Mohsen Abbasi

---

1 NAIVE BAYES

2 EM ALGORITHM

1. The Bayes rule states:

$$P(\lambda|x) = \frac{P(x|\lambda)P(\lambda)}{P(x)}$$

and by using the Maximum Likelihood definition we have:

$$\begin{aligned}
\lambda_{MLE} &= \underset{\lambda}{\text{argmax}}\, P(x|\lambda) \\
&= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{n} P(x_i|\lambda) \\
&= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{n} e^{-\lambda}\lambda^{x_i} \\
&= \underset{\lambda}{\text{argmax}} \prod_{i=1}^{n} e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i} = \prod_{i=1}^{n} ln(e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i}) \\
&= \underset{\lambda}{\text{argmax}} -n\lambda + \sum_{i=1}^{n} x_i ln(\lambda)
\end{aligned}$$

To find the maximum value, we take the derivative of the the last one:

$$-n + \sum_{i=1}^{n}(\frac{1}{\lambda}) = 0$$

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

---

2. In this part we care about the generation of the data based on the model and not the prediction of the labels. For a Smith's record, we consider λ_s and η which are the Poisson parameter and prior probability of belonging to Smith's respectively:

$$\begin{aligned}
P(X, Y = s|\lambda_s, \eta) &= P(Y = s|\lambda_s, \eta)P(X|Y = s, \lambda_s, \eta)) \\
&= \eta \prod_{i=1}^{n} \frac{\lambda_s^{x_i} e^{-\lambda_s}}{x_i!} \textit{ (By Naive Bayes assumption)}
\end{aligned}$$

For a Trader Joe's record, we consider λ_t and $1-\eta$ which are the Poisson parameter and prior probability of belonging to Trader Joe's respectively:

$$\begin{aligned} P(X, Y = t|\lambda_t, 1-\eta) &= P(Y = t|\lambda_t, 1-\eta)P(X|Y = t, \lambda_t, 1-\eta)) \\ &= (1-\eta)\prod_{i=1}^{n}\frac{\lambda_t^{x_i}e^{-\lambda_t}}{x_i!} \textit{ (By Naive Bayes assumption)} \end{aligned}$$

---

3. The probability of belonging to Smith's for each record is:

$$P(Y = s|x, \lambda_s, \eta) = \frac{P(x|Y = s, \lambda_s, \eta)P(Y = s|\lambda_s \eta)}{P(x|\lambda_s, \eta)}$$

The probability of belonging to Trader Joe's for each record is calculated the same way except for the parameters which change to λ_t and $1-\eta$. Therefore it's easy to figure out which probability is higher using:

$$\underset{Y}{\operatorname{argmax}} P(x|Y = s, \lambda_s, \eta)P(Y = s|\lambda_s, \eta)$$

$$\underset{Y}{\operatorname{argmax}} P(x|Y = t, \lambda_t, 1-\eta)P(Y = t|\lambda_t, 1-\eta)$$

And each record is put in the cluster which results in a higher probability.

3 Experiment