Bayes 方法学习笔记

李向阳   d1142845997@gmail.com

目录

1 引入

之前只学过基本的概率论与数理统计课程, 并没有系统的接触过贝叶斯统计学. 因此, 为了以后学习的方便, 把贝叶斯统计的主要方法记录下来, 以及在机器学习中的应用也稍微总结下.

2 贝叶斯统计

2.1 贝叶斯推断方法

在贝叶斯学派看来, 一切未知参数 $\boldsymbol{\theta}$ 都可以看成是一个随机变量, 有概率分布, 这个概率分布称为先验分布. 观测到样本以后, 利用样本分布以及 $\boldsymbol{\theta}$ 的先验分布, 我们可以导出 $\boldsymbol{\theta}$ 的后验分布, 统计推断基于后验分布进行.

由贝叶斯公式可得

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x},\boldsymbol{\theta})}{\displaystyle\int_{\boldsymbol{\theta}} p(\boldsymbol{x},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \\ &= \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\displaystyle\int_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}} \end{aligned} \tag{1}$$

这里使用了不太清晰的记法, 其中 $p(\boldsymbol{\theta}|\boldsymbol{x})$ 表示参数 $\boldsymbol{\theta}$ 的后验分布, $p(\boldsymbol{x},\boldsymbol{\theta})$ 表示样本与参数的联合概率密度, $p(\boldsymbol{x})$ 表示样本的边际密度 (边际密度可由联合密度积分而得, 其实这都是概率论里面随机变量的知识).

最关键的是 $p(\boldsymbol{x}|\boldsymbol{\theta})$, 我们用它表示的是观测样本 $\boldsymbol{x}$ 的分布. 从贝叶斯统计的观点看, 样本分布是给定 $\boldsymbol{\theta}$ 条件下的条件分布 $p(\boldsymbol{x}|\boldsymbol{\theta})$, 它是样本的联合概率函数, 也即是通常的似然函数 $L(\boldsymbol{\theta}|\boldsymbol{x})$(或记为 $L(\boldsymbol{\theta})$).

经典统计中, 称 $p(\boldsymbol{x};\boldsymbol{\theta})$ 为似然函数. 而在贝叶斯统计中, 样本分布 $p(\boldsymbol{x};\boldsymbol{\theta})$(即似然函数) 被看成是给定某 $\boldsymbol{\theta}$ 时样本 $\boldsymbol{x}$ 的条件分布, 因此记为 $p(\boldsymbol{x}|\boldsymbol{\theta})$.

本质上讲, 似然函数不是严格的概率分布, 但是在贝叶斯统计中, 我们将其看成是样本 $\boldsymbol{x}$ 的分布, 即看成是样本的概率 (密度) 函数, 因此公式 (1) 是成立的, 把似然函数看成是样本分布, 就容易理解了.

2.2 多参数模型

3 贝叶斯分类

3.1 模型预测

沿用以前的记号, 假设我们的训练数据集为 $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{Y}\}$, 其中 $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$ 表示 m 个样本, $\boldsymbol{Y} = \{y_1, y_2, \cdots, y_m\}$ 表示相应的类别标签. 一般我们用 $(\boldsymbol{x}, y)$ 表示单个的样本. 如果是训练数据, 那么其类别标签 y 就是已知的, 如果是待预测数据, 那么其类别标签就是未知的, 当然, 为了明确区分, 也可以把待预测数据记为 $(\tilde{\boldsymbol{x}}, \tilde{y})$, 其中 $\tilde{y}$ 未知.

所谓判别模型 (以下可以联想 Logistic 回归), 是给定新的预测样本 $\tilde{\boldsymbol{x}}$ 后, 想直接求出 $P(\tilde{y}|\tilde{\boldsymbol{x}})$, 即输出关于输入的条件 (概率) 分布, 当然, 从贝叶斯统计角度看, 这个分布的条件实际还有训练数据, 因为我们是看过训练数据之后, 学习到了对数据分布的后验认识, 然后根据这个认识对新的测试样本 $\tilde{\boldsymbol{x}}$ 做类别预测的, 也就是说可记为 $P(\tilde{y}|\tilde{\boldsymbol{x}}) = P(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{Y}, \boldsymbol{X})$.

我们建立模型, 认为这个条件 (概率) 分布是由参数 $\boldsymbol{\theta}$ 决定的 (不管是参数还是超参数, 都先包含在 $\boldsymbol{\theta}$ 里), 记为 $P(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\theta})$, 这个分布的形式是假设好的, 已知的.

