Exercise Session 5 – Solutions

1 Beta prior

We remember that maximum likelihood estimation of Bernoulli parameters can be easily solved by counting the data points. Using a prior $P(\theta) = Beta(\theta; \alpha, \beta)$ on the parameter, we get the following posterior distribution (remember that Beta distributions are **conjugate priors** for Bernoulli/binomial distributions):

$$P(\theta|D) = \frac{P(D|\theta)Beta(\theta; \alpha, \beta)}{P(D)} = Beta(\theta; \alpha + \#H, \beta + \#T)$$

from which we can compute θ_{Bayes} by computing the mean of the posterior:

$$\theta_{Bayes} = \langle\theta\rangle = \int \theta p(\theta|D)d\theta = \frac{\alpha + \#H}{\beta + \#T + \alpha + \#H}.$$

a) The MLE is:

$$\theta_{MLE} = argmax_\theta P(D|\theta) = \frac{\#H}{\#H + \#T} = \frac{3}{3+7} = 0.3$$

b) The weak prior has little to no effect compared to not giving a prior.

$$\theta_{Bayes} = \frac{0.0001 + 3}{0.0001 + 3 + 0.0001 + 7} = 0.300004$$

c) A very strong prior removes the effect of the observations to a great extent, where having $\alpha = \beta$ shifts the result towards 0.5.

$$\theta_{Bayes} = \frac{10000 + 3}{10000 + 3 + 10000 + 7} = 0.4999$$

2 MAP estimation with a Beta prior

The MAP estimates for a binomial with Beta priors are computed as:

$$\theta_{MAP} = \arg\max_\theta p(\theta|D) = \arg\max_\theta \frac{p(D|\theta)p(\theta)}{p(D)} = \arg\max_\theta p(D|\theta)p(\theta) = \frac{\#H + \alpha - 1}{\#H + \#T + \alpha + \beta - 2}$$

Computing a MAP estimate using a Beta prior is very similar to computing the Bayes estimate as we did in the previous exercise. Instead of computing computing the mean of the posterior Beta distribution, by maximizing we're in fact obtaining its mode (assuming both $\#H + \alpha$ and $\#T + \beta$ to be > 1.)

a) In absence of a prior, we assume $Beta(\theta; 1, 1)$, which is equivalent to the uniform prior $U(0, 1)$, i.e. $p(\theta) = 1$ for $\theta \in [0, 1]$. We see that the $p(\theta)$ term vanishes from the equation above and we are left with simple MLE.

$$\theta_{MAP} = \theta_{MLE} = \frac{\#H}{\#H + \#T} = 0.3$$

b)

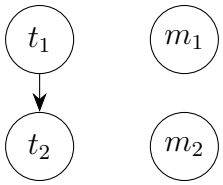
$$\theta_{MAP} = \frac{0.0001 + 3 - 1}{0.0001 + 3 + 0.0001 + 7 - 2} = 0.25$$

c)

$$\theta_{MAP} = \frac{10000 + 3 - 1}{10000 + 3 + 10000 + 7 - 2} = 0.49990$$

3 Missing at random

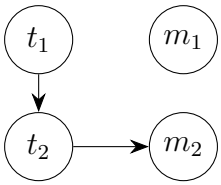
The base network is as follows



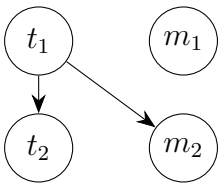
where links have to be added to m_1 and m_2 accordingly.

a) Data is MCAR, so the base network holds and $t_2 \perp\!\!\!\perp m_2$ and $t_1 \perp\!\!\!\perp m_1$.

b) Data is MNAR because the result of the second test determines whether the second measurement is missing and thus $t_2 \not\perp\!\!\!\perp m_2$.



c) Data is MAR as given information on the first test, the second measurement is missing. We can write that $t_2 \perp\!\!\!\perp m_2 | t_1$.



4 Expectation Maximization

4.1 Formula

- In the **E-step**, we're essentially filling in the missing values with all possibilities and weighing those hypothetical configurations in terms of the current parameter estimates.

