数据科学的数学基础
第一次理论作业

陈万祺　3220102895

2025 年 9 月 28 日

Problem 1.9

Consider two models M_1 and M_2, where from prior knowledge we believe that $Pr(M_1) = 0.25$ and $Pr(M_2) = 0.75$. We then observe a data set D. Given each model, we assess the likelihood of seeing that data, given the model, as $Pr(D|M_1) = 0.5$ and $Pr(D|M_2) = 0.01$. Now that we have the data, which model has a higher probability of being correct?

解： 由贝叶斯公式 $\Pr(M_i \mid D) = \frac{\Pr(D|M_i)\Pr(M_i)}{\Pr(D)}, i \in \{1,2\}$, 比较两个模型的后验概率只需比较分子项：

$$\begin{aligned}\Pr(D \mid M_1)\Pr(M_1) &= 0.5 \times 0.25 = 0.125,\\ \Pr(D \mid M_2)\Pr(M_2) &= 0.01 \times 0.75 = 0.0075.\end{aligned}$$

显然 $0.125 > 0.0075$，因此 $\Pr(M_1 \mid D) > \Pr(M_2 \mid D)$, 观测数据后模型 M_1 更有可能正确。

Problem 1.10

Assume I observe 3 data points x_1, x_2, and x_3 drawn independently from an unknown distribution. Given a model M, I can calculate the likelihood for each data point as $Pr(x_1|M) = 0.5$, $Pr(x_2|M) = 0.1$, and $Pr(x_3|M) = 0.2$. What is the likelihood of seeing all of these data points, given the model M: $Pr(x_1, x_2, x_3|M)$?

解： 由于数据点是独立同分布的，联合似然为各点似然的乘积：

$$\Pr(x_1, x_2, x_3|M) = \Pr(x_1|M) \cdot \Pr(x_2|M) \cdot \Pr(x_3|M) = 0.5 \times 0.1 \times 0.2 = 0.01.$$

因此，给定模型 M，观察到这三个数据点的联合似然为 0.01。

Problem 1.11

Consider a data set D with 10 data points (-1, 6, 0, 2, -1, 7, 7, 8, 4, -2). We want to find a model M from a restricted sample space $\Omega = \{0, 2, 4\}$. Assume the data has Laplace noise defined, so from a model M a data point's probability distribution is described as f(x) = (1/4) exp(-|M - x| / 2). Also assume we have a prior assumption on the models so that Pr(M = 0) = 0.25, Pr(M = 2) = 0.35, and Pr(M = 4) = 0.4. Assuming all data points in D are independent, which model is most likely?

解： 由于数据点独立，数据集 D 的似然函数为各数据点似然的乘积：

$$\Pr(D \mid M) = \prod_{i=1}^{10} f(x_i \mid M) = \prod_{i=1}^{10} \frac{1}{4} \exp\left(-\frac{|M - x_i|}{2}\right) = \left(\frac{1}{4}\right)^{10} \exp\left(-\frac{1}{2}\sum_{i=1}^{10} |M - x_i|\right).$$

根据贝叶斯定理，后验概率 $\Pr(M \mid D) \propto \Pr(D \mid M)\Pr(M)$。由于因子 $(1/4)^{10}$ 对所有模型都是常数，我们只需比较 $\Pr(M) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{10} |M - x_i|\right)$ 的大小。
计算每个模型的绝对偏差和：

- 对于 $M = 0$: $\sum |0 - x_i| = 1 + 6 + 0 + 2 + 1 + 7 + 7 + 8 + 4 + 2 = 38$
- 对于 $M = 2$: $\sum |2 - x_i| = 3 + 4 + 2 + 0 + 3 + 5 + 5 + 6 + 2 + 4 = 34$
- 对于 $M = 4$: $\sum |4 - x_i| = 5 + 2 + 4 + 2 + 5 + 3 + 3 + 4 + 0 + 6 = 34$

