

Discussion About D.A. & M.L.

Miao Feng

September 11, 2017






Before I start

Theme: 机器学习与资料同化的关系

PS.: 根据个人理解、在哈工大的学习整理所得。

Inferences:

-  周志华 (2016), 机器学习
-  李航 (2012), 统计学习方法
-  邹晓蕾 (2009), 资料同化理论与应用

B.T.W.:You may find many errors in this report,please help me correct it.

Outline

Basics of M.L.

The relationship of M.L.& D.A.

Some issues

Summary



Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

Some issues

Summary



What's M.L.

M.L.,D.L.,S.L.



① 1980s~1990s

中期

② 1990s 中期 ~

③ 2000s~

What's M.L.

M.L.,D.L.,S.L.

Definition

- M.L.

- ① **Mitchell:** 对于某类任务 T 和性能度量 P ，如果一个计算机程序在 T 上以 P 衡量的性能随着经验 E 而自我完善，那么我们称这个计算机程序在从经验 E 学习。
- ② **Samuel:** 在不直接针对问题进行编程的情况下，赋予计算机学习能力的一个研究领域。

- D.L.

狭义：多层神经网络

What's M.L.

S.L.,D.M.

Definition

- S.L.

统计学习，基于数据构建概率统计模型并运用模型对数据进行预测与分析

- D.M.

数据挖掘，指从大量的数据中通过算法搜索隐藏于其中信息的过程。算法主要使用 M.L. 的，但是重在数据中隐含的信息

What's M.L.

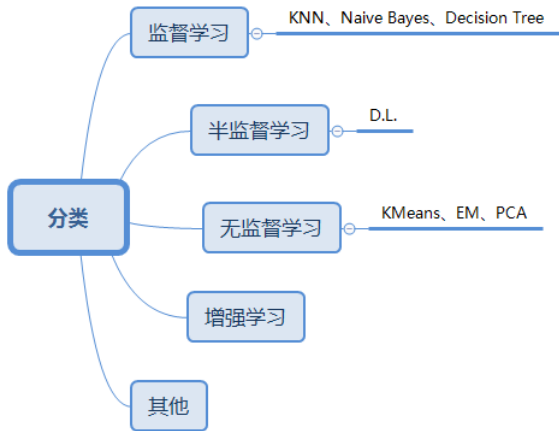
Forget things before

Be simple

- **M.L.** 基于数据进行模型训练，最终归纳出一个面向一种性能度量的决策。
- **M.L.** 现在的主流算法是基于连接主义的 **D.L.** 和基于统计学习的 **SVM**、**KNN**、**logistics** 回归、**EM**、**AdaBoost** 等

What's M.L.

Classifications



- 监督学习
测试集带 label。
- 半监督学习
测试集部分带 label。
- 无监督学习
测试集不带 label。
- 增强学习
没有规则化的数据集，多用于 A.I. 中机器人控制

Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

Some issues

Summary



M.L. Theory

Problem Definition

Definition

- 样本点（数据集）

输入数据与输出数据组成的样本对。数据集分为**训练集**与**测试集**。输出数据又被称为 **label**。训练集用于训练模型，测试集用于测试训练得到的模型的性能。

- 模型

输入空间到输出空间的一个映射

M.L. Theory

Still boring, try something practical

已知辐射值 $R = (r_1, r_2, r_3, \dots, r_n)^T$ ，对应温度为 $T = (t_1, t_2, t_3, \dots, t_n)^T$ ，欲知 $R' = (r_{n+1}, r_{n+2}, r_{n+3}, \dots, r_N)^T$ 时对应的温度

