# Probabilistic Models: Spring 2014
# Poisson Mixture Model Example Solutions

We are given the following dataset $\mathcal{D}$ of DNA sequences from a metagenomics sample. Use a Poisson mixture model (PMM) to cluster the sequences into two groups.

| Index | Sequence | Count |
|---|---|---|
| $D_1$ | AACCTGCCG | 1 |
| $D_2$ | CGCGCTCAA | 12 |
| $D_3$ | AGTGTGAGC | 3 |
| $D_4$ | TGGGTACAC | 11 |
| $D_5$ | GGCCGCGTG | 15 |
| $D_6$ | CCTTAAGAG | 2 |
| $D_7$ | GCGGAACTG | 9 |
| $D_8$ | GCGTTGTAG | 17 |
| $D_9$ | GTTGTAGCG | 20 |
| $D_{10}$ | ACACGTGAC | 16 |

Use the following initial parameters for the PMM.

- $p_1$: 0.4
- $p_2$: 0.6
- $\lambda_1$: 7
- $\lambda_2$: 8

1. Calculate the $P(k|l)$ values for $D_1$, $D_2$ and $D_3$. The others are as follows.
   To calculate $P(k|D_1)$, we first need the $q(k, 1)$ values.

$$\begin{aligned} q(k=1,1) &= p_1 \cdot g(D_1, \lambda_1) \\ &= 0.4 \cdot 0.0064 \\ &= 0.0026 \end{aligned}$$

$$\begin{aligned} q(k=2,1) &= p_2 \cdot g(D_1, \lambda_2) \\ &= 0.6 \cdot 0.0027 \\ &= 0.0016 \end{aligned}$$

With these values, we can see that $\sum_m^K q(m,1) = 0.0026 + 0.0016 = 0.0042$. We can now calculate the $P(k|D_1)$ values.

$$\begin{aligned} P(k=1|D_1) &= \frac{q(k=1,1)}{\sum_m^K q(m,1)} \\ &= \frac{0.0026}{0.0042} \\ &= 0.6133 \end{aligned}$$

$$\begin{aligned} P(k=2|\ D_1) &= \frac{q(k=2,1)}{\sum_m^K q(m,1)} \\ &= \frac{0.0016}{0.0042} \\ &= 0.3867 \end{aligned}$$

The calculations for $D_2$ and $D_3$ are similar.

| Index | $P(k=1\|\ D_l)$ | $P(k=2\|\ D_l)$ |
|---|---|---|
| $D_1$ | 0.6133 | 0.3867 |
| $D_2$ | 0.2674 | 0.7326 |
| $D_3$ | 0.5483 | 0.4517 |
| $D_4$ | 0.2944 | 0.7056 |
| $D_5$ | 0.1965 | 0.8035 |
| $D_6$ | 0.5811 | 0.4189 |
| $D_7$ | 0.3527 | 0.6473 |
| $D_8$ | 0.1577 | 0.8423 |
| $D_9$ | 0.1114 | 0.8886 |
| $D_{10}$ | 0.1762 | 0.8238 |

2. Use the $P(k|l)$ values to calculate $\sum_l^N P(k|l)$ and $\sum_l^N \{P(k|l) \cdot D_l\}$ for each $k$.

$$\begin{aligned} \sum_l^N P(k=1|l) &= .6133 + .2674 + .5483 + .2944 + .1965 + .5811 + .3527 + .1577 + .1114 + .1762 \\ &= 3.2990 \\ \sum_l^N P(k=2|l) &= 6.7001 \end{aligned}$$

$$\begin{aligned} \sum_l^N \{P(k=1|l) \cdot D_l\} &= 23.7183 \\ \sum_l^N \{P(k=2|l) \cdot D_l\} &= 82.2817 \end{aligned}$$

3. Calculate the updated values for $\lambda_k$ and $p_k$.

   We can find the new parameters directly from the update equations. The only thing to recall is that $Z(k) \coloneqq \sum_l^N P(k|l)$

$$\begin{aligned}
\lambda_1 &= \frac{\sum_l^N P(k=1|l) D_l}{Z(k=1)} \\
&= \frac{23.7183}{3.2990} \\
&= 7.1895 \\
\lambda_2 &= \frac{\sum_l^N P(k=2|l) D_l}{Z(k=2)} \\
&= \frac{82.2817}{6.7001} \\
&= 12.2791 \\
p_1 &= \frac{Z(k=1)}{N} \\
&= \frac{3.2990}{10} \\
&= 0.3299 \\
p_2 &= \frac{Z(k=2)}{N} \\
&= \frac{6.7001}{10} \\
&= 0.6700
\end{aligned}$$

4. Use the new parameters to calculate $P(k|l)$ for $D_1$.

   To calculate $P(k|D_1)$, we first need the $q(k,1)$ values.

