6.864 (Fall 2007)

The EM Algorithm, Part I

An Experiment/Some Intuition

- I have three coins in my pocket,

  Coin 0 has probability λ of heads;
  Coin 1 has probability p_1 of heads;
  Coin 2 has probability p_2 of heads

- For each trial I do the following:

  First I toss Coin 0
  If Coin 0 turns up **heads**, I toss **coin 1** three times
  If Coin 0 turns up **tails**, I toss **coin 2** three times

  I don't tell you whether Coin 0 came up heads or tails,
  or whether Coin 1 or 2 was tossed three times,
  but I do tell you how many heads/tails are seen at each trial

- you see the following sequence:

$$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$$

What would you estimate as the values for λ, p_1 and p_2?

Overview

- Maximum-Likelihood Estimation
- Models with hidden variables
- The EM algorithm for a simple example (3 coins)
- The general form of the EM algorithm
- Hidden Markov models

Maximum Likelihood Estimation

- We have data points $x_1, x_2, \ldots x_n$ drawn from some set $\mathcal{X}$
- We have a parameter vector Θ
- We have a parameter space Ω
- We have a distribution $P(x \mid \Theta)$ for any $\Theta \in \Omega$, such that

$$\sum_{x \in \mathcal{X}} P(x \mid \Theta) = 1 \text{ and } P(x \mid \Theta) \geq 0 \text{ for all } x$$

- We assume that our data points $x_1, x_2, \ldots x_n$ are drawn at random (independently, identically distributed) from a distribution $P(x \mid \Theta^*)$ for some $\Theta^* \in \Omega$

Log-Likelihood

- We have data points $x_1, x_2, \ldots x_n$ drawn from some set $\mathcal{X}$
- We have a parameter vector Θ, and a parameter space Ω
- We have a distribution $P(x \mid \Theta)$ for any $\Theta \in \Omega$
- The likelihood is

$$Likelihood(\Theta) = P(x_1, x_2, \ldots x_n \mid \Theta) = \prod_{i=1}^{n} P(x_i \mid \Theta)$$

- The log-likelihood is

$$L(\Theta) = \log Likelihood(\Theta) = \sum_{i=1}^{n} \log P(x_i \mid \Theta)$$

A First Example: Coin Tossing

- $\mathcal{X} = \{\mathtt{H}, \mathtt{T}\}$. Our data points $x_1, x_2, \ldots x_n$ are a sequence of heads and tails, e.g.

`HHTTHHHTHH`

- Parameter vector Θ is a single parameter, i.e., the probability of coin coming up heads
- Parameter space $\Omega = [0, 1]$
- Distribution $P(x \mid \Theta)$ is defined as

$$P(x \mid \Theta) = \begin{cases} \Theta & \text{If } x = \mathtt{H} \\ 1 - \Theta & \text{If } x = \mathtt{T} \end{cases}$$

Maximum Likelihood Estimation

- Given a sample $x_1, x_2, \ldots x_n$, choose

$$\Theta_{ML} = \text{argmax}_{\Theta \in \Omega} L(\Theta) = \text{argmax}_{\Theta \in \Omega} \sum_i \log P(x_i \mid \Theta)$$

- For example, take the coin example:
  say $x_1 \ldots x_n$ has $Count(H)$ heads, and $(n - Count(H))$ tails
  $\Rightarrow$

$$\begin{aligned} L(\Theta) &= \log \left(\Theta^{Count(H)} \times (1 - \Theta)^{n - Count(H)} \right) \\ &= Count(H) \log \Theta + (n - Count(H)) \log(1 - \Theta) \end{aligned}$$

- We now have

$$\Theta_{ML} = \frac{Count(H)}{n}$$

A Second Example: Probabilistic Context-Free Grammars

- $\mathcal{X}$ is the set of all parse trees generated by the underlying context-free grammar. Our sample is n trees $T_1 \ldots T_n$ such that each $T_i \in \mathcal{X}$.
- R is the set of rules in the context free grammar
  N is the set of non-terminals in the grammar
- Θ_r for $r \in R$ is the parameter for rule r
- Let $R(\alpha) \subset R$ be the rules of the form $\alpha \rightarrow \beta$ for some α
- The parameter space Ω is the set of $\Theta \in [0, 1]^{|R|}$ such that

$$\text{for all } \alpha \in N \sum_{r \in R(\alpha)} \Theta_r = 1$$

- We have

$$P(T \mid \Theta) = \prod_{r \in R} \Theta_r^{Count(T,r)}$$

where $Count(T, r)$ is the number of times rule r is seen in the tree T

$$\Rightarrow \quad \log P(T \mid \Theta) = \sum_{r \in R} Count(T, r) \log \Theta_r$$

