张志华-统计机器学习笔记
1.绪论

By Hao ZHAN
Edit 2019.7.26

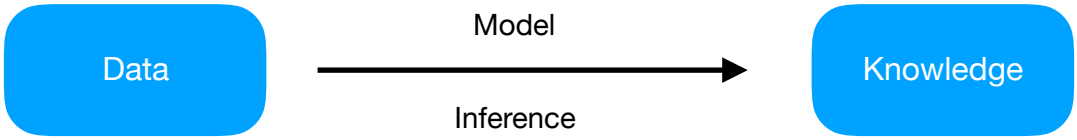
导言

第一次课属于导论性质，主要讨论以下八点类容：

- 知识的定义
- 机器学习与统计的关系
- 数据挖掘与机器学习的关系
- 为什么现在机器学习突然火了
- 什么是统计机器学习
- 如何学习机器学习
- 机器学习所要面对的问题
- 机器学习的基本方法

Part 01 知识的定义

关于知识的定义有很多，在机器学习中，一种普遍的看法是「知识是有用的信息」。计算机领域通常意义上的信息（information）就是我们所说的数据（data），从数据到「有用的数据」，即知识，往往需要经过建模和推理两个过程，而这个过程就可以看作是机器学习。



Part 02 机器学习与统计的关系

机器学习的目的在于获取知识，即试图得到一个有用的模型。而统计是为建模提供了完整的框架。一般认为机器学习时统计的分支，或者说机器学习就是应用统计学。

Larry Wasserman 曾在其著作 *All of Statistics* 中将统计学概念与机器学习概念进行对比：

Table.1 Statistics/Data Mining Dictionary

Statistics	Computer Science	Meaning

estimation	learning	using data to estimate an unknown quantity
Classification	Supervised learning	predicting a discrete *Y* from *X*
clustering	Unsupervised learning	putting data into groups
data	Training sample	$(X_1,Y_1),...,(X_n,Y_n)$
covariates	Features	X
classifier	Hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayes inference	statistical methods for using data to update beliefs
frequentist inference	—	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	Tong

PART 03 数据挖掘与机器学习的关系

数据挖掘和机器学习没有本质上的差别，如果非要进行区分的话，那么机器学习更偏向于完全的自动化，并且对数学的要求更高，而数据挖掘更偏向于半自动化，对模型的解释要求较高。

PART 04 为什么现在机器学习突然火了

近年来，机器学习突然火爆，这一潮流可以从理论因素和业界因素两个方面来进行阐释。

1.理论因素

John Hopcroft 将计算机的发展分为三个阶段：

（1）Work 阶段
在这一阶段中，计算机的主要目标是让计算机能够正常工作起来。在这一阶段中，主要的任务是构建程序语言、系统和数学基础。
（2）Make Computer Useful 阶段
在这一阶段中计算机的主要工作是使自身变得更加有用，在效率和准确度上超过人类。在这一阶段中，重点的方向是数据结构和算法。
（3）More Applications 阶段

在这一阶段中，计算机的主要工作是使自身应用在更多的领域和场景中。这一阶段的特征在于计算机所面对的数学基础从离散逐渐转移到概率和统计上来，即机器学习中所强调的概率和随机图。从图灵奖获得者的研究方向中也可以看出这一趋势。

2.工业界因素

越来越多的公司由于其业务结构开始涉足于人工智能领域。Google、Facebook等公司开始提出了Data Science 的概念，以及对应的数据科学家（data scientist）概念。数据科学家要求具有：（1）底层架构能力；（2）程序能力；（3）数学能力。

PART 05 什么是统计机器学习

Michael Jordan 将SML（statistics machine learning）定义为：”A field that bridges computation and statistics with information theorem, signal process, algorithm, control and optimization theorem.”

PART 06 如何学习机器学习

机器学习的知识架构主要分为四个部分：

矩阵：机器学习中的数据结构大多是以矩阵形式呈现的，机器学习也有很多特征分解等操作，因此矩阵的知识是必不可少的。

优化：最优化问题是机器学习模型计算的关键，求解模型需要优化方法的支持。

算法：算法不仅能够帮助我们计算出结果还可以提升效率。

统计：统计方法给出了建模的框架。

PART 07 机器学习面对的问题

1.降维问题

机器学习所面对的数据的维度通常较高，为了求解方便，我们通常需要对数据进行降维操作。

$$\overrightarrow{x_i} \in R^p \rightarrow \overrightarrow{z_i} \in R^q, \text{ where } p > q$$

降维的方式分为两种：

（1）线性降维

$$Z_{q\times 1} = A_{q\times p} X_{p\times 1}$$

$$\textit{where, } A^T A = I \quad \text{（列正交）}$$

（2）非线性降维

$$Z = f(x)$$

2.聚类问题

k-class问题

3.分类问题

包括二分类问题和多分类问题。

其中，我们需要划分：

Training data: for learning parameter
Validation data: for learning hyper parameter
Testing data: for predicting

4.回归问题 Regression

5.排序问题 Ranking

排序问题其实也是一种特殊的回归问题
Isotonic Regression

PART 08 机器学习的基本方法

1.频率派与贝叶斯派

（1）频率派
把Model Parameters 看作是未知的常数然后使用某种准则来进行估计
$y = X^T a$，估计 $min \sum_{i=1}^{n} \left(Y_i - X_i^T a\right)$，即最小二乘法
此外，最大似然法（Maximal Likelihood Estimation）的原理与最小二乘其实是相通的。

（2）贝叶斯派

其中σ 是一个随机变量

似然：$y \sim N\left(x^T a, \sigma\right)$
先验部分：$a \sim \ldots, \sigma \sim \ldots$
后验部分：$P\left(a \mid x\right)$

2.参数方法与非参数方法

（1）参数方法

The number of parameters is fixed.
一旦参数的个数固定袭来，那么就与训练数据的个数无关了

（2）非参数方法

参数不固定，依赖于数据的个数，参数可能无穷多

Example.01 最近邻算法 Nearest-neighbor
因为要计算每一个数据样本与待预测点的距离，因此参数的个数为数据点的个数，属于非参数方法。

Examole.02 Logistic Regression

参数的个数依赖于X的维度数量而不是数据数量，因此疏于参数方法。

频率派、贝叶斯派与参数方法、非参数方法互有交叉，可以是参数频率、非参数频率，也可以是参数贝叶斯，非参数贝叶斯。