Teresa Chu
CS436: Machine Learning
Homework #2

Homework #2

Declaration:
I have done this assignment completely on my own. I have not copied it, nor have I given my solution to anyone else. I understand that if I am involved in plagiarism or cheating I will have to sign an official form that I have cheated and that this form will be stored in my official university record. I also understand that I will receive a grade of 0 for the involved assignment for my first offense and that I will receive a grade of "F" for the course for any additional offense.
Teresa Chu

Naïve Bayes Answers and Results

1. Results from running program on remote.cs.binghamton.edu
   a.

Training	Test_with_stopwords	Test_without_stopwords
0.9807986609277858	0.9692307692307692	0.9616268788682582

      i. Results from training data had a higher accuracy of around ~0.981 while the test data had slightly lower accuracy with test data with stop words having an accuracy ~0.969 while test data without stop words had a lower accuracy of ~0.961
2. The accuracy of the filtered set without stop words has a slightly lower accuracy then the test set with the stop words. This might be the case because stop words in the case of determining if the email is spam or ham would be useful and not a word that should be filtered out. This is because spam emails usually have typos or incorrect grammar so these emails will typically not have a lot of stop words. Stop words usually appear in emails that seem legitimate and would not be expected to appear in spam emails. Meanwhile ham emails follow grammar rules so will often have these stop words in the email. Therefore, having more stop words is a good indication that this email is a ham email instead of a spam. As a result, by removing these stop words, there is a lower accuracy in determining if the email is legitimate or not.

Point Estimation Answers

1. MLE of parameter p, Bernoulli(p) sample of size n.
   a. Likelihood = $p(x_1, x_2 \ldots . x_n|p) = \prod_{i=1}^{n} \mathrm{p}^{(x_i)}(1-\mathrm{p})^{(1-x_i)} = \mathrm{p}^{(x_1+x_2 \ldots . x_n)}(1-\mathrm{p})^{\mathrm{n}-(x_1+x_2 \ldots . x_n)}$ where x_i is either 0 or 1
   b. Log-likelihood =
      i. $\ln(\mathrm{p}^{(x_1+x_2 \ldots . x_n)}(1-\mathrm{p})^{\mathrm{n}-(x_1+x_2 \ldots . x_n)})$
      ii. $(x_1 + x_2 \ldots . x_n)\ln(\mathrm{p}) + (\mathrm{n} - (x_1 + x_2 \ldots . x_n))(\ln(1-p))$
      iii. $(x_1 + x_2 \ldots . x_n)\ln(\mathrm{p}) + (\mathrm{n})\ln(1-p) - (x_1 + x_2 \ldots . x_n)\ln(1-p)$
   c. MLE =
      i. $\frac{d}{dp}((x_1 + x_2 \ldots . x_n)\ln(\mathrm{p}) + (\mathrm{n})\ln(1-p) - (x_1 + x_2 \ldots . x_n)\ln(1-p))=0$
      ii. $(x_1 + x_2 \ldots . x_n)\frac{1}{p} + (\frac{-n}{1-p}) - (x_1 + x_2 \ldots . x_n)(\frac{-1}{1-p}) = 0$

iii. $\frac{(x_1+x_2....x_n)}{p} + \left(\frac{-n}{1-p}\right) + (\frac{(x_1+x_2....x_n)}{1-p}) = 0$

iv. $\frac{(x_1+x_2....x_n)}{p} = -\left(\frac{(x_1+x_2....x_n)-n}{1-p}\right)$

v. $(x_1 + x_2 \ldots . x_n)(1-p) = -((x_1 + x_2 \ldots . x_n) - n)p$

vi. $(x_1 + x_2 \ldots . x_n) - (x_1 + x_2 \ldots . x_n)p = -(x_1 + x_2 \ldots . x_n)p + np$