$$\begin{aligned}
q(k=1,1) &= p_1 \cdot g(D_1, \lambda_1) \\
&= 0.3299 \cdot 0.0054 \\
&= 0.0018
\end{aligned}$$

$$\begin{aligned}
q(k=2,1) &= p_2 \cdot g(D_1, \lambda_2) \\
&= 0.6700 \cdot 5.7e-05 \\
&= 3.824e-05
\end{aligned}$$

   With these values, we can see that $\sum_m^K q(m,1) = 0.0018 + 3.824e-05 =$

0.001827. We can now calculate the $P(k|D_1)$ values.

$$\begin{aligned} P(k=1|D_1) &= \frac{q(k=1,1)}{\sum_m^K q(m,1)} \\ &= \frac{0.0018}{0.001827} \\ &= 0.9791 \end{aligned}$$

$$\begin{aligned} P(k=2|\ D_1) &= \frac{q(k=2,1)}{\sum_m^K q(m,1)} \\ &= \frac{3.824e-05}{0.001827} \\ &= 0.0209 \end{aligned}$$

5. After 5 iterations, we have the following values.

   - $p_1$: 0.2997
   - $p_2$: 0.7003
   - $\lambda_1$: 2.0035
   - $\lambda_2$: 14.2799
   - $P(k=1|3)$: 0.9961
   - $P(k=2|3)$: 0.0039
   - $\sum_l^N P(k=1|l)$: 2.9974
   - $\sum_l^N P(k=2|l)$: 7.0026
   - $\sum_l^N \{P(k=1|l) \cdot D_l\}$: 6.0049
   - $\sum_l^N \{P(k=2|l) \cdot D_l\}$: 99.9951

   Use these values to calculate new values for $\lambda_k$ and $p_k$. Then, use those to calculate $P(k|3)$ for each $k$.

   We can find the new parameters directly from the update equations. The

only thing to recall is that $Z(k) = \sum_l^N P(k|l)$.

$$\begin{aligned}
\lambda_1 &= \frac{\sum_l^N P(k=1|l)D_l}{Z(k=1)} \\
&= \frac{6.0049}{2.9974} \\
&= 2.0033 \\
\lambda_2 &= \frac{\sum_l^N P(k=2|l)D_l}{Z(k=2)} \\
&= \frac{99.9951}{7.0026} \\
&= 14.2797 \\
p_1 &= \frac{Z(k=1)}{N} \\
&= \frac{2.9974}{10} \\
&= 0.2997 \\
p_2 &= \frac{Z(k=2}{N} \\
&= \frac{7.0026}{10} \\
&= 0.7003
\end{aligned}$$

To calculate $P(k|D_3)$, we first need the $q(k,3)$ values.

$$\begin{aligned}
q(k=1,3) &= p_1 \cdot g(D_3, \lambda_1) \\
&= 0.2997 \cdot 0.1807 \\
&= 0.0542
\end{aligned}$$

$$\begin{aligned}
q(k=2,3) &= p_2 \cdot g(D_3, \lambda_2) \\
&= 0.7003 \cdot 0.0003 \\
&= 0.0002
\end{aligned}$$

With these values, we can see that $\sum_m^K q(m,3) = 0.0542 + 0.0002 = 0.0544$.

We can now calculate the $P(k|D_3)$ values.

$$\begin{aligned} P(k=1|D_3) &= \frac{q(k=1,3)}{\sum_m^K q(m,3)} \\ &= \frac{0.0542}{0.0544} \\ &= 0.9961 \end{aligned}$$

$$\begin{aligned} P(k=2|\ D_3) &= \frac{q(k=2,1)}{\sum_m^K q(m,1)} \\ &= \frac{0.0002}{0.0544} \\ &= 0.0039 \end{aligned}$$

Note that the $P(k|3)$ values did not change (up to the significant digits shown).

# Useful Equations and Algorithms

$$g(D_l : \lambda) = \frac{\lambda^{D_l}}{D_l!} \exp\{-\lambda\}$$

$$Z(k) := \sum_l^N P(k|l)$$

---

**procedure** PMMEM(data $\mathcal{D}$, inital $p_1 \dots p_K, \lambda_1 \dots \lambda_K$, convergence criteria $\mathcal{C}$)
  **while** $\mathcal{C}$ has not been met **do**
    ▷ Update the expectations
    $q(k,l) \leftarrow p_k \cdot g(D_l, \lambda_k)$
    $P(k|l) \leftarrow \frac{q(k,l)}{\sum_m^K q(m,l)}$
    ▷ Maximize the parameters
    $\lambda_k \leftarrow \frac{\sum_l^N P(k|l) D_l}{Z(k)}$
    $p_k \leftarrow \frac{Z(k)}{N}$
  **end while**
**end procedure**

---