Maximum Likelihood Estimation for PCFGs

- We have

$$\log P(T \mid \Theta) = \sum_{r \in R} Count(T, r) \log \Theta_r$$

where $Count(T, r)$ is the number of times rule r is seen in the tree T

- And,

$$L(\Theta) = \sum_i \log P(T_i \mid \Theta) = \sum_i \sum_{r \in R} Count(T_i, r) \log \Theta_r$$

- Solving $\Theta_{ML} = \text{argmax}_{\Theta \in \Omega} L(\Theta)$ gives

$$\Theta_r = \frac{\sum_i Count(T_i, r)}{\sum_i \sum_{s \in R(\alpha)} Count(T_i, s)}$$

where r is of the form $\alpha \rightarrow \beta$ for some β

Multinomial Distributions

- $\mathcal{X}$ is a finite set, e.g., $\mathcal{X} = \{\texttt{dog}, \texttt{cat}, \texttt{the}, \texttt{saw}\}$
- Our sample $x_1, x_2, \ldots x_n$ is drawn from $\mathcal{X}$
  e.g., $x_1, x_2, x_3 = \texttt{dog}, \texttt{the}, \texttt{saw}$
- The parameter Θ is a vector in $\mathbb{R}^m$ where $m = |\mathcal{X}|$
  e.g., $\Theta_1 = P(dog)$, $\Theta_2 = P(cat)$, $\Theta_3 = P(the)$, $\Theta_4 = P(saw)$
- The parameter space is

$$\Omega = \{\Theta : \sum_{i=1}^{m} \Theta_i = 1 \text{ and } \forall i, \Theta_i \geq 0\}$$

- If our sample is $x_1, x_2, x_3 = \texttt{dog}, \texttt{the}, \texttt{saw}$, then

$$L(\Theta) = \log P(x_1, x_2, x_3 = \texttt{dog}, \texttt{the}, \texttt{saw}) = \log \Theta_1 + \log \Theta_3 + \log \Theta_4$$

Overview

- Maximum-Likelihood Estimation
- Models with hidden variables
- The EM algorithm for a simple example (3 coins)
- The general form of the EM algorithm
- Hidden Markov models

Models with Hidden Variables

- Now say we have two sets $\mathcal{X}$ and $\mathcal{Y}$, and a joint distribution $P(x, y \mid \Theta)$
- If we had **fully observed data**, (x_i, y_i) pairs, then

$$L(\Theta) = \sum_i \log P(x_i, y_i \mid \Theta)$$

- If we have **partially observed data**, x_i examples, then

$$\begin{aligned} L(\Theta) &= \sum_i \log P(x_i \mid \Theta) \\ &= \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \Theta) \end{aligned}$$

- The **EM (Expectation Maximization) algorithm** is a method for finding

$$\Theta_{ML} = \text{argmax}_\Theta \sum_i \log \sum_{y \in \mathcal{Y}} P(x_i, y \mid \Theta)$$

Overview

- Maximum-Likelihood Estimation
- Models with hidden variables
- The EM algorithm for a simple example (3 coins)
- The general form of the EM algorithm
- Hidden Markov models