vii. $(x_1 + x_2 \ldots . x_n) = np$

viii. $p = \frac{1}{n}(x_1 + x_2 \ldots . x_n)$

2. Parameter p based on a Binomial(N, p) sample of size n. Compute your estimators if the observed sample is (3, 6, 2, 0, 0, 3) and N = 10.
   a. Parameter p based on a Binomial(N, p) sample of size n.
      i. Likelihood = $\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\ \mathrm{p}^{\mathrm{x_i}}(1-\mathrm{p})^{N-\mathrm{x_i}}$
      ii. Log-likelihood =
         1. $\ln\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\ \mathrm{p}^{\mathrm{x_i}}(1-\mathrm{p})^{N-\mathrm{x_i}}\right)$
         2. $\ln\left(\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\right) \mathrm{p}^{\sum_{i=1}^{n} x_i}(1-\mathrm{p})^{\mathrm{Nn}-\sum_{i=1}^{n} x_i}\right)$
         3. $\ln\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\right) + \ln\left(\mathrm{p}^{\sum_{i=1}^{n} x_i}\right) + \ln\left((1-\mathrm{p})^{\mathrm{Nn}-\sum_{i=1}^{n} x_i}\right)$
         4. $\ln\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\right) + \ln\left(\mathrm{p}^{\sum_{i=1}^{n} x_i}\right) + \ln\left((1-\mathrm{p})^{Nn-\sum_{i=1}^{n} x_i}\right)$
         5. $\ln\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\right) + \sum_{i=1}^{n} x_i \ln(\mathrm{p}) + (\mathrm{Nn} - \sum_{i=1}^{n} x_i)\ln((1-\mathrm{p})$
      iii. MLE
         1. $\frac{d}{dp}\left(\ln\left(\prod_{i=1}^{n} \frac{N!}{(N-x_i)!x_i!}\right) + \sum_{i=1}^{n} x_i \ln(\mathrm{p}) + (Nn - \sum_{i=1}^{n} x_i)\ln((1-\mathrm{p})\right) = 0$
         2. $\frac{d}{dp}\left(\sum_{i=1}^{n} x_i \ln(\mathrm{p}) + (Nn - \sum_{i=1}^{n} x_i)\ln((1-\mathrm{p})\right) = 0$
         3. $\frac{1}{p}\sum_{i=1}^{n} x_i + (Nn - \sum_{i=1}^{n} x_i)\frac{-1}{(1-\mathrm{p})} = 0$
         4. $\frac{1}{p}\sum_{i=1}^{n} x_i = (Nn - \sum_{i=1}^{n} x_i)\frac{1}{(1-\mathrm{p})}$
         5. $(1-p)\sum_{i=1}^{n} x_i = (Nn - \sum_{i=1}^{n} x_i)\mathrm{p}$
         6. $\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i = (Nn\mathrm{p} - \mathrm{p}\sum_{i=1}^{n} x_i$
         7. $\sum_{i=1}^{n} x_i = Nnp$
         8. $p = \left(\frac{1}{Nn}\right)\sum_{i=1}^{n} x_i$
   b. Estimators if the observed sample is (3, 6, 2, 0, 0, 3) and N = 10.
      i. $p = \left(\frac{1}{Nn}\right)\sum_{i=1}^{n} x_i = \left(\frac{1}{10*6}\right)\sum_{i=1}^{6} x_i$
      ii. $p = \left(\frac{1}{60}\right)(3+6+2+0+0+3)$
      iii. $p = \left(\frac{14}{60}\right)$
3. parameters a and b based on a Uniform (a, b) sample of size n.
      i. Likelihood = = $\prod_{i=1}^{n} \frac{1}{b-a} = \frac{1}{(b-a)^n}$ *where* $a \leq \min(x_i \ldots x_n)$ *and* $b \geq \max(x_i \ldots x_n)$ *since* $f(x) = 0$ *if* x *not in bounds* $a \leq x \leq b$
      ii. Log-likelihood = $\ln(\frac{1}{(b-a)^n}) = \ln((b-a)^{-n}) = -n\ ln(b-a))$

iii. MLE of parameter a
   1. $\frac{d}{da}(-n\ ln(b-a))$=0
   2. $\frac{-n}{b-a} * -1$=0
   3. $\frac{n}{b-a} = 0$
   4. a = $\min(x_1 \dots x_n)$ since we want to get the largest a where a≤ $\min(x_i \dots x_n)$ to maximize $\frac{n}{b-a}$ which is our MLE

iv. MLE of parameter b
   1. $\frac{d}{db}(-n\ ln(b-a))$=0
   2. $\frac{-n}{b-a} = 0$
   3. $\frac{-n}{a-b} = 0$
   4. b = $\max(x_1 \dots x_n)$ since we want to get the smallest b where b≥ $\max(x_i \dots x_n)$ to maximize $\frac{n}{b-a}$ which is our MLE

