Author：Liu Jian

Time：2021-01-26

机器学习的基本概念

- 统计机器学习，基于数据统计而非规则或逻辑推理，不讲**规则** (道理) 只讲 **规律** (统计) (NLP 举例)，统计与逻辑的结合是一个重要的发展方向 [NatureReview，知识图谱]；
- 使用模型 (SVM、NN等) 对数据进行拟合、分类 ，目的是得到泛化能力强的模型：
  - 模型 + 数据；
  - **拟合标准：损失函数**，**无监督学习**也会存在一个损失函数，这个损失函数表征了人们对无标记数据的内在结构的认知；
  - 泛化能力：对未知数据的预测能力，过拟合与欠拟合都是泛化能力弱的表现：
    - 过拟合：捕捉到了训练数据中的非一般性规律，比如，**认为白色人种才是人，黑人的命就不是命了**；
    - 欠拟合：没有捕捉到训练数据中蕴含的一般性规律；
    - **画图，拟合曲线依次通过被拟合的点，并不表示一定过拟合**，也有可能是数据量不足，看到这种情况严格的说法应该是，相比于简单且较好地穿过数据区域的拟合曲线是欠拟合，是其发生过拟合的概率更高，**要减少这个概率，增大数据量。**
    - **奥卡姆剃刀原则**：训练误差相等的情况下，选更简单的模型。因为更简单模型泛化能力高的**概率**更大 (更简单的模型不一定一定更好，对于这种不确定的事，就可以使用概率来描述，比如依概率收敛，大数定理，PAC可学习)，奥卡姆剃刀原则在 PAC 可学习理论的框架下是可以被证明的；
- 流程：
  - **特征工程**，确定输入特征与输出；
  - **确定模型和损失函数**：
    - 和人一样，不同模型的学习能力，也就是表达能力也不同 (高阶多项式模型>低阶多项式模型)；模型不能过于简单 (欠拟合)，也不能过于复杂 (过拟合)；
    - 损失函数：均方误差、绝对误差、对数似然损失 ($-\ln p(y|x,\theta)$，模型 $p(y|x,\theta)$ 没有直接给出输出，但由贝叶斯决策论，$y=argmin_y p(y|x,\theta)$)；
  - 最小化损失函数 (**+ 正则化项**)，对未知参数进行估计；
  - **模型评估（交叉验证)**：将学得的模型作用于新的数据集 (验证集/测试集) 上，评估模型的泛化能力，作为模型性能的最终指标。
  - 若模型表现不好，换一类模型或同类模型再进行调整，再按上面流程进行一遍；参数估计--**参数**，模型评估--**超参数** (无法使用求导进行优化的参数)。

- 强化学习：学习一个策略，延迟标记的监督学习；游戏比赛，AlphaGo等；

最小二乘法

- 最小化均方误差：$\min(y-f(x;\theta))^2$
- **L1/L2 正则化及其解释**：
  - L1 正则化：$\min_\theta(y-f(x;\theta))^2+\alpha\sum_i|\theta_i|$
  - L2 正则化：$\min_\theta(y-f(x;\theta))^2+\alpha\|\theta\|^2$，**比如：SVM，画图解释**

极大似然估计

- $y=f(x;\theta)+\varepsilon$，通过引入随机误差$\varepsilon$ (服从某个假定的概率分布)，将拟合套上概率的外衣；实际上就是给出了**似然函数** $p(y|x,\theta)$，将真实数据代入到模型中，**若模型使得真实数据出现的概率越大，说明根据模型作预测时，模型越准，越贴合真实分布，模型也就越好**，预测时 $y=argmin_y p(y|x,\theta)$ 。
- 若概率模型使用高斯分布，$\varepsilon\sim\mathcal{N}(0,\sigma^2)$，则

$$p(y|x,\theta)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-f(x;\theta))^2}{2\sigma^2}\right)$$

