Poisson Mixture Models

Brandon Malone

Much of this material is adapted from Bilmes 1998 and Tomasi 2004.

Many of the images were taken from the Internet

February 20, 2014

Poisson Mixture Models

Suppose we have a dataset $\mathcal{D}$ which consists of DNA sequences observed from a mixture of k bacteria. We do not know which sequence belongs to which species.

Sequence	Species	Count
CAGAGGAT	?	5
TCAGTGTC	?	13
CTCTGTGA	?	2
AACTGTCG	?	7
CGCGTGGA	?	15
GGATGAGA	?	1

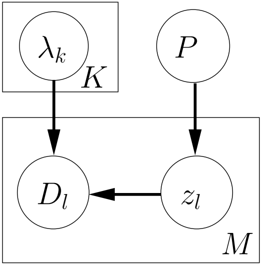
Which DNA sequences belong to the same species?

Poisson Mixture Models

Suppose we have a dataset $\mathcal{D}$ which consists of DNA sequences observed from a mixture of k bacteria. We do not know which sequence belongs to which species.

Sequence	Species	Count
CAGAGGAT	?	5
TCAGTGTC	?	13
CTCTGTGA	?	2
AACTGTCG	?	7
CGCGTGGA	?	15
GGATGAGA	?	1

$\Rightarrow$



Which DNA sequences belong to the same species?

This can be described by a **Poisson mixture model**.

1. The Poisson Distribution
2. Mixture Models
3. Expectation-Maximization
4. Wrap-up

Multiple Bernoulli trials

Suppose we have a **Bernoulli**-distributed variable (a weighted coin flip with parameter θ).

If we flip two coins, what is our probability of seeing *exactly* one H?

Multiple Bernoulli trials

Suppose we have a **Bernoulli**-distributed variable (a weighted coin flip with parameter θ).

If we flip two coins, what is our probability of seeing *exactly* one H?

C_1	C_2	$P(C_1, C_2)$
H	H	$\theta \cdot \theta$
H	T	$\theta \cdot (1 - \theta)$
T	H	$(1 - \theta) \cdot \theta$
T	T	$(1 - \theta) \cdot (1 - \theta)$

So, $P(\text{exactly one H}) = 2 \cdot \theta \cdot (1 - \theta)$.

Multiple Bernoulli trials

Suppose we have a **Bernoulli**-distributed variable (a weighted coin flip with parameter θ).

If we flip two coins, what is our probability of seeing *exactly* one H?

C_1	C_2	$P(C_1, C_2)$
H	H	$\theta \cdot \theta$
H	T	$\theta \cdot (1 - \theta)$
T	H	$(1 - \theta) \cdot \theta$
T	T	$(1 - \theta) \cdot (1 - \theta)$

So, $P(\text{exactly one H}) = 2 \cdot \theta \cdot (1 - \theta)$.

In general, $P(\text{exactly } m \text{ successes in } n \text{ trials}) = \binom{n}{m} \cdot \theta^m \cdot (1 - \theta)^{n-m}$.

Take it, to the limit, one more time

What if we have an infinite number of trials and expect to see λ successes?

$$\lim_{n \to \infty} P(\text{exactly } m \text{ successes in } n \text{ trials}) = \frac{\lambda^m}{m!} \exp\{-\lambda\}$$

This is called the **Poisson distribution**.

We will write $g(m : \lambda)$ to mean $P(\text{exactly } m \text{ successes given } \lambda)$.

(See the videos for a detailed derivation.)

Mixtures of distributions

Suppose we have K Poisson distributions (**components**) with parameters $\lambda_1 \dots \lambda_K$ **mixed** together with proportions $p_1 \dots p_K$.

We often write $P = \{p_1 \dots p_K\}$ and $\theta = \{\lambda_1 \dots \lambda_K, P\}$.

```
procedure GenerateDataset(Poisson parameters $\lambda_1 \dots \lambda_k$, mixing proportions $p_1 \dots p_k$, samples $N$)
    $\mathcal{D} \leftarrow \emptyset$
    for $l = 1$ to $N$ do
        component $z_l \leftarrow$sample(Mult($p_1 \dots p_K$))
        observation $D_l \leftarrow$sample(Poisson($\lambda_{z_l}$))
        $\mathcal{D} \leftarrow \mathcal{D} \cup D_l$
    end for
    return $\mathcal{D}$
end procedure
```

Mixtures of distributions

Suppose we have K Poisson distributions (**components**) with parameters $\lambda_1 \dots \lambda_K$ **mixed** together with proportions $p_1 \dots p_K$.

We often write $P = \{p_1 \dots p_K\}$ and $\theta = \{\lambda_1 \dots \lambda_K, P\}$.