4. parameter μ based on a Normal(μ, σ^2) sample of size n with known variance σ^2 and unknown mean μ.
   a. Likelihood = $\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$
   b. Log- Likelihood
      i. $\ln\left(\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$
      ii. $\ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right)$
      iii. $\ln\left(\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \ln(e)$
      iv. $n\ln(1) - n\ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
      v. $-n\ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
      vi. $-n\ln(\sigma) - n\ln\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
      vii. $-n\ln(\sigma) - \left(\ln\left(\sqrt{2\pi}\right)\left(\frac{n}{2}\right)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
   c. MLE
      i. $\frac{d}{d\mu}\left(-n\ln(\sigma) - \left(\ln\left(\sqrt{2\pi}\right)\left(\frac{n}{2}\right)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$= 0
      ii. $\frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$=0
      iii. $\left(-\frac{1}{2\sigma^2} * 2\sum_{i=1}^{n}(x_i-\mu) * -1\right) = 0$
      iv. $\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$
      v. Since $\frac{1}{\sigma^2} \neq 0, \sum_{i=1}^{n}(x_i-\mu) = 0$
      vi. $(x_1-\mu) + (x_2-\mu) + \dots (x_n-\mu) = 0$
      vii. $\sum_{i=1}^{n}(x_i) - \mu n = 0$
      viii. $\sum_{i=1}^{n}(x_i) = \mu n$
      ix. $\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i)$

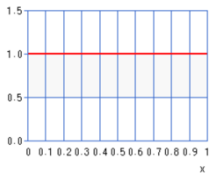
5. parameter σ based on a Normal(μ, σ 2) sample of size n with known mean μ and unknown variance σ 2 .
    a. Likelihood = $\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$
    b. Log- Likelihood = $-n\ln(\sigma) - (\ln(\sqrt{2\pi})\left(\frac{n}{2}\right)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
    c. MLE
        i. $\frac{d}{d\sigma}(-n\ln(\sigma) - (\ln(\sqrt{2\pi})\left(\frac{n}{2}\right)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) = 0$
        ii. $-\frac{n}{\sigma} - 0 - \frac{1}{2} * -2 * \sigma^{-3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$
        iii. $-\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$
        iv. $\frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$
        v. $\sigma^2 n = \sum_{i=1}^{n}(x_i-\mu)^2$
        vi. $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$
        vii. $\sigma = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^{\frac{1}{2}}$
6. parameters (μ, σ^2) based on a Normal(μ, σ ^{2}) sample of size n with unknown mean μ and variance σ^2.
    a. Likelihood = $\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$
    b. Log- Likelihood = $-n\ln(\sigma) - (\ln(\sqrt{2\pi})\left(\frac{n}{2}\right)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$
    c. MLE
        i. To get parameters of (μ, σ^2)
            1. $\frac{d}{d\mu}(-n\ln(\sigma) - (\ln(\sqrt{2\pi})\left(\frac{n}{2}\right)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) = 0$
            2. $\frac{d}{d\sigma}(-n\ln(\sigma) - (\ln(\sqrt{2\pi})\left(\frac{n}{2}\right)) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2) = 0$
        ii. Solving for μ, we see that $\frac{d}{d\mu}(Log - likelihood)$=0 only when $\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i)$
        iii. Solving for σ^2, we see that $\frac{d}{d\sigma}(Log - likelihood)$= 0 when $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$
    d. So parameter pair is $(\frac{1}{n}\sum_{i=1}^{n}(x_i), \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2)$

Coin and Thumbtack Answers

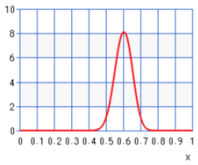
1. Derive the MLE and MAP estimates for the coin and the thumbtack
    a. MLE
        i. Likelihood of coin
            1. $\theta = \frac{\alpha_H}{\alpha_H+\alpha_T}$
            2. $\theta = \frac{60}{60+40} = \frac{60}{100} = \frac{3}{5}$
        ii. Likelihood of thumbtack
            1. $\theta = \frac{\alpha_H}{\alpha_H+\alpha_T}$