取负对数损失：

$$-\ln p(y|x,\theta)=\frac{(y-f(x;\theta))^2}{2\sigma^2}+C$$

其中，C 为常数。可见，**最小化均方误差等价于对高斯模型进行极大似然估计。**

- 若概率模型使用拉普拉斯分布，$\varepsilon\sim\mathcal{L}aplace(0,\sigma)$，$\sigma>0$：

$$p(y|x,\theta)=\frac{1}{2\sigma}\exp\left(-\frac{|y-f(x;\theta)|}{\sigma}\right)$$

取负对数损失：

$$-\ln p(y|x,\theta)=\frac{|y-f(x;\theta)|}{\sigma}+C$$

其中，C 为常数。可见，**最小化绝对误差等价于对拉普拉斯模型进行极大似然估计。**

最大后验估计

已知数据后，最大化参数的后验概率，贝叶斯公式：

$$p(A,B)=p(A|B)P(B)\Rightarrow p(A|B)=\frac{p(A,B)}{P(B)}$$

$$p(\theta|y)=\frac{p(y|\theta)p(\theta)}{p(y)}\propto p(y|\theta)p(\theta)$$

举例：猜测一个人是男是女，无任何信息前，我们猜 0.5 的概率是男，0.5 的概率是女 (先验概率 $p(\theta)$)；但若我们知道此人身高180cm，体重75kg后，我们猜 0.9 的概率是男，0.1 的概率是女 (后验概率 $p(\theta|y)$)，引入数据信息，对先验概率进行修正，得到后验概率。其中，

- $p(y|\theta)$ 似然函数，概率函数
- $p(\theta)$ 先验概率
- $p(\theta|y)$ 后验概率
- $p(y)$ 归一化因子，与 θ 无关，最大化后验概率时可不考虑：

$$\min_{\theta} p(\theta|y) \Leftrightarrow \min_{\theta} p(y|\theta)p(\theta)$$

先验分布$p(\theta)$：

- 均匀分布，等价于极大似然估计；
- 拉普拉斯分布，Ridge，等价于L1正则化
- 高斯分布，LASSO，等价于L2正则化

神经网络

- 感知机：最简单的神经网络，无法解决异或 (XOR) 问题 (明斯基，神经网络的寒冬)；
- 前馈神经网络：
  - 线性连接+激活函数；
  - BP 算法训练（**链式求导法则+梯度下降算法，梯度下降算法的形象化解释**）；
  - 通用近似定理；
  - 长程依赖问题：
    - 梯度消失：
      - Sigmoid 型函数的导数的值域都小于或等于1，误差经过每一层传递都会不断衰减．当网络层数很深时，梯度就会不停衰减，甚至消失，使得整个网络很难训练
    - 梯度爆炸：
      - 会造成系统训练不稳定
- CNN：
  - 卷积层 + 池化层；
  - 卷积层：**局部连接** (减少网络中连接的数量，神经元个数并没有显著减少)，**权重共享**；
  - 池化层：**降低特征维数，避免过拟合** (汇聚层也可以看作一个特殊的卷积层，卷积核为max 函数或mean 函数)；
  - 计算机视觉；
- RNN：处理时序问题，比如用于自然语言处理 (NLP) 的 LSTM。
- 借用了很多神经科学的名词，让人听起来很玄乎，但从数学的角度，神经网络模型是很简单的：
  - 线性函数+非线性激活函数的不断**复合**；
  - **定义好损失函数和结构后，使用 BP 算法优化参数即可**；
  - **神经网络的结构是超参数，需要人工设计调参**；
  - **神经网络的应用**：强大的拟合器；
  - 对神经网络能力的解释依然有待研究；

