

机器学习笔记

陈鸿峥

2019.07 *

目录

1	简介	1
1.1	分类	1
1.2	历史	2
2	基础概念	3
2.1	训练集与测试集	3
2.2	参数选择	4
2.3	性能度量	4
3	线性模型	5
3.1	线性回归	5
3.2	线性分类	6
3.3	线性判别分析	6
3.4	类别不平衡问题	7
4	参考资料	7

本笔记对应周志华的《机器学习》(西瓜书)。

1 简介

1.1 分类

机器学习(machine learning)通常可以分为以下几类:

- 监督学习(supervised learning): 有标签(label)
 - 回归(regression): 连续值

*Build 20190723

- 分类(classification): 离散值
- 无监督学习(unsupervised learning): 无标签
 - 降维
 - 聚类(clustering)
- 强化学习(reinforcement learning): 延后的标签

1.2 历史

- 推理期(1950-1970): 逻辑理论家(Logic Theorist)A. Newell & H. Simon(1975图灵奖)
- 知识期(1970s): 知识工程之父E. A. Feigenbaum(1994图灵奖)
- 学习期(1980s): 决策树、归纳逻辑程序设计(Prolog)

机器学习的五大学派(tribe):

- 符号主义(symbolist)
- 联结主义(connectionist)
- 进化主义(evolutionaries)
- 贝叶斯主义(bayesians)
- 类比主义(analogizers)



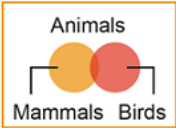

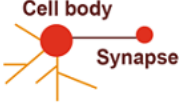


A look at

Machine learning evolution

Overview

For decades, individual “tribes” of artificial intelligence researchers have vied with one another for dominance. Is the time ripe now for tribes to collaborate? They may be forced to, as collaboration and algorithm blending are the only ways to reach true artificial general intelligence (AGI). Here’s a look back at how machine learning methods have evolved and what the future may look like.

What are the five tribes?

Symbolists	Bayesians	Connectionists	Evolutionaries	Analogizers
				
Use symbols, rules, and logic to represent knowledge and draw logical inference	Assess the likelihood of occurrence for probabilistic inference	Recognize and generalize patterns dynamically with matrices of probabilistic, weighted neurons	Generate variations and then assess the fitness of each for a given purpose	Optimize a function in light of constraints (“going as high as you can while staying on the road”)
Favored algorithm Rules and decision trees	Favored algorithm Naive Bayes or Markov	Favored algorithm Neural networks	Favored algorithm Genetic programs	Favored algorithm Support vectors

Source: Pedro Domingos, *The Master Algorithm*, 2015

2 基础概念

2.1 训练集与测试集

- 留出法(hold-out): 直接将数据集划分为两个互斥的集合, 一个作为训练集, 另一个作为测试集
 - 注意数据分布的一致性, 通过多次随机划分取平均保证
 - 通常用 $2/3 \sim 4/5$ 的样本用于训练, 其余用作测试
- 交叉验证法(cross validation)/ k 折(fold)交叉验证: 划分为 k 个互斥子集, 其中 $k - 1$ 个用于训练, 最后一个用于验证, 训练 k 次, 对这 k 次结果取平均
 - 通常采用10次10折交叉验证, 每次都换划分方式, 确保随机性
- 自助法(bootstrapping): 从原始数据集中放回采样得到新数据集作为测试集
 - 在数据集较小的、难以有效划分训练/测试集时比较有用

- 但改变了初始数据集分布，会引入估计偏差

注意：通常将习得模型实际使用中遇到的数据称为测试集，而将训练的数据划分为训练集与验证集(validation)

2.2 参数选择

参数通常包括

- 模型本身的参数(parameter)：通过学习改变
- 超参数(superparameter)：预先设定，调参实际上就是在选择算法

2.3 性能度量

- 回归：通常采用均方误差(MSE)
- 分类：错误率、精度

对于二分类问题，有混淆矩阵(confusion matrix)

	预测正	预测反
真正正	TP(真正例)	FN(假反例)
真实反	FP(假正例)	TN(真反例)

$$\text{查准率 } P = \frac{TP}{TP + FP}$$

$$\text{查全率 } R = \frac{TP}{TP + FN}$$

为了衡量机器学习算法的泛化性能，需要知道以下指标：

- 方差(variance)：度量同样大小的训练集的变动所导致的学习性能的变化，即刻画数据扰动所造成的影响

$$\mathbb{D}(\mathbf{x}) = \mathbb{E}_D \left[(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \right]$$

- 偏差(bias)：度量学习算法的期望预测和真实结果的偏离程度，即刻画学习算法本身的拟合能力

$$b^2(\mathbf{x}) = (\bar{f}(\mathbf{x}) - y)^2$$

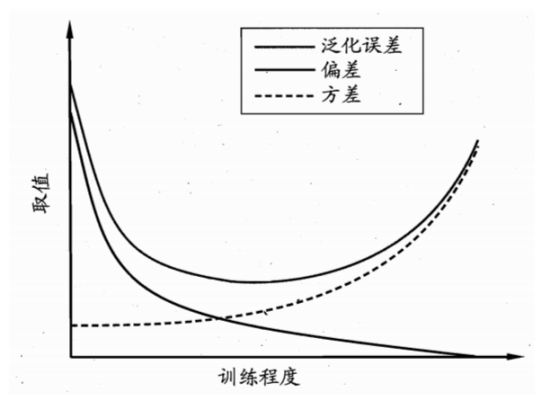
- 噪声(noise)：表达在当前任务上任何学习算法所能达到的期望泛化误差的下界