Filling in the **possibilities**, we get the following table, where the complete data has a weight of 1.

a	b	$weight$
1	1	$q^1(a=1,b=1)=1$
1	**1**	$q^2(a=1,b=1)$
1	**0**	$q^2(a=1,b=0)$
0	0	$q^3(a=0,b=0)=1$
1	0	$q^4(a=1,b=0)$
0	0	$q^4(a=0,b=0)$

Note that we write the q's in terms of all the variables in order to make more sense of Equation 1. Next, we write the unknown weights in terms of the current parameter values, i.e. we compute their **expected** probability given what we currently know about the model.

If a is given and b is guessed, this is simply the previous parameter value.

$$q^2(a=1,b=1) = p(b=1|a=1,\theta^{old}) = \theta^{old}_{b1}$$
$$q^2(a=1,b=0) = p(b=0|a=1,\theta^{old}) = (1-\theta^{old}_{b1})$$

If b is given and a is guessed, we rewrite in terms of parameter values.

$$\begin{aligned} q^4(a=0,b=0) &= p(a=0|b=0,\theta^{old}) \\ &= \frac{p(a=0,b=0|\theta^{old})}{p(b=0|\theta^{old})} \\ &= \frac{p(a=0|\theta^{old})p(b=0|a=0,\theta^{old})}{\sum_a p(a|\theta^{old})p(b=0|a,\theta^{old})} \\ &= \frac{(1-\theta^{old}_a)(1-\theta^{old}_{b2})}{\theta^{old}_a(1-\theta^{old}_{b1})+(1-\theta^{old}_a)(1-\theta^{old}_{b2})} \\ q^4(a=1,b=0) &= p(a=1|b=0,\theta^{old}) = 1-q^4(a=0,b=0) \end{aligned}$$

- For the **M-step**, we compute new estimates of the parameter values, given the new, weighted data. Maximising the expected log likelihood is equivalent to computing the probabilistic count using the weights obtained in the **E-step**. We thus use the following formula from the book/slides:

$$\theta^{new}_x = \frac{\sum_x q^n(x,pa(x))}{\sum_n q^n(pa(x))} \tag{1}$$

Then,

$$\begin{aligned} \theta^{new}_a &= \frac{\sum_n q^n(a=1)}{\sum_n q^n(a=1)+q^n(a=0)} = \frac{2+q^4(a=1,b=0)}{4} \\ \theta^{new}_{b1} &= \frac{\sum_n q^n(a=1,b=1)}{\sum_n q^n(a=1)} = \frac{1+q^2(a=1,b=1)}{2+q^4(a=1,b=0)} \\ \theta^{new}_{b2} &= \frac{\sum_n q^n(a=0,b=1)}{\sum_n q^n(a=0)} = \frac{0}{q^3(a=0,b=0)+q^4(a=0,b=0)}. \end{aligned}$$

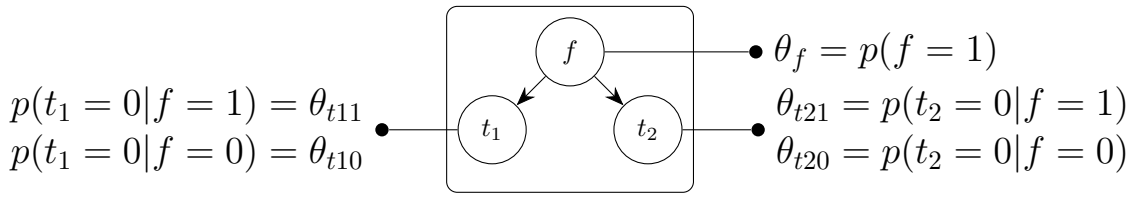
4.2 Algorithm

M-step			E-step			
θ_a	θ_{b1}	θ_{b2}	$q^2(a=1,b=1)$	$q^2(a=1,b=0)$	$q^4(a=1,b=0)$	$q^4(a=0,b=0)$
0.4	0.2	0.3	0.2	0.8	0.432	0.568
0.608	0.493	0	0.493	0.507	0.440	0.560
0.61	0.612	0	0.612	0.388	0.378	0.622

and so on...