以辐射值为输入数据，温度为输出数据，则

样本对: $Y = \{[R \ R'], T\}$

训练集: $Y_1 = \{R, T\}$

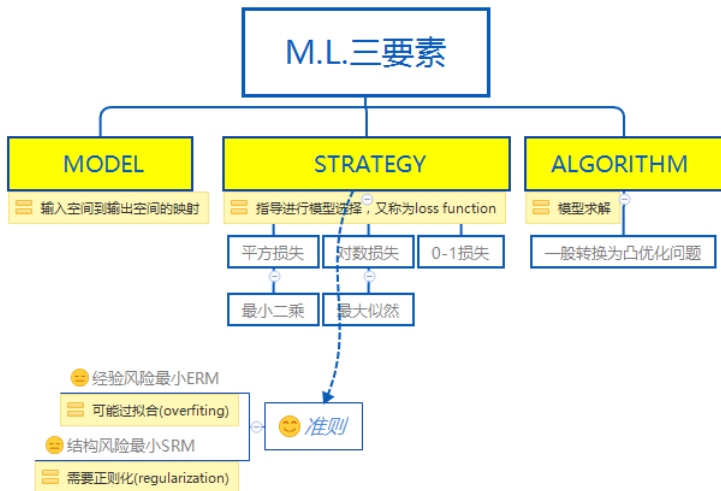
测试集: $Y_2 = \{R', [\]\}$

label: $T = (t_1, t_2, t_3, \dots, t_n)^T$

得到的模型等价于 **GMF**

M.L. Theory

Methods



M.L. Theory

ERM,SRM,Goal

- ① ERM, 说明与训练集的数据拟合的很好
- ② SRM, 说明对测试集的预测效果很好, 比如对已知的辐射值推测温度
- ③ 目标: 最小化结构风险的同时保证经验风险最小, 即确保模型的泛化能力, 在符合策略条件的众多模型中挑选最普适的模型

Therefore,we'd like to keep an eye on the **regularization**

M.L. Theory

Rules of choosing model

Principle: 奥卡姆剃刀 (Oscam's razor), 所有符合条件的模型中, 模型越简单, 泛化性能越好

Methods: 正则化 + 交叉验证
模型添加 penalty 数据集 trick

Regularization

$$\min_{f \in \mathcal{F}} \underbrace{\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))}_{\text{经验风险 (loss function)}} + \underbrace{\lambda R(f)}_{\text{penalty}} \quad \lambda \geq 0 \quad (1)$$

其中, λ 调节经验风险与正则化项之间的关系

$R(f)$ depends on your problems, 常用的有 $l_0, l_1, l_2, l_p, l_\infty$

M.L. Theory

Rules of choosing model

Cross Validation

- 思想
重复利用数据
- 方法
将给定数据切分组合为训练集与测试集，在此基础上反复训练、测试以及模型选择
- 分类
简单交叉验证、S折法、留一法

Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

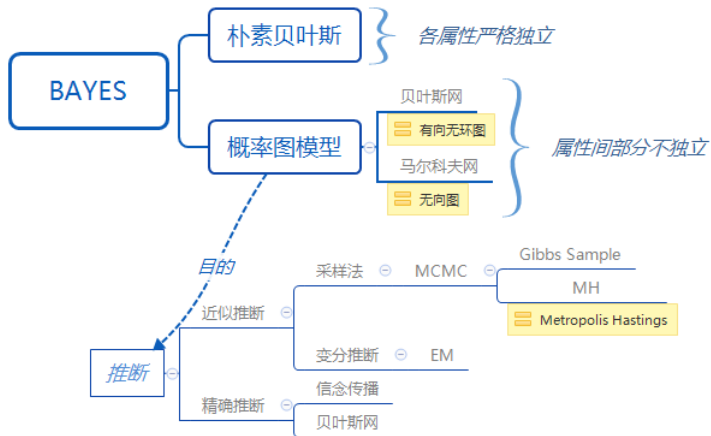
Some issues

Summary



Some M.L. algorithms

Bayesian Framework



- 近似推断: 避免精确推断的高计算需求

Some M.L. algorithms

Bayesian Framework

- 采样法

多采用马尔科夫链蒙特卡洛采样。当采样足够多时，认为可以近似表达分布，目标是直接通过采样计算期望

- 变分推断

使用已知简单分布逼近需推断的复杂分布，并通过限制近似分布的类型，得到一局部最优但有确定解的近似后验分布。

e.g. 高斯分布

- EM 算法

期望最大化算法。原理为 MAP（最大化后验概率）。分 E 步和 M 步，E 步求似然期望，M 步寻找能使 E 步产生的似然期望最大化的参数值。即，最大似然（ML）

Some M.L.algorithms

Linearity Regression

已知输入数据 \mathbf{X} , 输出数据 \mathbf{Y} , 求解 θ

$$f_{\theta}(\mathbf{X}) = \theta^T \mathbf{X} \quad (2)$$

$$\arg \min_{\theta} J_{\theta}(\mathbf{X}) + \lambda R(\theta) \quad (3)$$

$$J_{\theta}(\mathbf{X}) = \sum_{i=1}^N (f_{\theta}(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)})^2 \quad (4)$$

$$(4) + \begin{cases} 0, & \text{if } \lambda = 0 \Rightarrow \text{OLS} \\ \lambda \|\theta\|_1, & \text{if } \lambda > 0 \Rightarrow \text{LASSO} \\ \lambda \|\theta\|_2^2, & \text{if } \lambda > 0 \Rightarrow \text{Ridge} \end{cases}$$

Some M.L.algorithms

Linear Regression


(4) 根据 loss function 定, 平方 loss+ 问题 overdetermined

⇒ 最小二乘

通常, l_0, l_1 比 l_2 稀疏性好, 所得模型泛化能力更强, 但是 l_0, l_1 优化求解不如 l_2 简单

Some M.L.algorithms

Others:Many!!

