大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

目　录

- 样本Samples
  - 用于学习或评估的数据项或实例
- 特征Features
  - 属性集，通常表示为与样本相关的向量
- 标记Labes
  - 在样本上指定的值或类别
  - 分类问题中，样本被指定为特定的类别，称为标记
  - 在回归问题中，项被指定为实数的值，称为标记

- 训练样本Training sample
  - 用于训练学习算法的样本
  - 例如，对于垃圾邮件问题，训练样本由一组邮件样本以及给定的标签（Labels）组成
- 验证样本 Validation sample
  - 用于在使用标记数据时调整学习算法参数的样本
  - 学习算法通常具有一个或多个自由参数，因而验证样本用于为这些模型参数选择适当的值
- 测试样本Test sample
  - 用于评估学习算法性能的样本，这些样本具有给定的标签
  - 算法训练结束后，对测试样本进行预测，然后将这些预测与测试样本的标签进行对比，用于衡量算法的性能

- 损失函数Loss function
  - 用于度量预测标签和真实标签之间的差异或损失的函数
- 抽象 Abstraction
  - 其含义是将数据转化为更广泛的表示
- 泛化Generalization
  - 它形容将抽象知识转化为可用于动作形式的过程，它也是学习算法具有学习数据集的经验后，可以对未知样本正确地进行处理的能力

- 欠拟合是指模型不能在训练集上获得足够低的误差，
- 过拟合是指训练误差和测试误差之间的差距太大。

- 机器学习的主要挑战是我们的算法必须能够在先前未观测到的新输入上表现良好，而不只是在训练集上表现良好。而这种在先前未观测到的输入上表现良好的能力被称为泛化能力。
- 为了得到泛化能力好的学习器，我们应该从训练样本中尽可能学出适用于所有潜在样本的“普遍规律”，这样才能在遇到新样本时做出正确的判别。
- 如果我们把训练样本的一些特有的特点也当做潜在样本的一般性质，这样就会导致泛化能力下降，这也是我们常说的“过拟合”现象。

4、经典的机器学习方法

- 所谓监督学习与非监督学习，是指训练数据是否有标注类别，若有则为监督学习，若否则为非监督学习。监督学习是根据输入数据（训练数据）学习一个模型，能对后来的输入做预测。
- 在监督学习中，输入变量与输出变量可以是连续的，也可以是离散的。若输入变量与输出变量均为连续变量，则称为回归；输出变量为有限个离散变量，则称为分类。

4、经典的机器学习方法

- 4.1 分类算法原理
- 4.2 决策树算法
- 4.3 K-近邻分类算法（KNN算法）
- 4.4 K-均值聚类算法（K-means算法）
- 4.5 Apriori关联规则算法

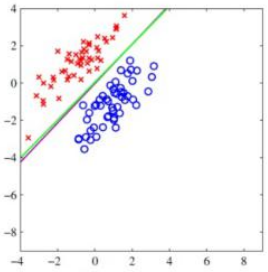
什么是分类？

- 分类是基于包含已知类别成员观测值的训练数据集，来辨识新的观测值属于哪一组类别的任务。
- 关于分类器（Classifier）
  - 一种实现分类、尤其是构成一种具体实现的算法，被称为分类器
  - “分类器”有时也指由分类算法所实现的数据函数，它将输入数据映射为一个类别。

