CSMath 课程学习笔记

王津航　11521030

1. 为什么是数据驱动

当今的计算机科学面临的最大挑战:第一是大数据,大公司正在搜集大数据(Google、Apple、 Facebook、IBN、Microsoft 正在做这样的事情);第二仍然是数据,我们长期致力于创造数据世界、数据生活,例如我们可以寻找新的方式来通信并且创造新的通信媒介;另外专家变得异常昂贵：科学家、工程师,电影制作者均是如此。第三是处理现存的数据并且从现存的数据中创造新的数据。可见是数据驱动计算机科学。

纯粹的程序合成 vs 纯数据

情形 1:创造电影中的角色移动

- 纯程序合成:显得很简单、但是很明显看得出来是仿造的,因此很少用于实际中
- 手工或者纯数据:高质量但是很低的灵活性
- 最好的方法：两者相结合

贝叶斯推理(Bayesian Reasoning)

- 不确定性的原理模型
- 非结构化数据的通用模型
- 数据拟合以及不确定分析的有效方法
- 现在贝叶斯推理通常用作黑盒式的方法

数据驱动词汇: 数据(data):数据驱动、数据挖掘;学习(learning):机器学习,数据学习;不确定性(uncertainty): 概率(probability)、似然性(likelihood);智能(intelligent):推断(inference)、决策(decision)、检测 (detection)、识别(decision)

数据驱动系统: 学习系统(learning system)并不是直接编程解决一个问题,相反是基于他们应有行为的实例, 解决问题的过程中的试错经历来开发自己的程序。其不同于传统的计算机科学,它只有通过 提供的样本输入输出对来实现未知的函数。

学习问题的主要分类(main categories): 根据训练实例中可提供的信息不同造成学习场景的不同

- 受监督的(supervised):能够改正输出:分类(classification),1-of-N 输出(语音识别、对象识别、医疗诊断);回归(regression),实数值输出(预测市场的价格、温度)
- 半受监督(semi-supervised):只有部分输出可用,排名(Ranking)
- 不受监督的(unsupervised):没有反馈,需要构造良好输出的方法:聚类(clustering),聚 类指的是把分割的数据划分成一致的数据集;异常检查(Novelty-detection),检查偏离正常数据集的新的数据点
- 强化(reinforcement):数量反馈、可能时间延迟

为什么使用数据驱动的方法

- 开发加强的计算机系统以自动地适用于用户,为用户定制,能够在线下大数据中发现模式
- 改善人类、生物学习由于计算分析能够提供强大的理论、预测。
- 时效性良好,由于可用的数据在不断增加、存在大量廉价以及强大的计算机、已研发出的理论和算法套件的支持

成功的数据驱动(data-driven)算法的特性:

- 计算高效(computationalefficiency)
- 鲁棒性(Robustness)
- 统计稳定(statisticalstability)

2. 点估计(point estimation) 本节要点:

- 点估计(pointestimation)
  - 最大似然估计(Maximal Likelihood Estimation, MLE)
  - 贝叶斯学习(Bayesian Learning)
  - 最大化后验(Maximize A posterior, MAP)
- 高斯估计(Gaussian estimation)
- 回归(Regression)

- 基函数 = 特征
- 优化和方差(optimizing sum squared error)
- 回归和高斯之间的关系

- 偏方差权衡(Bias-Variance trade-off)

PAC(Probably Approximate Correct) learning:概率近似正确学习 先验:实验之前的知识(prior: knowledge before experiments) 不再是估算一个单一值 θ,我们可以得到一个概率值 θ 的分布。

贝叶斯学习(Bayesian learning)

贝叶斯准则(Bayes rule): $Posterior \rightarrow P(\theta|D) = \frac{P(\theta)P(D|\theta)}{p(D)\leftarrow Datadistribution}$

或等价地表示为:$P(\theta|D) \propto P(\theta)P(D|\theta)$

即 posterior ∝ likelihood ∗ prior 我们例子中的贝叶斯学习(Bayesian Learning in our case):