$$\varepsilon^2 = \mathbb{E}_D \left[(y_D - y)^2 \right]$$

泛化误差可以分解为

$$E(f; D) = b^2(\mathbf{x}) + \mathbb{D}(\mathbf{x}) + \varepsilon^2$$

即泛化性能是由学习算法的能力、数据的充分性及学习任务本身的难度决定的



3 线性模型

线性模型

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + \mathbf{b}$$

由于 \mathbf{w} 直观表达了各属性在预测中的重要性，因此线性模型具有很好的可解释性(comprehensibility)。

3.1 线性回归

数据集 $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ ，每个样本 \mathbf{x}_i 都由 d 个属性描述，多元线性回归希望(multivariate linear regression)学到

$$f(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i + b, \text{ s.t. } f(\mathbf{x}_i) \simeq y_i$$

将 \mathbf{w} 和 b 写在一起变成 $\mathbf{w} \leftarrow \begin{bmatrix} \mathbf{w} & b \end{bmatrix}$ ，并设

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} & 1 \\ x_{21} & x_{22} & \cdots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{md} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T & 1 \\ \mathbf{x}_2^T & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^T & 1 \end{bmatrix}$$

为已知， \mathbf{w} 为需要训练的权重。再将标记写成向量形式 $\mathbf{y} = [y_1 \ y_2 \ \cdots \ y_m]$ ，进而得到最小二乘优化

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \|\mathbf{y} - X\mathbf{w}\|_2^2$$

对 \mathbf{w} 求导有

$$\nabla_{\mathbf{w}} E_{\mathbf{w}} = 2X^T(X\mathbf{w} - \mathbf{y})$$

当 $X^T X$ 满秩或正定时，令上式为0有

$$\mathbf{w}^* = (X^T X)^{-1} X^T \mathbf{y}$$

但现实中大多数时候 $X^T X$ 都非可逆阵，故常用正则化方法。

3.2 线性分类

单位阶跃(unit-step)函数

$$y = \begin{cases} 0 & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$$

不连续, 故用对数几率(logistic)函数替代

$$y = \frac{1}{1 + e^{-z}}$$

这是一种Sigmoid函数, 将线性表达式代入有

$$y = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

进而有

$$\mathbf{w}^T \mathbf{x} + b = \ln \frac{y}{1 - y}$$

将 y 视为样本 \mathbf{x} 为正例的可能性, 则 $1 - y$ 为反例可能性, 两者比值 $y/(1 - y)$ 称为几率(odds), 反映了 \mathbf{x} 作为正例的相对可能性。对数几率又称logit, 故这种方法又称为逻辑斯蒂(logistic)回归, 但其实是分类学习方法。

将 y 视为后验概率估计, 有

$$\ln \frac{\mathbb{P}(y = 1 | \mathbf{x})}{\mathbb{P}(y = 0 | \mathbf{x})} = \mathbf{w}^T \mathbf{x} + b$$

显然有

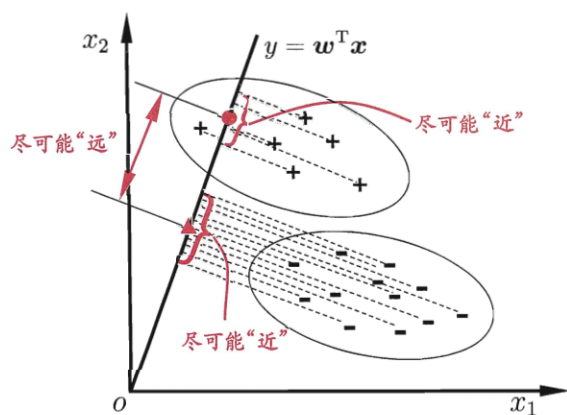
$$\begin{aligned} \mathbb{P}(y = 1 | \mathbf{x}) &= \frac{e^{\mathbf{w}^T \mathbf{x} + b}}{1 + e^{\mathbf{w}^T \mathbf{x} + b}} \\ \mathbb{P}(y = 0 | \mathbf{x}) &= \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x} + b}} \end{aligned}$$

给定数据集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, 由极大似然法估计 \mathbf{w} 和 b 有

$$\ell(\mathbf{w}, b) = \sum_{i=1}^m \ln \mathbb{P}(y_i | \mathbf{x}_i; \mathbf{w}, b)$$

3.3 线性判别分析

线性判别分析(Linear Discriminant Analysis, LDA)[Fisher, 1936]同样用于二分类, 希望将样例投影到一条直线上, 使得同类样例投影点尽可能近, 异类样例投影点尽可能远。



最优化广义瑞利商

$$J = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}$$

其中 S_w 为类内散度矩阵， S_b 为类间散度矩阵。

LDA 也是经典的监督降维技术。

3.4 类别不平衡问题

- 欠采样(undersampling): 减少样例
- 过采样(oversampling): 增加样例
- 阈值移动(threshold-moving)/再缩放(rescaling)

$$\frac{y'}{1 - y'} = \frac{y}{1 - y} \times \frac{m^-}{m^+}$$

4 参考资料

1. 周志华, 《机器学习》, 2016