

The-Elements-of-Statistical-Learning 学习笔记

李奥林

mr liaolin@outlook.com

目录

第一章 Model Inference and Averaging	1	1.5 The EM Algorithm	2
1.1 Introduction	1	1.5.1 Two-Component Mix- ture Model	2
1.2 The Bootstrap and Maximum Likelihood Methods	1	1.5.2 The EM Algorithm in General	2
1.2.1 A Smoothing Example .	1	1.5.3 EM as a Maximization- Maximization Procedure	3
1.2.2 Maximum Likelihood Inference	1	1.6 MCMC for Sampling from the Posterior	3
1.2.3 Bootstrap versus Maxi- mum Likelihood	1	1.7 Bagging(装袋)	3
1.3 Bayesian Methods	1	1.7.1 Example: Trees with Simulated Data	3
1.4 Relationship Between the Bootstrap and Bayesian Inference	2		

第一章

Model Inference and Averaging

本章为《The elements of Statistical Learning》第 8 章的笔记。

1.1 Introduction

1.2 The Bootstrap and Maximum Likelihood Methods

1.2.1 A Smoothing Example

自助法提供了一种评估不确定性的直接计算方法（置信区间）。

- 非参数自助法（与最小二乘法的置信区间类似）
- 参数自助法（对每个预测的 y 值加一个高斯噪声，参数为噪声的方差，此时估计出的函数的置信区间与最小二乘法的完全相同）

1.2.2 Maximum Likelihood Inference

参数自助法与最小二乘法是一致的，因为模型具有加法高斯误差。一般地，参数自助法并非与最小二乘法一致，而是与极大似然一致。

1.2.3 Bootstrap versus Maximum Likelihood

1.3 Bayesian Methods

在用于推理的贝叶斯方法中， $\Pr(\mathbf{Z}|\theta)$ 是采样模型，先验分布是 $\Pr(\theta)$ ，反映我们看到数据之前的关于 θ 的知识，后验分布为 $\Pr(\theta|\mathbf{Z})$ 是我们看到数据之后关于 θ 更新的知识。

与标准的“频率论”方法的区别是使用先验分布来表达看到数据之前的不确定性，并在看到数据之后允许残余的不确定性以后验分布形式来表示。

1.4 Relationship Between the Bootstrap and Bayesian Inference

自助法分布为我们的参数提供了一个（近似的）非参数的、无信息的后验分布。

1.5 The EM Algorithm

1.5.1 Two-Component Mixture Model

1.5.2 The EM Algorithm in General

公式 $E(\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)})$ 的定义是

$$E(\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)}) = \sum_{\mathbf{Z}^m} \ell_0(\theta'; \mathbf{T}) \Pr(\mathbf{Z}^m|\mathbf{Z}, \hat{\theta}^{(j)}) \quad (1.1)$$

根据条件概率的链式法则

$$\Pr(x_2, \dots, x_n|x_1) = \prod_{i=2}^n \Pr(x_i|x_1, \dots, x_{i-1}) \quad (1.2)$$

得 $\Pr(\mathbf{Z}^m, \mathbf{Z}|\theta') = \Pr(\mathbf{Z}^m|\theta') \Pr(\mathbf{Z}|\mathbf{Z}^m, \theta')$, 即

$$\Pr(\mathbf{Z}|\theta') = \frac{\Pr(\mathbf{Z}^m, \mathbf{Z}|\theta')}{\Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')} \quad (1.3)$$

用对数似然函数表示, $\ell(\theta'; \mathbf{Z}) = \ell_0(\theta'; \mathbf{Z}^m, \mathbf{Z}) - \ell_1(\theta'; \mathbf{Z}^m|\mathbf{Z})$, 其中 ℓ_1 基于条件概率密度 $\Pr(\mathbf{Z}^m|\mathbf{Z}, \theta')$ 。关于参数 θ 支配的 $\mathbf{T}|\mathbf{Z}$ 取条件期望, 得

$$\begin{aligned} \ell(\theta'; \mathbf{Z}) &= E[\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \theta] - E[\ell_1(\theta'; \mathbf{Z}^m|\mathbf{Z})|\mathbf{Z}, \theta] \\ &\equiv Q(\theta', \theta) - R(\theta', \theta) \end{aligned} \quad (1.4)$$

其中, $R(\theta', \theta)$ 的定义是

$$R(\theta', \theta) = \sum_{\mathbf{Z}^m} \ell_1(\theta'; \mathbf{Z}^m|\mathbf{Z}) \Pr(\mathbf{Z}^m|\mathbf{Z}, \theta) \quad (1.5)$$

is the expectation of a log-likelihood of a density(indexed by θ'), with respect to the same density indexed by θ , 当 $\theta' = \theta$ 时, 作为 θ' 的函数取最大值。因而, 当极大化 $Q(\theta', \theta)$ 时, 可以得出

$$\begin{aligned} \ell(\theta'; \mathbf{Z}) - \ell(\theta; \mathbf{Z}) &= [Q(\theta', \theta) - Q(\theta, \theta)] - [R(\theta', \theta) - R(\theta, \theta)] \\ &\geq 0 \end{aligned} \quad (1.6)$$

1.5.3 EM as a Maximization-Maximization Procedure

One does not need to maximize with respect to all of the latent data parameters at once, but could instead maximize over one of them at a time, 而可以在 M 步轮流一次极大化它们中的一个。

1.6 MCMC for Sampling from the Posterior

MCMC(Markov chain Monte Carlo) 马尔科夫链蒙特卡洛方法。

1.7 Bagging(装袋)

Bagging(Bootstrap aggregation) 对自助法样本集上的预测求平均，从而降低方差。真实装袋估计的定义是 $E_{\hat{p}} \hat{f}^*(x)$ ，其中 \hat{p} 表示经验分布，即从实际总体中而不是数据中抽取样本。

仅当原来的估计是非线性的，或者是数据的自适应函数时，装袋估计与 $\hat{f}(x)$ 不同。

对于数模型来说，类概率估计为在末端节点中的类比例，Bagging 平均类概率通常可降低方差。

1.7.1 Example: Trees with Simulated Data

由于预测子的相关性，这些树具有较高的方差。Bagging 成功地光滑了这种方差，从而降低了检验误差。

Bagging 可以降低均方误差，因为平均可以降低不稳定过程（如树）的方差，而保持偏倚不变。参考[知乎：为什么说 bagging 是减少 variance，而 boosting 是减少 bias](#)，由于子集样本集的相似性以及使用的是同种模型，因此各模型有近似相等的 bias 和 variance。由于 $E[\sum x_i/n] = E[x_i]$ ，以 bagging 后的 bias 和单个子模型的接近，一般来说不能显著降低 bias。另一方面，若各子模型独立， $\text{var}(\sum x_i/n) = \text{var}(x_i)/n$ ，此时可以显著降低 variance。

训练样本是从 P 分布中抽取的不相关的样本（不重复），而自助样本也是从 P 中采样得到的。

当 bag 一个模型时，模型中任何简单结构都将失去。如 bagged 树已不再是树，对于模型的解释，这显然是一个缺点。

由 bag 计算的期望类概率不能在任何一个 single replication 上实现，在这种其意义上，bag 一定程度上增大了各基分类器的模型空间。对于该例子 (single split 分类器 bag 拟合双

向拟合 $x_1 + x_2 = 1$) 或者其它例子, 模型需要放大时, **bag** 没有帮助。