- 核技巧（**kernel tricks**）以及相关的支持向量机（**SVM**）、高斯过程回归（**GPR**）等算法
 - 神经网络（**ANN**）和 **D.L.** e.g.: **CNN**、**RNN**
 - 降维算法，e.g. 主成分分析（**PCA**）
 - and so on and so forth
- 

Outline

Basics of M.L.

The relationship of M.L.& D.A.

Some issues

Summary



Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

Some issues

Summary



Why talking about this?

It's not nonsense

May I could say the Theory of D.A. comes from **control theory**, using **PDE, ODE** to make a perfect description of the natural phenomena, although it's **impossible**.

The natural phenomena is too complicated to describe, we need to **simplify** it again and again. Afterwards, we get our model, within plenty of **biases, errors**.

M.L. focuses on the **general model**. You just need to spend more time on the **mapping** between the input data and output data, without much consideration about the physics, biology etc. And it did works.

Why talking about this?

It's not nonsense

Then why didn't us try to keep an eye on the more general model without considering many physics process?

In fact, we are doing it. The **statistics** in D.A. supports me. If you look at Particle Filter, Hierarchical Bayesian network, you will find more.

We are **combining D.A. and M.L.**

But we need to be better.

Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

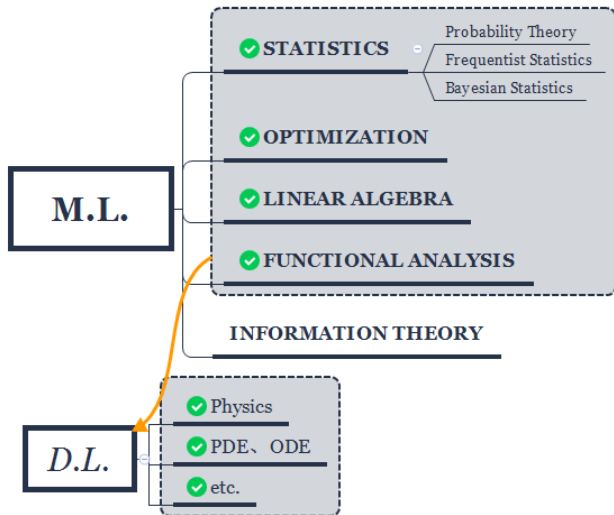
Some issues

Summary



Mathematical Basis

you may need to learn



Outline

Basics of M.L.

What's M.L.

M.L. Theory

Some M.L.algorithms

The relationship of M.L.& D.A.

Why talking about this?

Mathematical Basis

Methods

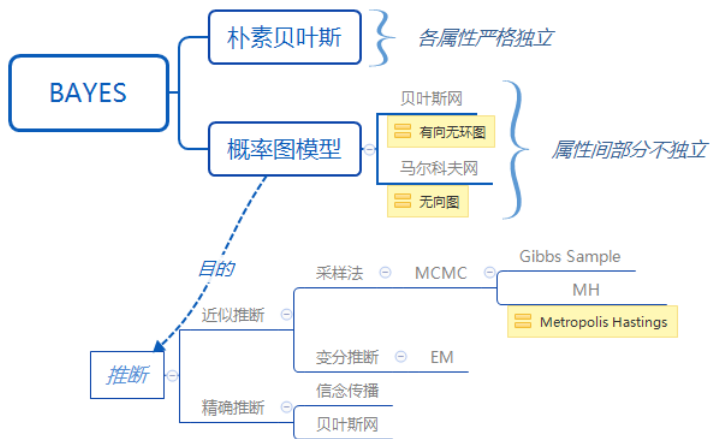
Some issues

Summary



Methods

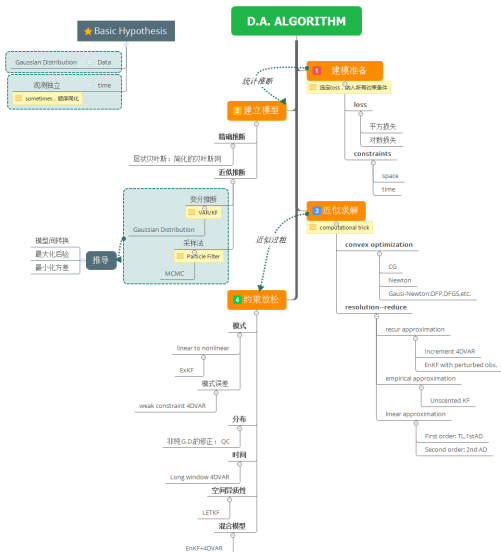
Let's go back firstly



According to the figure above, I'd like to make some extensions.

