COMP 330: Optimization–Expectation Maximization

Chris Jermaine and Kia Teymourian

Rice University

Missing Data

First, a few words...

- EM is a very widely-used MLE algorithm for dealing with missing data
- Perhaps the most intense thing we'll discuss this semester?
- This is some CRAZY S<missing data>!!
- But definitely understandable to a Rice UG
- So pay attention carefully!

Missing Data

Often, one has an optimization problem that would be easy...

- Except that some of the data are missing

Why imght data be missing?

- They were never recorded
- Wrong values recorded
- They are imaginary

Why Can't We Do Something Simple?

Like replace with the mean?

- Back to the regression example
- Want a line to fit points $\langle \mathbf{118}, \mathbf{122}, \mathbf{145}, \mathbf{149}, \mathbf{186}, ?, ? \rangle$
- Mean of observed data is 144
- Does $\langle \mathbf{118}, \mathbf{122}, \mathbf{145}, \mathbf{149}, \mathbf{186}, \mathbf{144}, \mathbf{144} \rangle$ make sense?
- No: given our regression model, we expect larger values

Can't We Just Drop the Data?

In our example, just learn from $\langle 118, 122, 145, 149, 186 \rangle$

- Might make sense here...
- But not in general

Hierarchical Models

In data science, often impossible to drop missing data

Happens with "hierarchical models"

Example:

- Have a bag with two coins
- First has prob p_1 of heads
- Second has prob p_2 of heads
- I repeatedly reach in, pull out a coin
- Identity is $z_i \in \{1, 2\}$
- Flip it 10 times and observe x_i heads
- How to compute $\Theta = \{p_1, p_2\}$

Each z_i is missing: we don't know identity of coin

- How to just drop missing data in this case?

Formal Problem Definition

Formally: we want to compute an MLE for $L(\Theta|x_1, x_2, ..., z_1, z_2, ...)$

- $x_1, x_2, ...$ are observed data
- $z_1, z_2, ...$ are missing data

Note that:

- $L(\Theta|x_1, x_2, ..., z_1, z_2, ...) = f(x_1, x_2, ..., z_1, z_2, ...|\Theta)$
- When the z's are missing, choose Θ to max

$$\int_{\langle z_1, z_2, ...\rangle} f(x_1, x_2, ..., z_1, z_2, ...|\Theta) d\langle z_1, z_2, ...\rangle$$

Formal Problem Definition

Formally: we want to compute an MLE for $L(\Theta|x_1, x_2, ..., z_1, z_2, ...)$

- $x_1, x_2, \ldots$ are observed data
- $z_1, z_2, \ldots$ are missing data

Note that:

- $L(\Theta|x_1, x_2, ..., z_1, z_2, ...) = f(x_1, x_2, ..., z_1, z_2, ...|\Theta)$
- When the z's are missing, choose Θ to max

$$\int_{\langle z_1, z_2, \ldots\rangle} f(x_1, x_2, ..., z_1, z_2, ...|\Theta) d\langle z_1, z_2, ...\rangle$$

- Why is this? Ex: say we have (height, weight) pairs
- Probs are:

$$\langle(\text{short, light}), .3\rangle, \langle(\text{short, heavy}), .1\rangle, \langle(\text{tall, light}), .2\rangle, \langle(\text{tall, heavy}), .4\rangle$$

- Probability they are "tall" is $0.6 = \sum_{\text{weight } w} Pr[(\text{tall}, w)]$
- Easy here, but difficult in the general case!

Expectation Maximization

Is an iterative algorithm for difficult missing-data MLEs

Basic idea...

- Have an estiamte Θ^{iter} for each iteration
- Repeatedly update Θ^{iter} until convergence
- Looks a lot like gradient descent, right?