敏感性分析

一个量(输入)的变化对另一个量(输出)变化影响程度的量化分析：

- 局部敏感性：导数
- 全局敏感性：多点的导数 (的绝对值) 取平均

基于方差的全局敏感性分析：

- **方差表征变量的变化剧烈程度，因此，可以将输出的总方差分解为各输入的贡献，以此衡量各输入对输出的影响**；
- Sobol' decomposition

$$M(x_1,x_2)=x_1^2+2x_2^2-2x_1x_2+x_1+1$$

$$M_0=\int M(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2=2$$

$$M_1(x_1)=x_1^2-\frac{1}{3}$$

$$M_2(x_2)=2x_2^2-x_2-\frac{1}{6}$$

$$M_{12}(x_1,x_2)=-2x_1x_2+x_1+x_2-\frac{1}{2}$$

进一步，可进行方差分解，定义Sobol'指数。

注意：

- **交叉项中可以含有一阶项和常数项**；
- **从数学的角度上来看**，$M_{12}+M_1$ 也是关于 x_1,x_2 的函数，为何不说 $M_{12}+M_1$ 表征了交互作用而只是 M_{12} 表征了交互作用 (**凭什么这样的定义就是对的，在数学上如何解释？**)?
  - **Sobol' 分解是对原函数最彻底的分解**
- 非 $[0,1]$ 上的均匀分布如何处理？ 变量替换后的变量的敏感性依然可以表征替换前变量的敏感性吗，**变换后的我还是我嘛**?
  - **Hoeffding–Sobol' decomposition**

Polynomial chaos expasion

- PCE：多项式混沌展开，**混沌--随机，不是指乱展开，而是输入是随机变量**；
- **画表格解释下 PCE**；
- 可视为对 Sobol' decomposition 的进一步分解，类似于**分子和原子之间的关系**；
- PCE 系数计算方法的分类：
  - **嵌入式方法。**这类方法将多项式混沌展开代入到待研究模型的控制方程中，利用多项式基函数的正交性构建一个以多项式系数为未知数的新的控制方程并予以求解，通用性较低；
  - **非嵌入式方法**：
    - 投影法：采样计算积分，进而计算系数；
    - 回归法：我们所采用的方法；
- PCE 构建算法：
  - 模型选择：使用哪些基函数进行拟合
    - KIC (AIC、BIC)；
    - KIC 的计算 -- **拉普拉斯近似**；
  - 参数估计：计算 PCE 的系数
    - **相比于 LASSO，只是多出来一个 σ_ε 需要估计**；
    - C_0 设置的理解，对比 L2 正则化；
  - 构建算法：核心点为两次排序及基函数的扩充策略
    - 也只是一个**局部最优搜索策略，贪心的**，不能考虑所有的可能性，当然也就不能保证最优；
    - **线性相关并不表示因果关系，只是一种统计关系**；
  - **伪相关性举例：收入、衣服消费、食物消费**；
- 思考：
  - 使用其他的模型评估指标也可以；
  - 不使用模型评估指标，直接进行交叉验证也可以；

    - 集成学习：Bayesian Model Averaging，好而不同
        - **弱学习器--> 强学习器**，三个臭皮匠顶个诸葛亮；**基于 PAC 学习理论可以被证明**；
        - **头脑风暴，取长补短，胡老师和大家一起讨论，再弱鸡的新人也有发言权，也会听取意见，只不过意见的权重不同**，中央政治局常委讨论决定 (**基于委员会的学习**)；
    - 基函数扩充策略的改进
    - **PCA + ICA + GPCE**：
        - PCA：降维去噪，奇异值分解/**总**最小二乘 (**画图说明**)；
        - ICA：**鸡尾酒会问题**，无监督学习，**最大化非高斯性/最小化互信息**，Sudret 做过！！
            - 信息熵解释内卷：封闭系统熵增，熵增就意味着混乱和无序，所以需要和外界交流；**课题组需要开拓新题目**，**帝国主义要向外扩张**；
        - **构建起一个完备的框架，胡老师项目可以用**！