The Three Coins Example

- e.g., in the three coins example:
  $\mathcal{Y} = \{\texttt{H}, \texttt{T}\}$
  $\mathcal{X} = \{\texttt{HHH}, \texttt{TTT}, \texttt{HTT}, \texttt{THH}, \texttt{HHT}, \texttt{TTH}, \texttt{HTH}, \texttt{THT}\}$
  $\Theta = \{\lambda, p_1, p_2\}$
- and

$$P(x, y \mid \Theta) = P(y \mid \Theta) P(x \mid y, \Theta)$$

where

$$P(y \mid \Theta) = \begin{cases} \lambda & \text{If } y = \texttt{H} \\ 1 - \lambda & \text{If } y = \texttt{T} \end{cases}$$

and

$$P(x \mid y, \Theta) = \begin{cases} p_1^h (1 - p_1)^t & \text{If } y = \texttt{H} \\ p_2^h (1 - p_2)^t & \text{If } y = \texttt{T} \end{cases}$$

where h = number of heads in x, t = number of tails in x

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(x = \mathsf{THT}, y = \mathsf{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(x = \mathsf{THT}, y = \mathsf{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

$$\begin{aligned} P(x = \mathsf{THT} \mid \Theta) &= P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) \\ &\quad + P(x = \mathsf{THT}, y = \mathsf{T} \mid \Theta) \\ &= \lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2 \end{aligned}$$

The Three Coins Example

- Various probabilities can be calculated, for example:

$$P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) = \lambda p_1 (1 - p_1)^2$$

$$P(x = \mathsf{THT}, y = \mathsf{T} \mid \Theta) = (1 - \lambda) p_2 (1 - p_2)^2$$

$$\begin{aligned} P(x = \mathsf{THT} \mid \Theta) &= P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta) \\ &\quad + P(x = \mathsf{THT}, y = \mathsf{T} \mid \Theta) \\ &= \lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2 \end{aligned}$$

$$\begin{aligned} P(y = \mathsf{H} \mid x = \mathsf{THT}, \Theta) &= \frac{P(x = \mathsf{THT}, y = \mathsf{H} \mid \Theta)}{P(x = \mathsf{THT} \mid \Theta)} \\ &= \frac{\lambda p_1 (1 - p_1)^2}{\lambda p_1 (1 - p_1)^2 + (1 - \lambda) p_2 (1 - p_2)^2} \end{aligned}$$

The Three Coins Example

- Fully observed data might look like:

$$(\langle HHH \rangle, H), (\langle TTT \rangle, T), (\langle HHH \rangle, H), (\langle TTT \rangle, T), (\langle HHH \rangle, H)$$

- In this case maximum likelihood estimates are:

$$\lambda = \frac{3}{5}$$

$$p_1 = \frac{9}{9}$$

$$p_2 = \frac{0}{6}$$

The Three Coins Example

- Partially observed data might look like:

$$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$$

- How do we find the maximum likelihood parameters?

The Three Coins Example

- Partially observed data might look like:

$$\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle$$

- If current parameters are λ, p_1, p_2

$$\begin{aligned} P(y = \mathsf{H} \mid x = \langle \mathsf{HHH} \rangle) &= \frac{P(\langle \mathsf{HHH} \rangle, \mathsf{H})}{P(\langle \mathsf{HHH} \rangle, \mathsf{H}) + P(\langle \mathsf{HHH} \rangle, \mathsf{T})} \\ &= \frac{\lambda p_1^3}{\lambda p_1^3 + (1-\lambda) p_2^3} \end{aligned}$$

$$\begin{aligned} P(y = \mathsf{H} \mid x = \langle \mathsf{TTT} \rangle) &= \frac{P(\langle \mathsf{TTT} \rangle, \mathsf{H})}{P(\langle \mathsf{TTT} \rangle, \mathsf{H}) + P(\langle \mathsf{TTT} \rangle, \mathsf{T})} \\ &= \frac{\lambda (1-p_1)^3}{\lambda (1-p_1)^3 + (1-\lambda)(1-p_2)^3} \end{aligned}$$

The Three Coins Example

- If current parameters are λ, p_1, p_2

$$P(y = \mathsf{H} \mid x = \langle \mathsf{HHH} \rangle) = \frac{\lambda p_1^3}{\lambda p_1^3 + (1-\lambda) p_2^3}$$

$$P(y = \mathsf{H} \mid x = \langle \mathsf{TTT} \rangle) = \frac{\lambda (1-p_1)^3}{\lambda (1-p_1)^3 + (1-\lambda)(1-p_2)^3}$$