But EM is unique in how it deals with missing data points

- The famous "Q function"

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = E\left[\log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}}) | x_1, x_2, \ldots, \Theta^{\text{iter}-1}\right]$$

Expectation Maximization

But EM is unique in how it deals with missing data points

- The famous "Q function"

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = E\left[\log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}}) | x_1, x_2, \ldots, \Theta^{\text{iter}-1}\right]$$

- What does this mean?
- Treat $z_1, z_2, \ldots$ as random variables
- With distribution $f(z_1, z_2, \ldots | x_1, x_2, \ldots, \Theta^{\text{iter}-1})$
- Kind of like Bayesian!!
- The Q-function is the expected value of the LLH wrt this distribution

Expectation Maximization

But EM is unique in how it deals with missing data points

- The famous "Q function"

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = E\left[\log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}}) | x_1, x_2, \ldots, \Theta^{\text{iter}-1}\right]$$

What is expected value?

- Recall: exp val of $g(z)$ when z has dist $f(z)$ is $\sum_z f(z)g(z)$ or $\int_z f(z)g(z)dz$
- Example... sample (A, B) from

$$\langle (1,2), .3\rangle, \langle (3,5), .1\rangle, \langle (2,6), .2\rangle, \langle (-3,6), .4\rangle$$

- $E[A+B] = .3 \times (1+2) + .1 \times (3+5) + .2 \times (2+6) + .4 \times (-3+6)$

Expectation Maximization

But EM is unique in how it deals with missing data points

- The famous "Q function"

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = E\left[\log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})|x_1, x_2, \ldots, \Theta^{\text{iter}-1}\right]$$

- If Q is discrete (usually is):

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = \sum_{\langle z_1, z_2, \ldots\rangle} f(z_1, z_2, \ldots|x_1, x_2, \ldots, \Theta^{\text{iter}-1}) \log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})$$

Expectation Maximization

But EM is unique in how it deals with missing data points

- The famous "Q function"

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = E\left[\log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}}) | x_1, x_2, \ldots, \Theta^{\text{iter}-1}\right]$$

In EM, in each iteration:

- Choose Θ^{iter} to maximize the expected value of the log-likelihood

Back to the Example

Best to return to our example...

- Bag with two coins
- $\Theta = \{p_1, p_2\}$: prob of each being heads
- $z_i \in \{1, 2\}$: identity of coin flip the ith time I reach in the bag
- $x_i \in \{1, \ldots, 10\}$: number of heads for the ith trial

So where in the H<missing data> do we start?

Back to the Example

Best to return to our example...

- Bag with two coins
- $\Theta = \{p_1, p_2\}$: prob of each being heads
- $z_i \in \{1, 2\}$: identity of coin flip the ith time I reach in the bag
- $x_i \in \{1, ..., 10\}$: number of heads for the ith trial

So where in the H<missing data> do we start?

- With the likelihood function!

$$\begin{aligned} L(\Theta|x_1, x_2, ..., z_1, z_2, ...) &= f(x_1, x_2, ..., z_1, z_2, ...|\Theta) \\ &= \prod_i f(x_i, z_i|\Theta) \\ &= \prod_i \frac{1}{2}\text{Binomial}(x_i|p_{z_i}, 10) \end{aligned}$$

- Why?

What About the Posterior for the Missing Data

Need $f(z_1, z_2, ...|x_1, x_2, ..., \Theta^{\text{iter}-1}) = \prod_i f(z_i|x_i\Theta^{\text{iter}-1})$

- Use Bayes' rule!

$$f(z_i|x_i\Theta^{\text{iter}-1}) = \frac{f(x_i, z_i|\Theta^{\text{iter}-1})}{f(x_i|\Theta^{\text{iter}-1})}$$
$$= \frac{\frac{1}{2}\text{Binomial}(x_i|p_{z_i}^{\text{iter}-1}, 10)}{f(x_i|\Theta^{\text{iter}-1})}$$

- Not too bad. But what is $f(x_i|\Theta^{\text{iter}-1})$?