2. $\theta = \frac{70}{70+30} = \frac{70}{100} = \frac{7}{10}$

b. MAP
   i. Coin
      1. Beta(1,1)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{60+1-1}{100+2-2} = \frac{6}{10}$
      2. Beta(40, 60)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{60+40-1}{100+100-2} = \frac{99}{198}$
      3. Beta(30, 70)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{60+30-1}{100+100-2} = \frac{89}{198}$
   ii. Thumbtack
      1. Beta(100, 100)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{70+100-1}{100+200-2} = \frac{169}{298}$
      2. Beta(1000, 1000)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{70+1000-1}{100+2000-2} = \frac{1069}{2098}$
      3. Beta(100,000, 100,000)
         a. $\theta = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + B_T - 2}$
         b. $\theta = \frac{70+100,000-1}{100+200,000-2} = \frac{100,069}{200,098}$

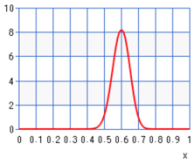
2. With the help of figures identify how the different priors affect the estimated parameter values.
   a. Coin
      i. With prior: Beta(1,1)
         1. Prior has little effect on the estimated parameter values since the prior is so weak with $B_H = 1$ and $B_T = 1$ in comparison to the data which has 60 Heads and 40 tails. Therefore, the curve looks more similar to the data curve even with the input of the prior
         2. Prior Beta(1,1)
            a. 

               i. The curve appears as a line given that there are so few samples with $B_H = 1$ and $B_T = 1$
         3. Data Beta(60,40)

a. 

i. The curve appears centered slightly to the right because data has 60 Heads and 40 tails.

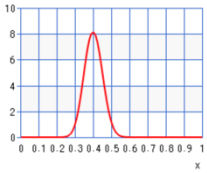
4. Posterior Beta(61,41)

a. 

i. The curve appears almost identical to the data curve it has similar values to the data given that the prior is so weak.

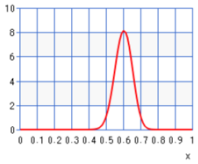
ii. With prior: Beta(40,60)

1. Prior has a strong effect on the estimated parameter values since values of the prior is similar to the data with B_H =40 and B_T =60 in comparison to the data which has 60 Heads and 40 tails. With this input, the curve is now much more centered and narrower than the likelihood curve
2. Prior Beta(40,60)

a. 

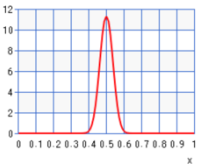
i. The curve appears centered slightly to the left because data has 40 Heads and 60 tails.

3. Data Beta(60,40)

a. 

i. The curve appears centered slightly to the right because data has 60 Heads and 40 tails.

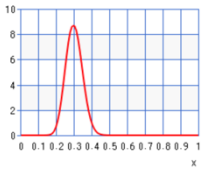
4. Posterior Beta(100,100)

a. 

i. The posterior curve appears centered and narrower because data has 100 Heads and 100 tails.

iii. With prior: Beta(30,70)

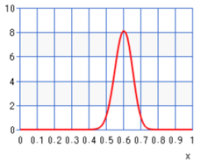
1. Prior has a strong effect on the estimated parameter values since values of the prior is similar to the data with B_H =30 and B_T =70 in comparison to the data which has 60 Heads and 40 tails. With this input, the curve is more centered but leaning to the left due to there being a higher number of tails. In addition with more samples, the curve becomes more narrow then the likelihood curve
2. Prior Beta(30,70)



a.

i. The prior curve appears centered slightly to the left because prior data has 70 Heads and 30 tails.

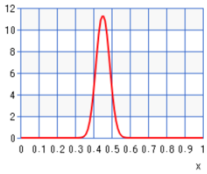
3. Data Beta(60,40)



a.

i. The curve appears centered slightly to the right because data has 60 Heads and 40 tails.