- If $\lambda = 0.3, p_1 = 0.3, p_2 = 0.6$:

$$P(y = \mathsf{H} \mid x = \langle \mathsf{HHH} \rangle) = 0.0508$$

$$P(y = \mathsf{H} \mid x = \langle \mathsf{TTT} \rangle) = 0.6967$$

The Three Coins Example

- After filling in hidden variables for each example, partially observed data might look like:

$$
\begin{array}{ll}
(\langle \mathsf{HHH} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{HHH}) = 0.0508 \\
(\langle \mathsf{HHH} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{HHH}) = 0.9492 \\
(\langle \mathsf{TTT} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{TTT}) = 0.6967 \\
(\langle \mathsf{TTT} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{TTT}) = 0.3033 \\
(\langle \mathsf{HHH} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{HHH}) = 0.0508 \\
(\langle \mathsf{HHH} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{HHH}) = 0.9492 \\
(\langle \mathsf{TTT} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{TTT}) = 0.6967 \\
(\langle \mathsf{TTT} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{TTT}) = 0.3033 \\
(\langle \mathsf{HHH} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{HHH}) = 0.0508 \\
(\langle \mathsf{HHH} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{HHH}) = 0.9492
\end{array}
$$

The Three Coins Example

- New Estimates:

$$
\begin{array}{ll}
(\langle \mathsf{HHH} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{HHH}) = 0.0508 \\
(\langle \mathsf{HHH} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{HHH}) = 0.9492 \\
(\langle \mathsf{TTT} \rangle, H) & P(y = \mathsf{H} \mid \mathsf{TTT}) = 0.6967 \\
(\langle \mathsf{TTT} \rangle, T) & P(y = \mathsf{T} \mid \mathsf{TTT}) = 0.3033 \\
 & \ldots
\end{array}
$$

$$\lambda = \frac{3 \times 0.0508 + 2 \times 0.6967}{5} = 0.3092$$

$$p_1 = \frac{3 \times 3 \times 0.0508 + 0 \times 2 \times 0.6967}{3 \times 3 \times 0.0508 + 3 \times 2 \times 0.6967} = 0.0987$$

$$p_2 = \frac{3 \times 3 \times 0.9492 + 0 \times 2 \times 0.3033}{3 \times 3 \times 0.9492 + 3 \times 2 \times 0.3033} = 0.8244$$

The Three Coins Example: Summary

- Begin with parameters $\lambda = 0.3, p_1 = 0.3, p_2 = 0.6$
- Fill in hidden variables, using

$$P(y = \mathsf{H} \mid x = \langle \mathsf{HHH} \rangle) = 0.0508$$

$$P(y = \mathsf{H} \mid x = \langle \mathsf{TTT} \rangle) = 0.6967$$

- Re-estimate parameters to be $\lambda = 0.3092, p_1 = 0.0987, p_2 = 0.8244$

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967
1	0.3738	0.0680	0.7578	0.0004	0.9714	0.0004	0.9714
2	0.4859	0.0004	0.9722	0.0000	1.0000	0.0000	1.0000
3	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $\mathbf{y} = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$. The solution that EM reaches is intuitively correct: the coin-tosser has two coins, one which always shows up heads, the other which always shows tails, and is picking between them with equal probability ($\lambda = 0.5$). The posterior probabilities $\tilde{p}_i$ show that we are certain that coin 1 (tail-biased) generated y_2 and y_4, whereas coin 2 generated y_1 and y_3.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$	$\tilde{p}_5$
0	0.3000	0.3000	0.6000	0.0508	0.6967	0.0508	0.6967	0.0508
1	0.3092	0.0987	0.8244	0.0008	0.9837	0.0008	0.9837	0.0008
2	0.3940	0.0012	0.9893	0.0000	1.0000	0.0000	1.0000	0.0000
3	0.4000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

The coin example for $\{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle\}$. λ is now 0.4, indicating that the coin-tosser has probability 0.4 of selecting the tail-biased coin.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.3000	0.6000	0.1579	0.6967	0.0508	0.6967
1	0.4005	0.0974	0.6300	0.0375	0.9065	0.0025	0.9065
2	0.4632	0.0148	0.7635	0.0014	0.9842	0.0000	0.9842
3	0.4924	0.0005	0.8205	0.0000	0.9941	0.0000	0.9941
4	0.4970	0.0000	0.8284	0.0000	0.9949	0.0000	0.9949