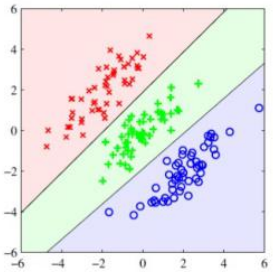
Classes 类别

$$y_k(\mathbf{x}) = \mathbf{w}_k \cdot \mathbf{x} + b$$

- Two classes: 二元分类： $k = 2$
- Multiple classes: 多元分类： $k > 2$



Two classes
二元分类



Three classes
三元分类

- 分类问题一般包括两个步骤

1、模型构建（归纳）

通过对训练集合的归纳，建立分类模型。

学习模型可以用分类规则、判定树或数学公式的形式提供

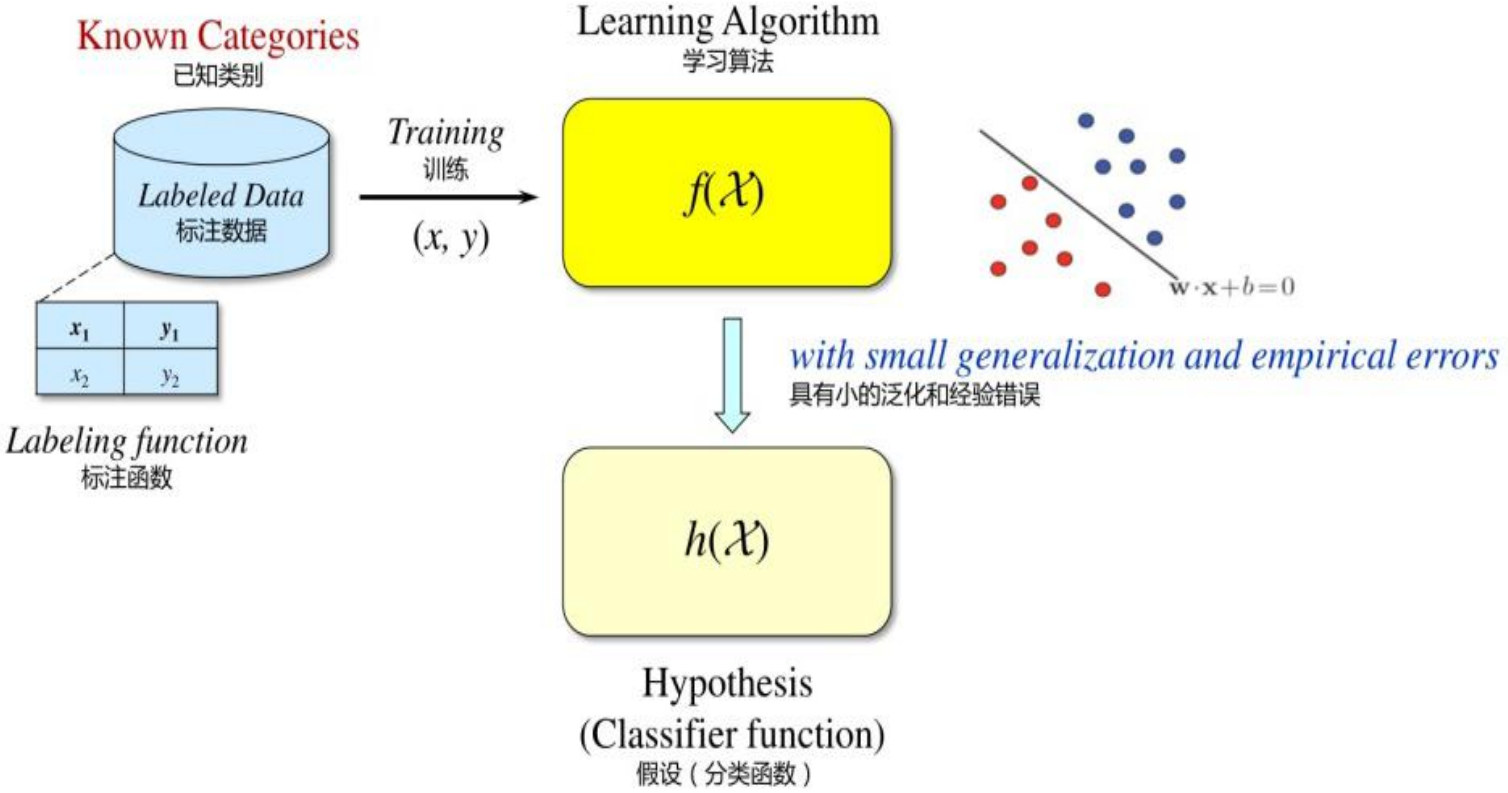
2、预测应用（推论）

