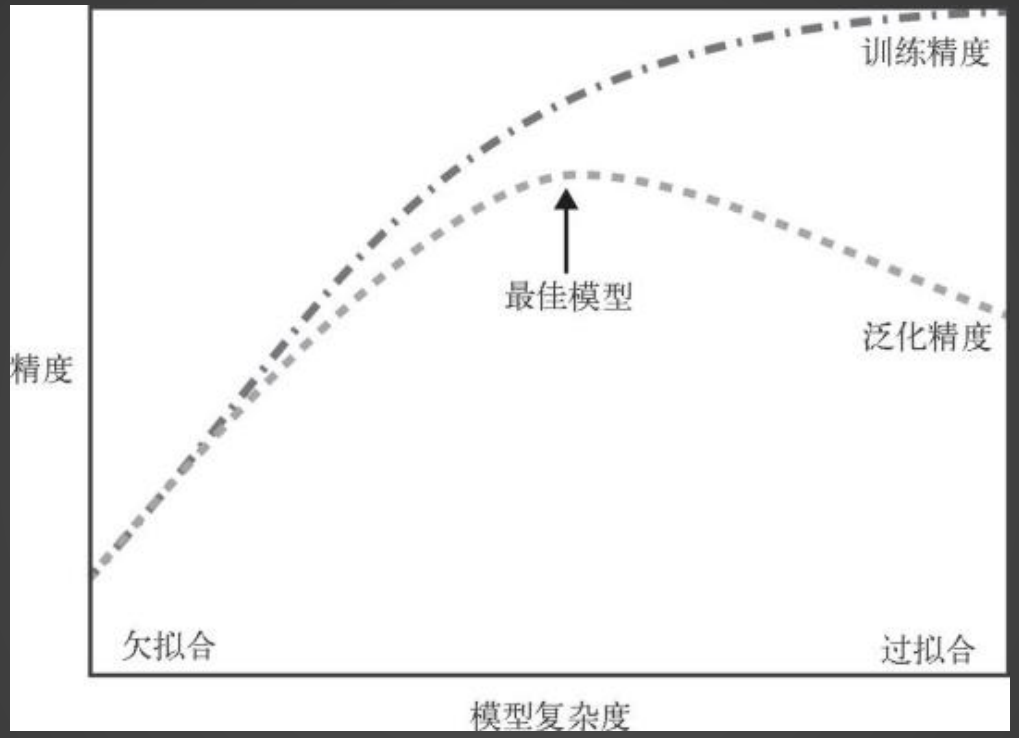
过拟合与欠拟合

- **过拟合（overfitting）：** 如果拟合模型的时候，过分关注训练集的细节，得到了一个在训练集上表现好，但不能泛化到新的数据集上的模型，那么存在过拟合。
- **欠拟合（underfitting）：** 如果模型过于简单，可能无法抓住数据的全部内容及数据中的变化，得到的模型在训练集上的表现就很差，选择过于简单的模型被称为欠拟合。



图片摘自吴恩达机器学习课程讲义

过拟合和欠拟合之间的权衡



我们需要在模型复杂度与训练集精度和测试集精度之间做一个权衡。