ST495/590 Assignment 3 - Solutions

Bayes theorem:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

(1) *Develop spell checker. The dictionary contains* 6 *words:* {*cat, hat, fat, cap, cup, hot*}.

- Define the check rule. (Note: This is not the only case, and one can define his/her own check rule.)

  Assume the typed word always has the same length with the intended word, i.e., in this problem, both are of length 3. Let X be the number of wrong letters, so $X \in \{0, 1, 2, 3\}$. Compare the letter at the same position, if they don't match, then X should add 1. In particular, if the two letters next to each other are only switched, then X should add 1 instead of 2. The relevant R function is `stringdist` in `stringdist` package.

  For example, if the intended word is "hat", and when the typed word is "cat", $X = 1$; when the typed word is "aht", $X = 1$.

- Suppose $A_i = \{$The intended word is the ith word in the dictionary$\}$ for $i = 1, \cdots, 6$, and $B = \{$The word is the typed word$\}$. The goal is to get $P(A_i|B)$.

- Calculate $P(A_j)$ for $j = 1, \cdots, 6$. Since in the dictionary, all words are equally likely except that "hot" is α times as likely as the other words, then

$$P(A_j) = \begin{cases} \frac{1}{5+\alpha} & (j = 1, 2, 3, 4, 5) \\ \frac{\alpha}{5+\alpha} & (j = 6) \end{cases}$$

- Calculate $P(B|A_j)$. Suppose the number of wrong letters is X_j when the intended word is the jth word in the dictionary.

$$P(B|A_j) = \theta^{X_j}(1-\theta)^{3-X_j}$$

- By Bayes theorem,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)} = \frac{\theta^{X_i}(1-\theta)^{3-X_i}P(A_i)}{\sum_j \theta^{X_j}(1-\theta)^{3-X_j}P(A_j)}.$$

- Say the typed word is "hat". For the three cases of α and θ, the results are as follows. It shows that as α goes larger the probability that the intended word is "hot" will increase, this is because the prior probability of the "hot" in the

dictionary increases. As θ increases to 0.95, this means it is very likely to type the wrong letters, so when the typed word is "hat", the most possible intended word would be the word whose all 3 letters are different from "hat", i.e., the probability of the intended word is "cup" in the dictionary is 0.939.

intended word	cat	hat	fat	cap	cup	hot
probability ($\alpha = 2$, $\theta = 0.1$)	0.076	0.686	0.076	0.008	0.0009	0.152
probability ($\alpha = 50$, $\theta = 0.1$)	0.016	0.147	0.016	0.002	0.0002	0.818
probability ($\alpha = 2$, $\theta = 0.95$)	0.0026	0.0001	0.003	0.049	0.939	0.005

(2) *Poisson likelihood and discrete uniform prior.*

- The prior is a discrete uniform distribution for λ with the following

$$P(\lambda = i) = \frac{1}{21} \ (i = 0, 1, \cdots, 20).$$

- Compute the prior mean and standard deviation.

$$\text{prior mean} = \sum_{i=0}^{20} iP(\lambda = i) = 10, \ \text{prior std} = \sqrt{\sum_{i=0}^{20}(i-10)^2 P(\lambda = i)} = \sqrt{37} = 6.06$$

- Find an interval so that λ in this interval with prior probability of 0.9.

$$P(k \le \lambda \le l) = 0.9 = \sum_{i=k}^{l} P(\lambda = i) = \frac{1}{21}(l - k + 1)$$

So $l - k = 17.9$, the interval $(2, 20)$ satisfies this, and the probability of λ is in this interval is about 0.9.

Note: The interval is not unique. It can be as symmetric as possible, or you can fix k at 0, then to find l.

- $Y|\lambda$ is following a Poisson(λ) distribution.

$$P(Y = Y|\lambda) = e^{-\lambda}\frac{\lambda^Y}{Y!}$$

- Find the posterior distribution. For $i = 0, \cdots, 20$,

$$P(\lambda = i|Y = 2) = \frac{P(Y|\lambda = i)P(\lambda = i)}{\sum_{j=0}^{20} P(Y|\lambda = j)P(\lambda = j)} = \frac{e^{-i}\frac{i^Y}{Y!}\frac{1}{21}}{\sum_{j=0}^{20} e^{-j}\frac{j^Y}{Y!}\frac{1}{21}} = \frac{e^{-i}i^2}{\sum_{j=0}^{20} e^{-j}j^2}.$$

- Compute the posterior mean and standard deviation.

$$\text{posterior mean} = \sum_{i=0}^{20} iP(\lambda = i|Y = 2) = \frac{\sum_{i=0}^{20} ie^{-i}i^2}{\sum_{j=0}^{20} e^{-j}j^2} = 3.01$$

$$\text{posterior std} = \sqrt{\sum_{i=0}^{20}(i-3.01)^2 P(\lambda = i|Y=2)} = 1.72$$

- Find an interval so that λ in this interval with posterior probability of 0.9.

$$P(k \leq \lambda \leq l|Y=2) = 0.9 = \sum_{i=k}^{l} P(\lambda = i|Y) = \frac{\sum_{i=k}^{l} e^{-i} i^2}{\sum_{j=0}^{20} e^{-j} j^2}$$

By enumerating, we find the interval is $(2, 6)$. (Note: This is not unique!)