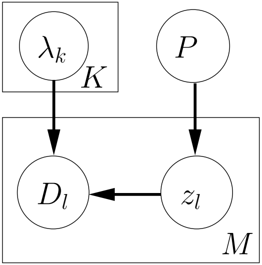


Figure: Generative model for a Poisson mixture model (PMM)

Likelihood of data

We can write the (log) probability of any mixture model as follows.

$$P(\mathcal{D} : \theta) = \sum_k^K p_k g(\mathcal{D} : \lambda_k)$$

$$P(\mathcal{D} : \theta) = \prod_l^N \sum_k^K p_k g(D_l : \lambda_k)$$

$$\ell(\mathcal{D} : \theta) = \log \prod_l^N \sum_k^K p_k g(D_l : \lambda_k)$$

$$\ell(\mathcal{D} : \theta) = \sum_l^N \log \sum_k^K p_k g(D_l : \lambda_k)$$

The learning problem can be formulated as follows.

$$\theta^* = \arg\max_\theta \ell(\mathcal{D} : \theta)$$

Membership probabilities

Notation

$$q(k,l) := p_k g(D_l : \lambda_k) \qquad \text{joint probability of } D_l \text{ and component } k$$
$$P(k|l) := P(z_l = k|D_l) \qquad \text{conditional probability of component } k \text{ given } D_l$$

The probability that D_l came from comonent k is expressed as follows.

$$P(k|l) = \frac{q(k,l)}{\sum_m^K q(m,l)}$$

Also, we know each observation came from *some* component.

$$\sum_k P(k|l) = 1$$

Jensen's Inequality

Recall the likelihood of the mixture model.

$$\ell(\mathcal{D} : \theta) = \sum_{l}^{N} \log \sum_{k}^{K} q(k, l)$$

Jensen's inequality shows the following.

$$\log \sum_{k}^{K} \pi_k \alpha_k \geq \sum_{k}^{K} \pi_k \log \alpha_k \qquad \text{when } \pi \text{ is a distribution}$$

We can make this work for any values.

$$\log \sum_{k}^{K} c_k = \log \sum_{k}^{K} c_k \frac{\pi_k}{\pi_k} = \log \sum_{k}^{K} \pi_k \frac{c_k}{\pi_k} \geq \sum_{k}^{K} \pi_k \log \frac{c_k}{\pi_k}$$

Expectation-Maximization (EM)

Our learning problem is formulated as follows.

$$\theta^* = \arg\max_{\theta} \ell(\mathcal{D} : \theta)$$

EM begins with a (bad) set of estimates for θ.

1. Use Jensen's inequality to estimate a bound b on ℓ called the **expectation** of ℓ
2. Find values of θ which **maximize** b

EM is guaranteed to find θs which do not decrease b.

Expectation and the Q function

Recall the definition of ℓ and Jensen's inequality.

$$\begin{aligned}\ell(\mathcal{D} : \theta) &= \sum_{l}^{N} \log \sum_{k}^{K} q(k, l) \\ &\geq \sum_{l}^{N} \sum_{k}^{K} P(k|l) \log \frac{q(k, l)}{P(k|l)}\end{aligned}$$

This gives the **expectation** of ℓ with our current parameters θ.

Based on this equation, we define $Q(\theta)$ which we want to maximize.

$$Q(\theta) = \sum_{l}^{N} \sum_{k}^{K} P(k|n) \log q(k, l)$$

(See the handout for a detailed derivation of Q.)

Maximization and the Q function

We use the following process to maximize Q for a particular parameter θ_i.

1. Differentiate Q w.r.t θ_i
2. Set the derivative equal to 0
3. Solve for θ_i

(See the handout for detailed derivations.)

$$\lambda_k = \frac{\sum_l^N P(k|l) D_l}{Z(k)}$$

$$p_k = \frac{Z(k)}{N}$$

The EM algorithm for PMMs

```
procedure PMMEM(data $\mathcal{D}$, inital $p_1 \ldots p_K, \lambda_1 \ldots \lambda_K$, 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
```

Grouping the DNA sequences into clusters

After running EM, we have several useful pieces of information about our metagenomics sample.

- $P(k|I)$. The distribution over species for each sequence.
- p_k. The relative genome sizes of the species.
- λ_k. The abundance of the species.

Other questions...

- Do we really know how many species there are?
- Can we differentiate species with similar abundances?
- How do we pick “good” initial parameters?
- When have we converged?

More on EM

EM is a general framework that is useful whenever data is missing.

- If used to estimate class probabilities in naive Bayes models, it is called Bayesian clustering
- If used in HMMs, it is called the Baum-Welch algorithm
- Can be used in general Bayesian networks to calculate parameters when some data is missing
- If used with structure learning algorthms, it is called Structural EM
- Many, many others...

We maximize likelihood with EM. What if we want MAP parameters?

Recap

During this part of the course, we have discussed:

- Mixture models as a probabilistic clustering method
- Expectation-maximization as a framework for estimating parameters when variables are hidden

Next in probabilistic models

We will see a Bayesian version of EM.

- Estimating parameters in topic models