那么如何由 $P(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\theta})$ 得到 $P(\tilde{y}|\tilde{\boldsymbol{x}})$ 呢?

经典统计的观点是假设参数 $\boldsymbol{\theta}$ 是固定不变的, 因此求得了其估计值, 也就得到 $P(\tilde{y}|\tilde{\boldsymbol{x}})$ 了. 我们这里采用的是贝叶斯统计的观点, 即认为参数 $\boldsymbol{\theta}$ 是一个随机变量, 它也是有概率分布的, 假如我们可以根据训练数据求出 $\boldsymbol{\theta}$ 的后验分布 $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})$, 那么就有很多种方法了, 比如可以采用 $\boldsymbol{\theta}$ 的最大后验估计作为点估计值, 这样就跟经典统计的方法类似, 此外, 也可以在整个参数空间上做平均, 即

$$
\begin{aligned}
P(\tilde{y}|\tilde{\boldsymbol{x}}) &= P(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{Y}, \boldsymbol{X}) && (2) \\
&= \int P(\tilde{y}, \boldsymbol{\theta}|\tilde{\boldsymbol{x}}, \boldsymbol{Y}, \boldsymbol{X}) \,\mathrm{d}\boldsymbol{\theta} = \int P(\tilde{y}|\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \cdot P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X}) \,\mathrm{d}\boldsymbol{\theta} && (3)
\end{aligned}
$$

从贝叶斯角度看, 最关键的是求出参数 $\boldsymbol{\theta}$ 的后验分布 $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})$, 有了后验分布, 可以用点估计做预测, 也可以在整个参数空间上进行积分做预测, 后者是 full bayesian 的观点.

3.2 模型训练

如何求出参数 $\boldsymbol{\theta}$ 的后验分布 $P(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{X})$ 呢?

其实这在贝叶斯统计中已经很普通了, 我们知道, 后验分布是正比于先验分布与似然函数的乘积的. 训练数据的似然函数为

$$P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}) = L(\boldsymbol{\theta}) = \prod_{i=1}^{m} P(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}) \tag{4}$$

根据贝叶斯公式, 有

$$P(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X}) = \frac{P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})\cdot P(\boldsymbol{\theta})}{P(\boldsymbol{Y}|\boldsymbol{X})} \tag{5}$$

其中

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \int P(\boldsymbol{Y},\boldsymbol{\theta}|\boldsymbol{X})\,\mathrm{d}\boldsymbol{\theta} = \int P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})\cdot P(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{6}$$

4 变分贝叶斯

变分贝叶斯的基本框架在介绍 Logistic 回归时已经讲过了, 这里再明确一点, 变分推断是干什么的?

上面我们看到, 用 full bayesian 的观点去做模型预测, 需要对后验分布做积分. 事实上, 即便不用 full bayesian 的观点, 也就是使用点估计, 那么在二次损失函数下, 参数 $\boldsymbol{\theta}$ 的贝叶斯点估计为后验均值, 也就是

$$\hat{\boldsymbol{\theta}} = \int \boldsymbol{\theta}\cdot P(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{X})\,\mathrm{d}\boldsymbol{\theta} \tag{7}$$

因此, 对后验分布的积分计算是无法避免的, 而解析的计算这些积分又很困难, 所以才有了各种近似算法, 比如 MCMC 采样算法. 而所谓变分推断, 便是寻求后验分布的近似表达式, 或者说用一个分布去近似后验分布, 使得积分的计算能够处理.

5 总结

5.1 参考资料

(1) 知乎: https://www.zhihu.com/collection/45422299, 魏晋的回答, 理清了一些关系, 虽然是借着生成模型与判别模型的区别回答的.

(2) 博客: http://www.flickering.cn/%E6%95%B0%E5%AD%A6%E4%B9%8B%E7%BE%8E/2014/06/lda%E6%95%B0%E5%AD%A6%E5%85%AB%E5%8D%A6mcmc-%E5%92%8C-gibbs-sampling/, 理论介绍的更为详细一些.

参考文献

[1] 李荣华. 偏微分方程数值解法. 高等教育出版社 (2010)

[2] Zhilin Li,Zhonghua Qiao,Tao Tang.*Numerical Solutions of Partial Differential Equations-An Introduction to Finite Difference and Finite Element Methods.*(2011)

[3] 孙志忠. 偏微分方程数值解法. 科学出版社 (2011)

附录