5 Yet another EM exercise

a) With f for having the flu and t_1 and t_2 the results of the tests, we get the following network. We can see that the parameters represent probability tables that we're trying to learn from incomplete data.



b) The **completed** data (now using the shorter notation for q^n) is

f	t_1	t_2	*weight*
1	1	**1**	$q^1(t_2 = 1)$
1	1	**0**	$q^1(t_2 = 0)$
1	1	1	1
0	**1**	0	$q^3(t_1 = 1)$
0	**0**	0	$q^3(t_1 = 0)$
1	1	0	1
1	**1**	1	$q^5(f = 1, t_1 = 1)$
1	**0**	1	$q^5(f = 1, t_1 = 0)$
0	**1**	1	$q^5(f = 0, t_1 = 1)$
0	**0**	1	$q^5(f = 0, t_1 = 0)$

with the weights defined in terms of the parameters (**E-step**) as

$$
\begin{aligned}
q^1(t_2 = 1) &= p(t_2 = 1|f = 1, t_1 = 1) \\
&= p(t_2 = 1|f = 1) = 1 - \theta_{t21} && \text{(because } t_1 \perp\!\!\!\perp t_2|f\text{)} \\
q^1(t_2 = 0) &= p(t_2 = 0|f = 1) = \theta_{t21} \\
q^3(t_1 = 1) &= 1 - \theta_{t10} \\
q^3(t_1 = 0) &= \theta_{t10} \\
q^5(f = 1, t_1 = 1) &= p(f = 1, t_1 = 1|t_2 = 1) \\
&= \frac{p(f = 1, t_1 = 1, t_2 = 1)}{\sum_{t_1, f} p(f = 1, t_1 = 1, t_2 = 1)} \\
&= \frac{p(t_1 = 1|f = 1)p(t_2 = 1|f = 1)p(f = 1)}{\sum_f p(t_2 = 1|f)p(f)} && (t_1 \text{ part sums to 1}) \\
&= \frac{(1 - \theta_{t11})(1 - \theta_{t21})\theta_f}{\theta_f(1 - \theta_{t21}) + (1 - \theta_f)(1 - \theta_{t20})}
\end{aligned}
$$

and similarly

$$q^5(f=1,t_1=0) = \frac{\theta_{t11}(1-\theta_{t21})\theta_f}{\theta_f(1-\theta_{t21})+(1-\theta_f)(1-\theta_{t20})}$$
$$q^5(f=0,t_1=1) = \frac{(1-\theta_{t10})(1-\theta_{t20})(1-\theta_f)}{\theta_f(1-\theta_{t21})+(1-\theta_f)(1-\theta_{t20})}$$
$$q^5(f=0,t_1=0) = \frac{\theta_{t10}(1-\theta_{t20})(1-\theta_f)}{\theta_f(1-\theta_{t21})+(1-\theta_f)(1-\theta_{t20})}$$

Next, for the **M-step**, we write the parameters back in terms of the completed data. For example, for θ_f we count the weights of datapoints where $f=1$ (because $\theta_f = p(f=1)$) and divide by those where f is either 0 or 1, i.e. all data weights. We can see that the M-step thus amounts to weighted counting, very similar to classic maximum likelihood.

$$\theta_f = \frac{3 + q^5(f=1,t_1=1) + q^5(f=1,t_1=0)}{5}$$

Similarly for θ_{t11}, we look only at data points where $f=1$ and count the fraction of those where $t_1=0$ versus those with any value for t_1.

$$\theta_{t11} = \frac{q^5(f=1,t_1=0)}{3 + q^5(f=1,t_1=1) + q^5(f=1,t_1=0)}$$
$$\theta_{t10} = \frac{q^3(t_1=0) + q^5(f=0,t_1=0)}{1 + q^5(f=0,t_1=0) + q^5(f=0,t_1=1)}$$
$$\theta_{t21} = \frac{1 + q^1(t_2=0)}{3 + q^5(f=1,t_1=1) + q^5(f=1,t_1=0)}$$
$$\theta_{t20} = \frac{1}{1 + q^5(f=0,t_1=0) + q^5(f=0,t_1=1)}$$

c) After 1 iteration of EM algorithm, we get the follwoing parameter estimates:

- $p(\text{outcome test } 1 = neg | \text{patient has the flu} = no) = p(t_1=0|f=0) \approx \theta^1_{t10} = 0.8$
- $p(\text{outcome test } 1 = neg | \text{patient has the flu} = yes) = p(t_1=0|f=1) \approx \theta^1_{t11} = 0.034$

The Excel file for complete EM algorihtm is attached.