Expectation Maximization

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

31 October 2006

- In this lecture, we outline the haplotype phasing algorithm of Excoffier & Slatin [].

- In this lecture, we outline the haplotype phasing algorithm of Excoffier & Slatin [].
- Our domain is the set $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ of n samples, where each sample x_i is drawn independently from a probability density $p(x|\theta)$.

- In this lecture, we outline the haplotype phasing algorithm of Excoffier & Slatin [].
- Our domain is the set $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ of n samples, where each sample x_i is drawn independently from a probability density $p(x|\theta)$.
- The vector of parameters $\theta = (\theta_1, ..., \theta_d) \in \mathbb{R}^d$ is fixed but unknown and we would like to estimate θ from the sampled data.

- In this lecture, we outline the haplotype phasing algorithm of Excoffier & Slatin [].
- Our domain is the set $\mathcal{D} = \{x_1, x_2, ..., x_n\}$ of n samples, where each sample x_i is drawn independently from a probability density $p(x|\theta)$.
- The vector of parameters $\theta = (\theta_1, ..., \theta_d) \in \mathbb{R}^d$ is fixed but unknown and we would like to estimate θ from the sampled data.

Recall that the log likelihood of θ with respect to $\mathcal{D}$ is given by the likelihood function

$$l(\theta) = \log p(\mathcal{D}|\theta) = \sum_{k=1}^{n} p(x_k|\theta)$$

Recall that the log likelihood of θ with respect to $\mathcal{D}$ is given by the likelihood function

$$l(\theta) = \log p(\mathcal{D}|\theta) = \sum_{k=1}^{n} p(x_k|\theta)$$

and the *maximum log-likelihood estimate* of θ, denoted $\hat{\theta}$, is the value of θ that maximizes $p(\mathcal{D}|\theta)$

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta).$$

- In the missing data problem, we have both observed and missing features $x_k = \{x_{kg}, x_{kb}\}$.

- In the missing data problem, we have both observed and missing features $x_k = \{x_{kg}, x_{kb}\}$.
- For missing features, we need to perform statistical inference to infer the most likely value of the missing feature.

- In the missing data problem, we have both observed and missing features $x_k = \{x_{kg}, x_{kb}\}$.
- For missing features, we need to perform statistical inference to infer the most likely value of the missing feature.
- Our central equation is the following.

$$\boxed{Q(\theta;\theta^i) = \mathcal{E}_{\mathcal{D}_b}[lnp(\mathcal{D}_g, \mathcal{D}_b|\theta)|\mathcal{D}_b;\theta^i]}$$

- In the missing data problem, we have both observed and missing features $x_k = \{x_{kg}, x_{kb}\}$.
- For missing features, we need to perform statistical inference to infer the most likely value of the missing feature.
- Our central equation is the following.

$$\boxed{Q(\theta; \theta^i) = \mathcal{E}_{\mathcal{D}_b}[lnp(\mathcal{D}_g, \mathcal{D}_b|\theta)|\mathcal{D}_b; \theta^i]}$$

- Q is a function of θ with θ^i assumed fixed.

- In the missing data problem, we have both observed and missing features $x_k = \{x_{kg}, x_{kb}\}$.
- For missing features, we need to perform statistical inference to infer the most likely value of the missing feature.
- Our central equation is the following.

$$\boxed{Q(\theta; \theta^i) = \mathcal{E}_{\mathcal{D}_b}[lnp(\mathcal{D}_g, \mathcal{D}_b|\theta)|\mathcal{D}_b; \theta^i]}$$

- Q is a function of θ with θ^i assumed fixed.

- Suppose we have an initial parameter vector θ^i that is the current best estimate for the full distribution.

- Suppose we have an initial parameter vector θ^i that is the current best estimate for the full distribution.
- We will use the candidate vector θ^i to obtain an improved estimate θ^{i+1}.

- Suppose we have an initial parameter vector θ^i that is the current best estimate for the full distribution.
- We will use the candidate vector θ^i to obtain an improved estimate θ^{i+1}.
- Given θ^i, the bound on the right hand side calculates the likelihood of the data including the unknown $\mathcal{D}_b$ marginalized with respect to the current distribution (given by θ^i).