比较未归一化的后验概率：

模型 M	绝对偏差和	先验概率 $\Pr(M)$	相对后验概率
0	38	0.25	$0.25 \cdot e^{-19} \approx 1.40 \times 10^{-9}$
2	34	0.35	$0.35 \cdot e^{-17} \approx 1.45 \times 10^{-8}$
4	34	0.40	$0.40 \cdot e^{-17} \approx 1.66 \times 10^{-8}$

模型 $M = 4$ 具有最大的后验概率，因此是最有可能的模型。

Problem 1.13

A glassblower creates an intricate artistic glass bottle and attempts to measure its volume in liters; the 10 measures are: [1.82, 1.71, 2.34, 2.21, 2.01, 1.95, 1.76, 1.94, 2.02, 1.89). To sell the bottle, by regulation, she must label its volume up to O.1 Liters (it could be 1.7L or 1.8L or 1.9L, and so on). Her prior estimate is that it is 2.0L, but with a normal distribution with a standard deviation of 0.1; that is, her prior for the volume V being x is described by the pdf fv(x) = C • exp(-(2.0 - x) /(2 . 0.12) for some unknown constant C (since it is only valid at increments of O.1). Assuming the 10 empirical estimates of the volume are unbiased, but have a normal error with a standard deviation of O.2, what is the most likely model for the volume V?

解：

先验分布为 $V \sim N(2.0, 0.1^2)$，即先验概率密度函数为：

$$f_V(v) \propto \exp\left(-\frac{(v-2.0)^2}{2\times 0.1^2}\right) = \exp\left(-\frac{(v-2.0)^2}{0.02}\right).$$

给定 10 个独立测量值 $D = [1.82, 1.71, 2.34, 2.21, 2.01, 1.95, 1.76, 1.94, 2.02, 1.89]$，每个测量值在给定真实体积 $V = v$ 的条件下服从 $N(v, 0.2^2)$ 分布，因此似然函数为：

$$L(v) = \prod_{i=1}^{10} \frac{1}{0.2\sqrt{2\pi}} \exp\left(-\frac{(x_i - v)^2}{2\times 0.2^2}\right) \propto \exp\left(-\frac{\sum_{i=1}^{10}(x_i - v)^2}{0.08}\right).$$

后验概率满足：

$$P(V = v \mid D) \propto f_V(v)\cdot L(v) \propto \exp\left(-\frac{(v-2.0)^2}{0.02} - \frac{\sum_{i=1}^{10}(x_i - v)^2}{0.08}\right).$$

计算测量值的样本均值：

$$\bar{x} = \frac{1.82 + 1.71 + 2.34 + 2.21 + 2.01 + 1.95 + 1.76 + 1.94 + 2.02 + 1.89}{10} = 1.965.$$

利用正态分布共轭先验的性质，后验分布为 $N(\mu_{\text{post}}, \sigma^2_{\text{post}})$，其中：

$$\frac{1}{\sigma^2_{\text{post}}} = \frac{1}{0.1^2} + \frac{10}{0.2^2} = 100 + 250 = 350, \quad \sigma^2_{\text{post}} = \frac{1}{350},$$

$$\mu_{\text{post}} = \frac{\frac{2.0}{0.1^2} + \frac{10\times 1.965}{0.2^2}}{350} = \frac{200 + 491.25}{350} = \frac{691.25}{350} = 1.975.$$

因此，后验概率密度函数与 $\exp\left(-175(v-1.975)^2\right)$ 成正比。由于体积 V 必须为 0.1 升的倍数，比较候选值 $v = 1.9$、$v = 2.0$ 和 $v = 2.1$ 的后验概率：

对于 $v = 1.9$：$(1.975 - 1.9)^2 = 0.005625$，指数项为 $175 \times 0.005625 = 0.984375$。

对于 $v = 2.0$：$(1.975 - 2.0)^2 = 0.000625$，指数项为 $175 \times 0.000625 = 0.109375$。

对于 $v = 2.1$：$(1.975 - 2.1)^2 = 0.015625$，指数项为 $175 \times 0.015625 = 2.734375$。

后验概率与 $\exp(-\text{指数项})$ 成正比，因此 $v = 2.0$ 对应的指数项最小，后验概率最大。故最可能的模型为 $V = 2.0\,\text{L}$。