二向分布中的似然函数(Likelihood function is simply Binomial):

$$P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

先验函数代表的是专家知识,是简单的后验形式。共轭先验(conjugate priors)是封闭式的后验 表达,对于二项分布,共轭先验就是 Beta 分布

Beta 先验分布(Beta prior distribution – $P(\theta)$)

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \sim Beta(\theta|\beta_H,\beta_T) = \frac{\tau(\beta)}{\tau(\beta_H)\tau(\beta_T)}\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}$$

$$\tau(x) = x\tau(x), \tau(1) = 1$$

似然性(二项分布):

$$P(D|\theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

则后验(posterior):

$$P(\theta|D) \propto P(\theta)P(D|\theta) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\theta^{\beta_H-1}(1-\theta)^{\beta_T-1} \sim Beta(\alpha_H+\beta_H,\alpha_T+\beta_T)$$

MAP(Maximum a posteriori approximation)最大后验相似:

$$P(\theta|D) \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

$$E[f(\theta)] = \int_0^1 f(\theta) P(\theta|D) d_\theta$$

MAP 的思想:使用可能性最大的参数使得后验最大,即

$$\theta' = \arg max_\theta P(\theta|D)\ E[f(\theta)] \approx f(\theta')$$

对于 Beta 分布的 MAP:

$$P(\theta|D) \sim \text{Beta}(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

$$\theta' = \arg\max_\theta P(\theta|D) E[f(\theta)] = \frac{\alpha_T + \beta_T - 1}{\alpha_T + \beta_T + \alpha_H + \beta_H - 2}$$

当 $N = \alpha_T + \alpha_H \to \infty$时,先验就被"忽略"了,但是对于小样本量,先验是至关重要的 似然性除了二项分布还有:

多项式分布(Multinomial distribution):

共轭先验(狄利克雷分布)[Conjugate prior(Dirichlet distribution)]:

$$P(\theta) = \text{Dir}(\theta|\alpha_1, \ldots, \alpha_r) = \frac{\tau(\alpha)}{\prod_{k=1}^r \tau(\alpha_k)} \prod_{k=1}^r \prod {\theta_k}^{\alpha_k - 1} \ , \alpha = \sum_{k=1}^r \alpha_k$$

高斯分布(Gaussian distribution)

连续的随机变量情况下的高斯分布:

$$P(x|\mu, \delta) \sim \frac{1}{\delta\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\delta^2}}$$

其中δ表示方差, μ 表示均值

高斯分布的最大似然估计(MLE for Gaussian)

对于i.i.d. (Independent Identically distributed)独立一致性分布的示例 $D=\{x_1, x_2, \ldots, x_N\}$

似然性:

$$P(D|\mu, \sigma) = \left(\frac{1}{\delta\sqrt{2\pi}}\right)^N \prod_{I=1}^{N} e^{\frac{-(x-\mu)^2}{2\delta^2}}$$

MLE for Gaussian 所得到的结果:

$$\mu = \frac{1}{N}\sum_i x_i,\ \sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$

$$\mu' = \frac{1}{N}\sum_i x_i,\ \sigma'^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$

回归问题(The regression problem)描述:

实例:$<\boldsymbol{x}_i, t_i>$

学习:从 x 到 t(x)映射

- 假设空间(Hypothesis space):$t(x) \approx f'(x) = \sum_{i=1}^{k} w_i h_i$
- 给定基函数(basis function): $H = \{h_1, \ldots, h_k\}$

寻找因子$\mathbf{w}=\{w_1, \ldots, w_k\}$

问题公式化:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \sum [t(x_j) - \sum_i^k w_i h_i(x)]^2$$

曲线拟合:

和方差函数(Sum-of-Squares Error Function) $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$

方均根误差(Root-Mean-Square, RMS) Error $E_{RMS} = \sqrt{2E(w^*)/N}$

惩罚大系数值(penalize large coefficient values)

$$E'(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$