4. Posterior Beta(90,110)



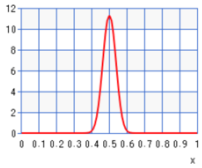
a.

i. The posterior curve appears centered slightly to the left because the posterior has 90 Heads and 110 tails. In comparison to the 2 curves above, it is also narrower since it has more data

b. Thumbtack

i. With prior: Beta(100, 100)

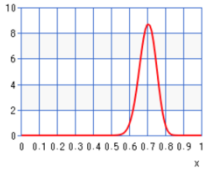
1. Prior has a very strong effect on the estimated parameter values since there is more prior data in comparison to the data with B_H =100 and B_T =100 while data only has 70 Heads and 30 tails. With this input, though the curve is to right, it is more centered due to the prior data.
2. Prior Beta(100,100)



a.

i. The prior curve appears centered and narrow because data has 100 Heads and 100 tails.

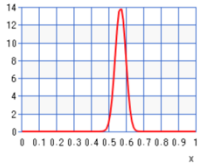
3. Data Beta(70,30)



a.

i. The curve appears centered to the right because data has 70 Heads and 30 tails.

4. Posterior Beta(170,130)



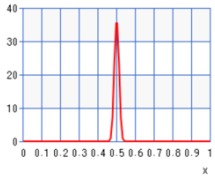
a.

i. The curve appears more to center though slightly to the right because data has 170 Heads and 130 tails.

ii.

ii. With prior: Beta(1000, 1000)

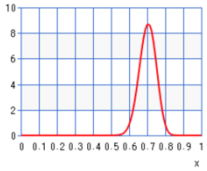
1. Prior has a very strong effect on the estimated parameter values since there is so much more prior data in comparison to the data with B_H =1000 and B_T =1000 while data only has 70 Heads and 30 tails. With this input, the data has little effect on the posterior and the posterior resembles the prior more
2. Prior



a.

i. The prior curve appears centered and narrower in comparison to previous prior curves because data consists of 1000 Heads and 1000 tails.

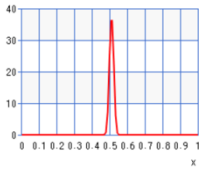
3. Data



a.

i. The curve appears centered to the right because data has 70 Heads and 30 tails.

4. Posterior

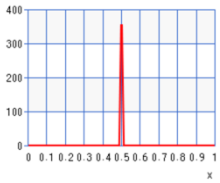


a.

i. The posterior curve appears almost identical to the prior curve since data has so little effect given it only has 70 Heads and 30 tails while prior is 1000 Heads and 1000 tails.

iii. With prior: Beta(100,000, 100,000))

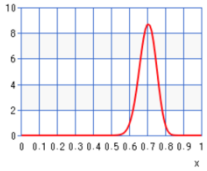
1. Prior has a very strong effect on the estimated parameter values since there is so much more prior data in comparison to the data with B_H =100,000 and B_T =100,000 while data only has 70 Heads and 30 tails. With this input, the data has very little effect on the posterior and resembles the prior
2. Prior



a.

i. The prior curve appears centered and narrower in comparison to previous prior curves because data consists of 100,000 Heads and 100,000 tails.

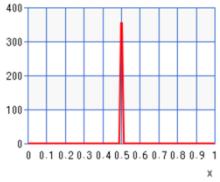
3. Data



a.

i. The curve appears centered to the right because data has 70 Heads and 30 tails.

4. Posterior



a.

i. The posterior curve appears almost identical to the prior curve since data has so little effect given it only has 70 Heads and 30 tails while prior is 100,000 Heads and 100,000 tails.

3. False. MLE estimate will not approach MAP estimate even if more data is collected for MLE since MAP estimate utilizes priors. If MAP however collected more data, it would reach MLE since prior becomes weak causing the data to dominate. Therefore, MAP estimate will appear more like the MLE estimate.
4. True. MLE estimates can be different but MAP estimates can be the same when a large prior is used. Similar to how if data becomes infinite, MAP approaches MLE since data becomes the dominate factor for θ, if the prior is very strong, the data does not play a large role. Therefore since MAP estimate utilizes mostly prior instead of data, the coin and thumbtack can have similar MAP estimates despite MLE estimates being different.