The coin example for $\mathbf{y} = \{\langle HHT\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. EM selects a tails-only coin, and a coin which is heavily heads-biased ($p_2 = 0.8284$). It's certain that y_1 and y_3 were generated by coin 2, as they contain heads. y_2 and y_4 could have been generated by either coin, but coin 1 is far more likely.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7000	0.7000	0.3000	0.3000	0.3000	0.3000
1	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
2	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
3	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
4	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
5	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000
6	0.3000	0.5000	0.5000	0.3000	0.3000	0.3000	0.3000

The coin example for $\mathbf{y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$, with p_1 and p_2 initialised to the same value. EM is stuck at a saddle point

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.7001	0.7000	0.3001	0.2998	0.3001	0.2998
1	0.2999	0.5003	0.4999	0.3004	0.2995	0.3004	0.2995
2	0.2999	0.5008	0.4997	0.3013	0.2986	0.3013	0.2986
3	0.2999	0.5023	0.4990	0.3040	0.2959	0.3040	0.2959
4	0.3000	0.5068	0.4971	0.3122	0.2879	0.3122	0.2879
5	0.3000	0.5202	0.4913	0.3373	0.2645	0.3373	0.2645
6	0.3009	0.5605	0.4740	0.4157	0.2007	0.4157	0.2007
7	0.3082	0.6744	0.4223	0.6447	0.0739	0.6447	0.0739
8	0.3593	0.8972	0.2773	0.9500	0.0016	0.9500	0.0016
9	0.4758	0.9983	0.0477	0.9999	0.0000	0.9999	0.0000
10	0.4999	1.0000	0.0001	1.0000	0.0000	1.0000	0.0000
11	0.5000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000

The coin example for $\mathbf{y} = \{\langle HHH\rangle, \langle TTT\rangle, \langle HHH\rangle, \langle TTT\rangle\}$. If we initialise p_1 and p_2 to be a small amount away from the saddle point $p_1 = p_2$, the algorithm diverges from the saddle point and eventually reaches the global maximum.

Iteration	λ	p_1	p_2	$\tilde{p}_1$	$\tilde{p}_2$	$\tilde{p}_3$	$\tilde{p}_4$
0	0.3000	0.6999	0.7000	0.2999	0.3002	0.2999	0.3002
1	0.3001	0.4998	0.5001	0.2996	0.3005	0.2996	0.3005
2	0.3001	0.4993	0.5003	0.2987	0.3014	0.2987	0.3014
3	0.3001	0.4978	0.5010	0.2960	0.3041	0.2960	0.3041
4	0.3001	0.4933	0.5029	0.2880	0.3123	0.2880	0.3123
5	0.3002	0.4798	0.5087	0.2646	0.3374	0.2646	0.3374
6	0.3010	0.4396	0.5260	0.2008	0.4158	0.2008	0.4158
7	0.3083	0.3257	0.5777	0.0739	0.6448	0.0739	0.6448
8	0.3594	0.1029	0.7228	0.0016	0.9500	0.0016	0.9500
9	0.4758	0.0017	0.9523	0.0000	0.9999	0.0000	0.9999
10	0.4999	0.0000	0.9999	0.0000	1.0000	0.0000	1.0000
11	0.5000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000

The coin example for $\mathbf{y} = \{\langle HHH \rangle, \langle TTT \rangle, \langle HHH \rangle, \langle TTT \rangle\}$. If we initialise p_1 and p_2 to be a small amount away from the saddle point $p_1 = p_2$, the algorithm diverges from the saddle point and eventually reaches the global maximum.

Overview

- Maximum-Likelihood Estimation
- Models with hidden variables
- The EM algorithm for a simple example (3 coins)
- The general form of the EM algorithm
- Hidden Markov models

The EM Algorithm

- Θ^t is the parameter vector at t'th iteration
- Choose Θ^0 (at random, or using various heuristics)
- Iterative procedure is defined as

$$\Theta^t = \text{argmax}_\Theta Q(\Theta, \Theta^{t-1})$$

where

$$Q(\Theta, \Theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y \mid x_i, \Theta^{t-1}) \log P(x_i, y \mid \Theta)$$

