Parameter estimation for text analysis - 笔记

冯柏淋

2018 年 10 月 23 日

目录

1 引入 **2**

2 基本问题与贝叶斯公式 **2**

2.1 两类基本问题 2

2.2 贝叶斯公式 2

2.3 记号 2

3 极大似然估计 MLE **2**

3.1 参数估计 2

3.2 预测 3

3.3 一个例子 3

4 最大后验概率 MAP **3**

4.1 参数估计 3

4.2 预测 4

4.3 一个例子 4

5 贝叶斯估计 **4**

5.1 参数估计 4

5.2 预测 5

5.3 一个例子 5

1 引入

此笔记是本人在阅读论文 Gregor Heinrich:*Parameter estimation for text analysis* 的过程的一点笔记，大多是对于原文的翻译和摘录。另有一部分自己的理解，因而可能存在错误。

2 基本问题与贝叶斯公式

2.1 两类基本问题

- **估计** 估计分布的参数以解释观测值
- **预测/回归** 基于以往观测值，预测新的观测值的概率

2.2 贝叶斯公式

$$p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{p(\mathcal{X})} \tag{1}$$

对以上公式中的每一部分，我们有如下名称:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}} \tag{2}$$

2.3 记号

数据集:

$$\mathcal{X} \triangleq \{x\}_i^{|\mathcal{X}|}$$

3 极大似然估计 MLE

3.1 参数估计

极大似然函数:

$$L(\theta; \mathcal{X}) \triangleq p(\mathcal{X} \mid \theta) = \prod_{x \in \mathcal{X}} p(x \mid \theta) \tag{3}$$

极大似然估计 (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta; \mathcal{X}) = \arg\max_{\theta} \sum_{x \in \mathcal{X}} \log p(x \mid \theta) \tag{4}$$

3.2 预测

对于新的观测值的概率估计过程如下：[1]

$$
\begin{aligned}
p(\tilde{x} \mid \mathcal{X}) &= \int_{\theta \in \Theta} p(\tilde{x} \mid \theta) \cdot p(\theta \mid \mathcal{X}) \, \mathrm{d}\theta && (5) \\
&\approx \int_{\theta \in \Theta} p(\tilde{x} \mid \hat{\theta}_{\mathrm{MLE}}) \cdot p(\theta \mid \mathcal{X}) \, \mathrm{d}\theta && (6) \\
&= p(\tilde{x} \mid \hat{\theta}_{\mathrm{MLE}}). && (7)
\end{aligned}
$$

3.3 一个例子

暂时略。

4 最大后验概率 MAP

4.1 参数估计

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &= \arg\max_{\theta} \, p(\theta \mid \mathcal{X}) && (8) \\
&= \arg\max_{\theta} \, \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{p(\mathcal{X})} && (9) \\
&= \arg\max_{\theta} \, p(\mathcal{X} \mid \theta) \cdot p(\theta) && (10) \\
&= \arg\max_{\theta} \, \{\mathcal{L}(\theta \mid \mathcal{X}) + \log p(\theta)\} && (11) \\
&= \arg\max_{\theta} \, \{\sum_{x \in \mathcal{X}} \log p(x \mid \theta) + \log p(\theta)\} && (12)
\end{aligned}
$$

MAP 相比于 MLE 而言，添加了关于先验分布的额外信息 $p(\theta)$。实践中，可以通过假定一些简单的先验分布 $p(\theta)$ 来防止过拟合。[2]

[1] 此处，$p(\tilde{x} \mid \hat{\theta}_{\mathrm{MLE}})$ 视为常数，并且 $\int_{\theta \in \Theta} p(\theta \mid \mathcal{X}) \, \mathrm{d}\theta = 1$

[2] Occam's razor 原理

4.2 预测

$$p(\tilde{x} \mid \mathcal{X}) \approx \int_{\theta \in \Theta} p(\tilde{x} \mid \hat{\theta}_{\mathrm{MAP}}) \cdot p(\theta \mid \mathcal{X}) \, \mathrm{d}\theta \tag{13}$$

$$= p(\tilde{x} \mid \hat{\theta}_{\mathrm{MAP}}). \tag{14}$$

4.3 一个例子

暂时略。

5 贝叶斯估计

5.1 参数估计

贝叶斯估计是对 MAP 的一个拓展。MAP 对于参数只产生一个确定的估计值 $\hat{\theta}_{\mathrm{MAP}}$，而贝叶斯估计则产生对于参数的一个分布。从而，贝叶斯分布则提供了关于参数的额外信息，比如期望和方差。

主要步骤为计算后验分布：

$$p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{p(\mathcal{X})} \tag{15}$$

在前述的 MAP 方法中，只需要最大化此式，则无需求出分母 $p(\mathcal{X})$，而在此处则需要计算，且：

$$p(\mathcal{X}) = \int_{\theta \in \Theta} p(\mathcal{X}|\theta) \cdot p(\theta) \, \mathrm{d}\theta \tag{16}$$

因为没最大化参数的过程，参数估计的过程也就不是一个具体的值，而是求取参数的分布：

$$p(\theta \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{p(\mathcal{X})} \tag{17}$$

$$= \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{\int_{\theta \in \Theta} p(\mathcal{X}|\theta) \cdot p(\theta) \, \mathrm{d}\theta} \tag{18}$$

5.2 预测

$$p(\tilde{x} \mid \mathcal{X}) = \int_{\theta \in \Theta} p(\tilde{x} \mid \theta) \cdot p(\theta \mid \mathcal{X}) \, \mathrm{d}\theta \tag{19}$$

$$= \int_{\theta \in \Theta} p(\tilde{x} \mid \theta) \cdot \frac{p(\mathcal{X} \mid \theta) \cdot p(\theta)}{p(\mathcal{X})} \, \mathrm{d}\theta. \tag{20}$$

5.3 一个例子

暂时略。