- Suppose we have an initial parameter vector θ^i that is the current best estimate for the full distribution.
- We will use the candidate vector θ^i to obtain an improved estimate θ^{i+1}.
- Given θ^i, the bound on the right hand side calculates the likelihood of the data including the unknown $\mathcal{D}_b$ marginalized with respect to the current distribution (given by θ^i).
- The EM algorithm will select θ^{i+1} as the best such candidate, i.e.,

$$\theta^{i+1} = \arg\max_{\theta} Q(\theta; \theta^i).$$

- Suppose we have an initial parameter vector θ^i that is the current best estimate for the full distribution.
- We will use the candidate vector θ^i to obtain an improved estimate θ^{i+1}.
- Given θ^i, the bound on the right hand side calculates the likelihood of the data including the unknown $\mathcal{D}_b$ marginalized with respect to the current distribution (given by θ^i).
- The EM algorithm will select θ^{i+1} as the best such candidate, i.e.,

$$\theta^{i+1} = \arg\max_{\theta} Q(\theta; \theta^i).$$

- Note that by choosing different initial candidates θ^0, the output of the EM algorithm possibly varies.

- Note that by choosing different initial candidates θ^0, the output of the EM algorithm possibly varies.
- Therefore, in order to compute a good candidate for a global optimal solution, the EM algorithm is often run many times from different initial values θ^0.

- Note that by choosing different initial candidates θ^0, the output of the EM algorithm possibly varies.
- Therefore, in order to compute a good candidate for a global optimal solution, the EM algorithm is often run many times from different initial values θ^0.

The entire algorithm is given as follows.

The entire algorithm is given as follows.

Input: Observed data $\mathcal{D}$, initial estimate θ^0, margine of error ϵ
Output: Maximum likelihood parameters $\hat{\theta}$

1. Initialize $i = 0$.
   While $Q(\theta^{i+1}, \theta^i) - Q(\theta^i, \theta^{i-1}) \leq \epsilon$
   (i) E-Step: Compute $Q(\theta; \theta^i)$
   (ii) Max step: $\theta^{i+1} = \arg\max_\theta Q(\theta; \theta^i)$
   (iii) $i = i + 1$

   Return n $\hat{\theta} \leftarrow \theta^{i+1}$

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.
- Given observed samples $n_1, n_2 \ldots n_m$ of the phenotypes, the likelihood of the samples given the frequencies are given by the likelihood equation

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.
- Given observed samples $n_1, n_2 \ldots n_m$ of the phenotypes, the likelihood of the samples given the frequencies are given by the likelihood equation

$$P(Samples|P_1, \ldots, P_m) = \frac{n!}{n_1!n_2! \cdots n_m!} P_1^{n_1} P_2^{n_2} \ldots P_m^{n_m}$$

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.
- Given observed samples $n_1, n_2 \ldots n_m$ of the phenotypes, the likelihood of the samples given the frequencies are given by the likelihood equation

$$P(Samples|P_1, \ldots, P_m) = \frac{n!}{n_1!n_2! \cdots n_m!} P_1^{n_1} P_2^{n_2} \ldots P_m^{n_m}$$

- Under the assumption of Hardy-Weinberg equilibrium, or random mating, the likelihood of the haplotype frequencies given the phenotypic counts is given by

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.
- Given observed samples $n_1, n_2 \ldots n_m$ of the phenotypes, the likelihood of the samples given the frequencies are given by the likelihood equation

$$P(Samples|P_1, \ldots, P_m) = \frac{n!}{n_1!n_2! \cdots n_m!} P_1^{n_1} P_2^{n_2} \ldots P_m^{n_m}$$

- Under the assumption of Hardy-Weinberg equilibrium, or random mating, the likelihood of the haplotype frequencies given the phenotypic counts is given by

$$l(P_1, ..., P_h) = a_1 \cdot \prod_{j=1}^{m} (\sum_{i=1}^{C_j} P(h_{ik} h_{ie}))^{n_i}$$

- In this problem, we have m different phenotype with frequencies $P_1, P_2, \ldots P_m$.
- Given observed samples $n_1, n_2 \ldots n_m$ of the phenotypes, the likelihood of the samples given the frequencies are given by the likelihood equation

$$P(Samples|P_1, \ldots, P_m) = \frac{n!}{n_1!n_2! \cdots n_m!} P_1^{n_1} P_2^{n_2} \ldots P_m^{n_m}$$

- Under the assumption of Hardy-Weinberg equilibrium, or random mating, the likelihood of the haplotype frequencies given the phenotypic counts is given by

$$l(P_1, ..., P_h) = a_1 \cdot \prod_{j=1}^{m} (\sum_{i=1}^{C_j} P(h_{ik} h_{ie}))^{n_i}$$