再回首

- 两类问题：
    - 方程的求解：见下文牛顿法与ANM；
    - 优化问题：**带约束的优化问题：拉格朗日乘数法，KKT 条件 (大号的拉格朗日乘数法)；算法：SGD，Adam等**
- 合适的基函数 (暴露了问题的特征) + 近似思想 (无法解析求解，就只能近似)：
    - 大问题 --> 小问题
    - 局部近似：泰勒展开 (求 KIC 时的**拉普拉斯近似**)；
    - 宏观近似：傅里叶展开；
- **桥域多尺度方法也是类似的优化问题**
- PCE：
    - **不管自变量服从何种分布，被展开的基函数是可以任意选取的，但选择合适的基函数进行展开才有助于解决问题**
    - **广义的多维傅里叶展开，求 PCE 投影法就是高等数学教科书中求傅里叶展开系数的方法**；
    - **卷积神经网络 (编码-->解码)，BPCE 的无线逼近 -- 通用近似定理**；
- **群哥的傅里叶展开**，利用了傅里叶基函数的宏观近似，不严格是包络线；
- **小波分析**：衰减的信号；
- **杰哥的无网格方法/瑞利里兹法**，在域内以已满足条件的基函数构建问题的解，只需考虑边界；精度不满足时，划分网格单元；
- **有限元法**：
    - **微分方程** --> 伽辽金加权转化为**积分形式 (弱形式)**
    - **划分网格 -- 数学上来说就是积分的分段可加性**
    - 插值函数 -- 用于暴露特征的近似基函数
        - **胡老师要我做过一个问题，三阶插值函数无法表征力矩**
- 选择合适的观测角度 (输入特征，基函数) 能降低问题的难度，即特征工程：
    - **二圆分类问题**：降维、升维、极坐标映射举例；
    - 维度过高 （**高维立方体内嵌球体的体积趋于0**）-- 维度灾难 (需要数据量大)；维度过低，难以求解
    - 核方法 (kernal trick)；
    - 流形学习：天地有正气、杂然赋流形；**高维空间中的低维嵌入，局部空间与欧式空间同胚**；
        - 举例：瑞士卷、地球
- 神经网络 -- 表示学习；NatureReview；

牛顿法和ANM

- ANM：**一个中心 (合并同类项)，两个基本点 (对原方程的二阶近似，对方程解的泰勒展开近似)**，是一种嵌入式的求解方法；
- 牛顿法，是一种非嵌入式的求解方法，迭代公式：

$$x_{k+1} = x_k - \frac{f(x)}{f'(x)}$$

- 牛顿法通过迭代收敛得到原方程的解，**理论上是不需要预测步的，可以任意给定迭代收敛点只要最终收敛即可**，若在迭代时总是置 λ 为某个固定的值，那就得到了**固定步长的牛顿法**；
- ANM求解的并不是原方程，而是原方程的近似方程，所以**ANM更容易求解就很好理解了，求解的只是原问题的简化**；若非二次型，就算取无穷阶，也得不到精确解；
- 摄动法引入了 $U \sim a$ 和 $\lambda \sim a$ 的间接关系 (直接关系是 $U \sim \lambda$)，**因此需要有补充方程**；对于不同的问题，取不同的补充方程 (**坐标系**) 收敛性当然会不同；因为是间接关系，**这也就解释了为何 U, λ 对 a 的泰勒展开不唯一 (问过胡老师)**；
- 刘老师论文上写的有误，牛顿法和ANM **并不是特殊和一般的关系**，二者侧重点压根不同，**ANM可以为牛顿法提供好的初始迭代点**；
- 最后，个人认为 ANM 推导稍微麻烦影响了其使用，除了方法的性能，方法使用简单才能得到推广，比如神经网络。