What About the Posterior for the Missing Data

Need $f(z_1, z_2, ...|x_1, x_2, ..., \Theta^{\text{iter}-1}) = \prod_i f(z_i|x_i\Theta^{\text{iter}-1})$

- Use Bayes' rule!

$$f(z_i|x_i\Theta^{\text{iter}-1}) = \frac{f(x_i, z_i|\Theta^{\text{iter}-1})}{f(x_i|\Theta^{\text{iter}-1})}$$
$$= \frac{\frac{1}{2}\text{Binomial}(x_i|p_{z_i}^{\text{iter}-1}, 10)}{f(x_i|\Theta^{\text{iter}-1})}$$

- Not too bad. But what is $f(x_i|\Theta^{\text{iter}-1})$?

$$f(x_i|\Theta^{\text{iter}-1}) = \frac{1}{2}\text{Binomial}(x_i|p_1^{\text{iter}-1}, 10) + \frac{1}{2}\text{Binomial}(x_i|p_2^{\text{iter}-1}, 10)$$

- So:

$$f(z_i|x_i\Theta^{\text{iter}-1}) = \frac{\frac{1}{2}\text{Binomial}(x_i|p_{z_i}^{\text{iter}-1}, 10)}{\frac{1}{2}\text{Binomial}(x_i|p_1^{\text{iter}-1}, 10) + \frac{1}{2}\text{Binomial}(x_i|p_2^{\text{iter}-1}, 10)}$$

- Computing $f(z_i|x_i\Theta^{\text{iter}-1})$ requires a pass over the data: called "E-Step"

OK, got the E-Step, Now What?

Now we need the M-Step, where max Q wrt. Θ^{iter}

- Back to the Q function:

$$Q(\Theta^{\text{iter}}, \Theta^{\text{iter}-1}) = \sum_{\langle z_1, z_2, \ldots \rangle} f(z_1, z_2, \ldots | x_1, x_2, \ldots, \Theta^{\text{iter}-1}) \log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}})$$

- Let $c_{i,j}$ denote $f(z_i = j | x_i \Theta^{\text{iter}-1})$
- Write Q function as:

$$\sum_{z_1} \sum_{z_2} \sum_{z_3} \cdots \left(\prod_i c_{i,z_i} \right) \log f(x_1, x_2, \ldots, z_1, z_2, \ldots | \Theta^{\text{iter}})$$

OK, got the E-Step, Now What?

- Write Q function as:

$$\sum_{z_1}\sum_{z_2}\sum_{z_3}\cdots\left(\prod_i c_{i,z_i}\right)\log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})$$

- Where $\log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})$ is:

$$\begin{aligned}\log\prod_i \text{Binomial}(x_i|p_{z_i}^{\text{iter}}) &= \sum_i \log \text{Binomial}(x_i|p_{z_i}^{\text{iter}}) \\ &\propto \sum_i \log\left((p_{z_i}^{\text{iter}})^{x_i} \times (1-p_{z_i}^{\text{iter}})^{10-x_i}\right) \\ &= \sum_i x_i \log(p_{z_i}) + (10-x_i)\log(1-p_{z_i})\end{aligned}$$

OK, got the E-Step, Now What?

- Write Q function as:

$$\sum_{z_1}\sum_{z_2}\sum_{z_3}\cdots\left(\prod_i c_{i,z_i}\right)\log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})$$

- Where $\log f(x_1, x_2, \ldots, z_1, z_2, \ldots|\Theta^{\text{iter}})$ is:

$$\begin{aligned}\log\prod_i \text{Binomial}(x_i|p_{z_i}^{\text{iter}}) &= \sum_i \log \text{Binomial}(x_i|p_{z_i}^{\text{iter}}) \\ &\propto \sum_i \log\left((p_{z_i}^{\text{iter}})^{x_i}\times(1-p_{z_i}^{\text{iter}})^{10-x_i}\right) \\ &= \sum_i x_i\log(p_{z_i}^{\text{iter}}) + (10-x_i)\log(1-p_{z_i}^{\text{iter}})\end{aligned}$$

- Dropping the "iter" on $p_{z_i}^{\text{iter}}$, the Q function is:

$$\sum_{z_1}\sum_{z_2}\sum_{z_3}\cdots\left(\prod_i c_{i,z_i}\right)\sum_i x_i\log(p_{z_i}) + (10-x_i)\log(1-p_{z_i})$$