The EM Algorithm

- Iterative procedure is defined as $\Theta^t = \text{argmax}_\Theta Q(\Theta, \Theta^{t-1})$, where

$$Q(\Theta, \Theta^{t-1}) = \sum_i \sum_{y \in \mathcal{Y}} P(y \mid x_i, \Theta^{t-1}) \log P(x_i, y \mid \Theta)$$

- Key points:
  - Intuition: fill in hidden variables y according to $P(y \mid x_i, \Theta)$
  - EM is guaranteed to converge to a local maximum, or saddle-point, of the likelihood function
  - In general, if

$$\text{argmax}_\Theta \sum_i \log P(x_i, y_i \mid \Theta)$$

has a simple (analytic) solution, then

$$\text{argmax}_\Theta \sum_i \sum_y P(y \mid x_i, \Theta) \log P(x_i, y \mid \Theta)$$

also has a simple (analytic) solution.

Overview

- Maximum-Likelihood Estimation
- Models with hidden variables
- The EM algorithm for a simple example (3 coins)
- The general form of the EM algorithm
- Hidden Markov models

The Structure of Hidden Markov Models

- Have N states, states $1 \ldots N$
- Without loss of generality, take N to be the final or stop state
- Have an alphabet K. For example $K = \{a, b\}$
- Parameter π_i for $i = 1 \ldots N$ is probability of starting in state i
- Parameter $a_{i,j}$ for $i = 1 \ldots (N-1)$, and $j = 1 \ldots N$ is probability of state j following state i
- Parameter $b_i(o)$ for $i = 1 \ldots (N-1)$, and $o \in K$ is probability of state i emitting symbol o

An Example

- Take $N = 3$ states. States are $\{1, 2, 3\}$. Final state is state 3.
- Alphabet $K = \{the, dog\}$.
- Distribution over initial state is $\pi_1 = 1.0, \pi_2 = 0, \pi_3 = 0$.
- Parameters $a_{i,j}$ are

	j=1	j=2	j=3
i=1	0.5	0.5	0
i=2	0	0.5	0.5

- Parameters $b_i(o)$ are

	o=the	o=dog
i=1	0.9	0.1
i=2	0.1	0.9

A Generative Process

- Pick the start state s_1 to be state i for $i = 1 \ldots N$ with probability π_i.
- Set $t = 1$
- Repeat while current state s_t is not the stop state (N):
  - Emit a symbol $o_t \in K$ with probability $b_{s_t}(o_t)$
  - Pick the next state s_{t+1} as state j with probability $a_{s_t,j}$.
  - $t = t + 1$

Probabilities Over Sequences

- An **output sequence** is a sequence of observations $o_1 \ldots o_T$ where each $o_i \in K$
  e.g. the dog the dog dog the
- A **state sequence** is a sequence of states $s_1 \ldots s_T$ where each $s_i \in \{1 \ldots N\}$
  e.g. 1 2 1 2 2 1
- HMM defines a probability for each state/output sequence pair

e.g. the/1 dog/2 the/1 dog/2 the/2 dog/1 has probability

$\pi_1\ b_1(\text{the})\ a_{1,2}\ b_2(\text{dog})\ a_{2,1}\ b_1(\text{the})\ a_{1,2}\ b_2(\text{dog})\ a_{2,2}\ b_2(\text{the})\ a_{2,1}\ b_1(\text{dog}) a_{1,3}$

Formally:

$$P(s_1 \ldots s_T, o_1 \ldots o_T) = \pi_{s_1} \times \left(\prod_{i=2}^{T} P(s_i \mid s_{i-1}) \right) \times \left(\prod_{i=1}^{T} P(o_i \mid s_i) \right) \times P(N \mid s_T)$$

A Hidden Variable Problem

- We have an HMM with $N = 3$, $K = \{e, f, g, h\}$
- We see the following **output sequences** in training data

e g
e h
f h
f g

- How would you choose the parameter values for π_i, $a_{i,j}$, and $b_i(o)$?

Another Hidden Variable Problem

- We have an HMM with $N = 3$, $K = \{e, f, g, h\}$
- We see the following **output sequences** in training data

e g h
e h
f h g
f g g
e h

- How would you choose the parameter values for π_i, $a_{i,j}$, and $b_i(o)$?