Methods

D.A. Framework, from the perspective of M.L.



Just for an instance, as for I'm a green hand in D.A.& M.L., you may find many errors in it

Methods

change the perspective

from 邹晓蕾, 资料同化与应用 page43

3DVAR、4DVAR、EnKF, 分析增量均为观测增量的加权平均。不同的 D.A. 方法只是获得的后验权重 (W) 估计不同, 所有的推导都是下式的直接或间接的表达

$$x^a = x^b + W(y^o - y) \quad (5)$$

参考线性回归, to be **more general**, 可以增加正则项

Methods

Tikhonov regularization

from John M. Lewis *Dynamic D.A.* page 115

an unified model which solve both the under-determined and over-determined problems

$$f(x) = \frac{1}{2} \|(z - Hx)\|^2 + \frac{\alpha}{2} x^T x \quad (6)$$

What if changing the norm to ℓ_0, ℓ_1 norm, will it make sense?

A.K.A. Will the ℓ_0, ℓ_1 norm, which is popularly used in M.L., help us get a more general model in D.A. ??

Outline

Basics of M.L.

The relationship of M.L.& D.A.

Some issues

Summary



Some issues

Some problems in D.A.

Problems in D.A.

- 计算需求
- 模式误差
- 观测误差（数据集）
- 观测独立假设（实际不全为真）
- 预报性



Some issues

Just for an example

You may get some inspirations from M.L.

- 计算需求 \Leftarrow 降维: kernel trick, sparse matrix reconstruction
- 模式误差 \Leftarrow D.L. 是否可以与模式结合?? LSTM 擅长分析时间序列
- 观测误差 (数据集) \Leftarrow D.M. 中的分箱、回归、离群点检测 etc.
- 观测独立假设 (实际不全为真) \Leftarrow 随机过程取代随机变量: GPR? LSTM?
- 预报性 (a.k.a. 泛化性能) \Leftarrow 正则化: norm selection

Some issues

You may be interested in

Data need huge storage space.

Hadoop and Spark works well in the D.M. of big data

- ① **Hadoop**: Apache 的项目，分布式系统架构，开源。下设 HDFS、HBase、map/reduce 等子项目。
 - HDFS 实现对于大数据的分布式文件存储。
 - HBase 不同于 Mysql、SQLSERVER、Oracle 等关系型数据库，是 No-sql 数据库。擅长文件数据存储检索。常用的 no-sql 数据库还有 mongoDB。
 - map/reduce 为数据并行计算框架，用于作业调度。map 步将大作业分散为小作业，reduce 步将小作业的输出结果组织为最终结果。

Some issues

You may be interested in

- ① **Spark:** Apache 项目，由 UC Berkeley AMP lab 开发的类 map/reduce 的通用并行框架，开源。
M.L. 的可行框架之一，计算速率优于 map/reduce。
可以在 HDFS 中并行运行，且提供 M.L. 的 API。也可单独运行。
- ② **Ray:** 由 UC Berkeley RISE Lab 开发，尚在开发中。旨在让基于 Python 的 M.L. 和 D.L. 工作负载能够实时执行，并具有类似消息传递接口（MPI）的性能和细粒度。据 Michael I.Jordan 说，Ray 性能优于 Spark，将取代 Spark。

B.T.W. Michael I.Jordan 曾先后担任 AMP、RISE 的顾问，Spark 和 Ray 都有做指导。研究方向为统计学习

Outline

Basics of M.L.

The relationship of M.L.& D.A.

Some issues

Summary



Summary

Outline again

relationship

Algorithm:

- 1 Bayesian network
- 2 linear regression

May be useful:

- 1 dimension reduce and sparse matrix reconstruction: kernel tricks, PCA
- 2 computation efficiency: D.L., Ray
- 3 storage space: HDFS

Summary

The combination of D.A.& M.L. is **unavoidable**