力学中的数据驱动

- Ortiz 的方法实际上**等价于**带约束的优化问题 (最小化势能)，约束 (本构关系) 以离散的数据点的形式给出，并使用拉个朗日乘数法求解：

$$\begin{gathered}
\min \text{最小势能} \\
s.t. \text{满足本构关系}(\text{以离散的数据点的形式给出}) \\
\Updownarrow \\
\min \text{离散数据点距离} \\
s.t. \text{平衡方程满足} \\
(\text{ps 或许存在一个更高层次的优化目标，上述问题只是其使用 EM 算法求解的结果})
\end{gathered}$$

  - 改进：
    - 基于 kd 树优化目前的算法 (来源于k近邻算法，构造kd数，迭代搜索)；
    - **噪声**的影响，**欧式距离改为曼哈顿距离**；
  - 缺点：
    - 陷入局部最优；问题复杂时，可能敏感不收敛；
    - 查找效率低，**理论上可以直接暴力破解，一个个数据点代进去试一试**；
    - **数据本来就有噪声，找到的不一定是最好的，这也是机器学习的出发点，使用模型拟合去噪**；
    - **无论是流形学习，还是分层聚类，做到最后就是拟合本构关系！**
  - **类比举例：求解方程 x-2=0，直接使用数据点反而把问题搞复杂了**
  - 本质上没有跳出传统有限元的基本理论和框架 (势能最小+本构关系)，只是求解方法从**嵌入式**变为**非嵌入式**！
- **群哥给看的使用深度学习求解有限元的论文**：
  - 输入：高斯点坐标，输出：位移场的值，使用 NN 进行拟合；
  - 最小化势能 (NN 的损失函数，本构关系已知且直接代入进去)，本质上还是没有跳出有限元的框架；
  - 对比传统有限元：

        - 区别在于**位移场的拟合函数不同**，传统 -- 多项式分段函数，这里 -- 一个神经网络全部搞定；
        - 求解的方式：传统的是**嵌入式**的方法，这里是**非嵌入式的**，**不需要将位移场模型代入控制方程进行繁杂的推导**！
    - 可能存在的问题：需要大量数据，过拟合，优化求解复杂，嵌入局部最优，因此，简简单单的分段插值函数有时可能更香。
- **综上所述，个人总结如下**：
    - 传统有限元是一种**嵌入式 (代入本构关系，代入位移场拟合函数到控制方程中，推导化简求解)** 的求解方法；
    - Ortiz 的方法和用深度学习解有限元实际上都是一种**非嵌入式的求解方法：一个是没有代入本构，一个是没有代入位移场，使用数值迭代优化求解**；
    - Ortiz 直接使用数据点的做法比较原始和低效：
        - 对数据进行分层、或kd树进行组织做到最后就是对数据进行拟合，得到拟合的本构关系，这样才能高效利用数据信息，这也是机器学习的出发点；
        - Ortiz 相当于直接保留已有的数据点作为本构关系，这也可以看做是用机器学习得到的一个模型，只不过这个模型很“弱”；
        - 当然不排除这个很“弱”的本构模型在某些问题上表现得异常好，但这只是特例，无关紧要，并不是发展方向.
    - **个人认为的终极框架**：
        - 传统有限元：**嵌入式方法**；
        - 近年来引入数据 + 机器学习的方法：**非嵌入式方法**
            - 使用机器学习拟合一个本构，将本构作为一个约束条件，使用非嵌入式的方法求解势能最小的约束优化问题，个人认为这是 Ortiz 方法的终极形态；
                - 当然，也有拟合出一个本构后，将本构方程又回代到控制方程中的，这就又回到了传统有限元嵌入式的方法，又需要繁杂的推导，因此不作推荐；
            - 使用 NN (或许也可以用其他机器学习模型) 拟合位移场，使用梯度下降算法反向传播最小化势能；
            - 一种可以的尝试，或许可以称之为**全非嵌入式方法**，即上述二者的结合：本构和位移场均拟合得出，使用非嵌入式的方法求解。
- 在大家各自的领域如何引入机器学习：
    - 拟合器：比如非常火的 GNN 使用神经网络拟合了两个函数：生成器和判别器。