The M-Step

- Now, we need to maximize:

$$\sum_{z_1}\sum_{z_2}\sum_{z_3}\ldots\left(\prod_i c_{i,z_i}\right)\sum_i x_i\log(p_{z_i})+(10-x_i)\log(1-p_{z_i})$$

- Nasty!! Or is it? Consider just one varible, z_1. Write as:

$$\sum_{z_1}\sum_{\langle z_2,z_3,\ldots\rangle} c_{1,z_1}a(\langle z_2,z_3,\ldots\rangle)\left(\log x_i(p_{z_1})+(10-x_i)\log(1-p_{z_1})+\sum_{i=2}^{n} b(\langle z_2,z_3,\ldots\rangle\right.$$

$$=\sum_{\langle z_2,z_3,\ldots\rangle} c_{1,1}a(\langle z_2,z_3,\ldots\rangle)\left(x_i\log(p_1)+(10-x_i)\log(1-p_1)+\sum_{i=2}^{n} b(\langle z_2,z_3,\ldots\rangle)\right)$$

$$+\sum_{\langle z_2,z_3,\ldots\rangle} c_{1,2}a(\langle z_2,z_3,\ldots\rangle)\left(x_i\log(p_2)+(10-x_i)\log(1-p_2)+\sum_{i=2}^{n} b(\langle z_2,z_3,\ldots\rangle)\right)$$

The M-Step

- Nasty!! Or is it? Consider just one varible, z_1; try to separate out. Write as:

$$
\begin{aligned}
&= c_{1,1} \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \left(x_i \log(p_1) + (10 - x_i) \log(1 - p_1) + \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle) \right) \\
&+ c_{1,2} \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \left(x_i \log(p_2) + (10 - x_i) \log(1 - p_2) + \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle) \right) \\
&= c_{1,1} \left(x_i \log(p_1) + (10 - x_i) \log(1 - p_1) \right) \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \\
&+ \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle) \\
&+ c_{1,2} \left(x_i \log(p_2) + (10 - x_i) \log(1 - p_2) \right) \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \\
&+ \sum_{\langle z_2, z_3, \ldots \rangle} a(\langle z_2, z_3, \ldots \rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots \rangle)
\end{aligned}
\tag{1}
$$

The M-Step

- Nasty!! Or is it? Consider just one varible, z_1; try to separate out. Write as:

$$= c_{1,1}\left(x_i \log(p_1) + (10 - x_i)\log(1 - p_1)\right) \sum_{\langle z_2, z_3, \ldots\rangle} a(\langle z_2, z_3, \ldots\rangle)$$

$$+ \sum_{\langle z_2, z_3, \ldots\rangle} a(\langle z_2, z_3, \ldots\rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots\rangle)$$

$$+ c_{1,2}\left(x_i \log(p_2) + (10 - x_i)\log(1 - p_2)\right) \sum_{\langle z_2, z_3, \ldots\rangle} a(\langle z_2, z_3, \ldots\rangle)$$

$$+ \sum_{\langle z_2, z_3, \ldots\rangle} a(\langle z_2, z_3, \ldots\rangle) \sum_{i=2}^{n} b(\langle z_2, z_3, \ldots\rangle)$$

$$= c_{1,1}\left(x_i \log(p_1) + (10 - x_i)\log(1 - p_1)\right) + \text{other terms w/o } z_1$$

$$+ c_{1,2}\left(x_i \log(p_2) + (10 - x_i)\log(1 - p_2)\right) + \text{other terms w/o } z_1$$

The M-Step

- So in the end, the Q function is simply

$$\sum_i c_{i,1} \left(x_i \log(p_1) + (10 - x_i) \log(1 - p_1) \right) + c_{i,2} \left(x_i \log(p_2) + (10 - x_i) \log(1 - p_2) \right)$$

- When we max (exercise for the board?), we find $p_1 = \sum_i c_{i,1} \times \frac{x_i}{10}$
- And $p_2 = \sum_i c_{i,2} \times \frac{x_i}{10}$
- Very simple!!
- But a long way to get there :-)

Questions?