根据建立的分类模型，对测试集合进行测试。

模型在给定测试集上的准确率是正确被模型分类的测试样本的百分比

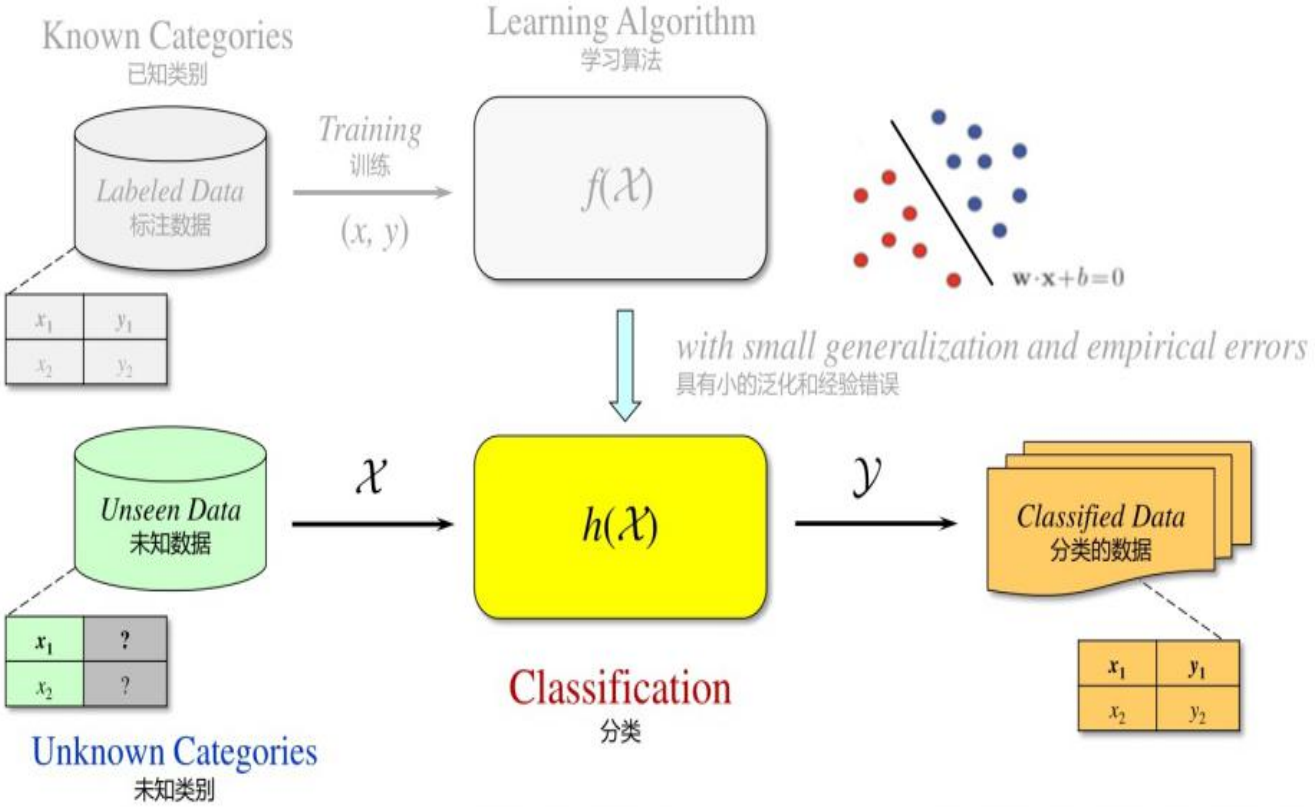
如何进行分类？

- 第一步：训练Training



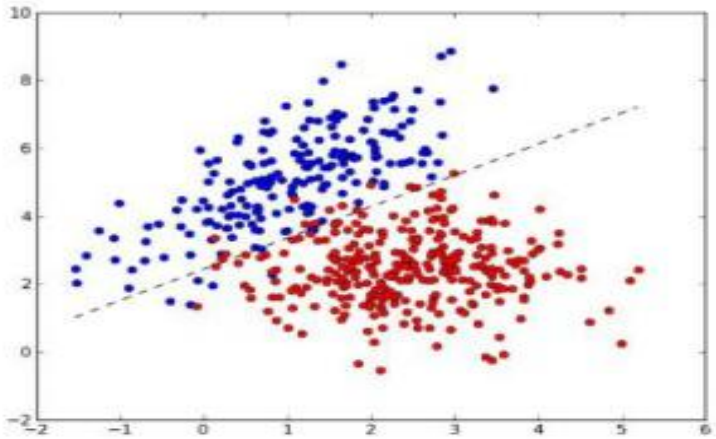
如何进行分类？

- 第二步：测试Testing



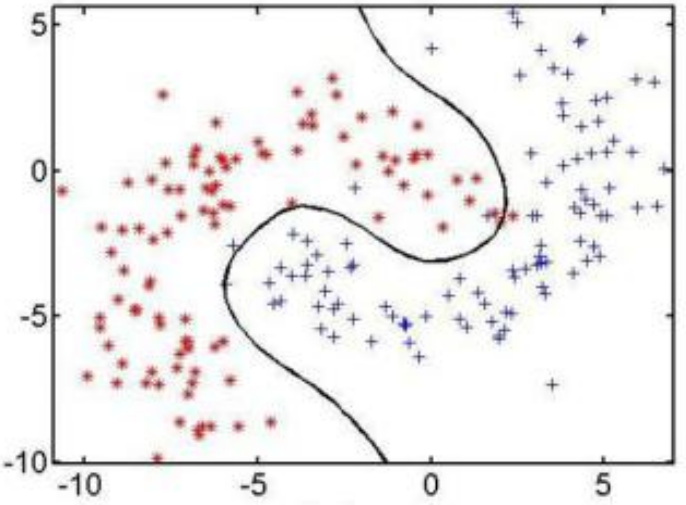
线性分类和非线性分类

- 线性分类通过线性分类器进行分类



- 线性分类器是具有一个线性决策边界的线性判断函数

- 非线性分类通过非线性分类器进行分类



- 非线性分类器具有若干非线性决定边界，并且可能是非连续决定边界