


COURSEWORK

AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

QUANTUM LEAP AFRICA

Foundations of Machine Learning

Instructions

This assignment has both writing and coding components. The coding needs to be in Python. You should submit the code as a separate file. The code must compile on a standard Linux installation.
You are **not** permitted to use any symbolic manipulation libraries (e.g., `sympy`) or automatic differentiation tools (e.g., `tensorflow`) for your submitted code (though, of course, you may find these useful for checking your answers). Your code will be checked for imports. You should not need to import anything other than `numpy` for the submitted code for this assignment.
The writing assignment requires plots, which you can create using any method of your choice. You should not submit the code used to create these plots.
No aspect of your submission may be hand-drawn. You are strongly encouraged to use LaTeX to create the written component.
In addition to this document a markscheme is provided as well as template files for both the writing and coding components.
In summary, you are required to submit the following:

- A file `write_up_<your-name>.pdf` for your written answers.
- A file `coding_answers_<your-name>.py` that implements all the methods for the coding exercises.

Please create a zip-file and send it to `mdeisenroth@aimsammi.org`.
The submission deadline is **October 16, 23:59**.

1 Discrete Models

Andrew is the hot-seat contestant on the popular TV dating show, *Probabilistic Attraction*. In the first round of this game show, he is presented with a hat containing 10 balls, all of which are either black or white. He then listens to three generative processes described by three challengers, Bethany, Charlotte and Davina. After considering their arguments and taking a sample of balls from the hat, he is able to choose which challenger he will choose to take out for dinner.
Bethany is up first:
My model may be simple but I think you'll find it effective. I flipped a fair coin: if it came up two heads I put in 10 white balls; if it came up heads I put in 10 black balls; otherwise I put in 5 of each. I think you'll find my posterior is so sharply peaked you can even take the MAP for making predictions and you'll hardly notice the difference.
Next is Charlotte:
To select the balls for you I rolled a fair die, subtracted one and doubled the result to choose the number of white balls for you. The rest I made up with black balls. I hope youll find my model has a good trade-off between power and flexibility
Finally, Davina gets her turn to describe a generative process:
I selected number of black balls uniformly at random across all possible options and filled the rest with white balls. You find my model the most expressive, and therefore surely the best.
None of the girls nor the contestant know the true generative process.

a) Write a python function `predD(nTotal, nWhite)` that returns the posterior over the number of white balls in the bag for Davina's model, given a sample of size `nTotal`, of which `nWhite` are white. Your function should return a vector of length 11 where the ith element is the probability that the ball contains i white balls.

   - Returns a probability distribution over 11 values for all valid inputs (you don't need to check the inputs for validity)
   - Correct answer. Will be checked only for small (< 30) values

b) Write a python function evidence `evidenceC(nTotal, nWhite)` that returns the model evidence for Charlotte's model.

   - Returns a positive scalar for all valid inputs
   - Correct answer. Will be checked only for small (< 30) values

   Andrew decides to select 20 balls (with replacement).

c) Show the evidence for each model for every possible data set (i.e. all the 21 possible outcomes) using a scatter graph. Use the same graph for each. Don't use a log-scale or join up the points.

   - Scatter graph with three sets of points, clearly labeled (use colors, don't bother making it readable in B&W)

- All correct

Andrew selects 20 balls (with replacement) and 13 are white.

d) Make a table of the posterior probabilities over the 11 different possibilities under each model, given the observed data. State also the predictive distribution of the next ball using the full posterior and also the plug-in MAP estimate for each model.

- Table with three probability distributions over 11 values. Sensible use of standard form where appropriate (no more than 4 significant digits)
- All correct probabilities
- Posteriors all correct for p(next is white—data)
- MAP estimates all correct

e) Justify the claims of the three girls in light of the data. In your answer, you should compare the model evidence for each, and discuss how the situation would have been different if Andrew had drawn 130 white balls out of 200.

- A point about Bethany's claim with brief justification
- A point about Charlotte's claim with brief justification
- A point about Davina's claim with brief justification

2 Continuous Models

It is round two of *Probabilistic Attraction* and this time it's Elizabeth in the hot-seat. For this round, she is faced with a random number generator and three contestants, Fred, George and Harry. It is known that the random numbers x_i will be independent samples from an unknown univariate normal distribution $\mathcal{N}(x_i|\mu,\sigma^2)$. The contestants vie for her affection by presenting their probabilistic models for her consideration.
It is Fred's turn first:
I've gone for a simple but effective option: I'm going to stick to just the Gaussian likelihood model and leave the parameters entirely free. You can just update them when you see the data. You'll find my way always wins in the end
Next it's George
I've been watching this show for years and I've a hunch that the mean of these number is going to be positive. I'd say around 10, but I can't be absolutely sure so I'm going to place a Gaussian prior on μ with mean 10 and a fairly large variance of 25. You'll have to find σ^2 with maximum likelihood
Finally it's Harry's turn to describe his model
I've also been watching this show for a while and I'm pretty sure the numbers are going to be quite spread out. I'm going to add this to my model by placing a inverse gamma prior on σ^2, with shape 2 and scale 5.

a) Derive Fred's maximum likelihood solution for the parameters of his model.

4 marks for

- Explanation of what you are doing
- A brief discussion of why you can take logs
- Tidy algebra and correct notation
- Clear derivation of result

b) Using only the properties of expectation and variance of linear combinations of iid random variables, show that the maximum likehood estimator σ^2_{ML} is biased (that is, taking the expectation of the maximum likelihood estimator with respect $\mathcal{N}(\mu, \sigma^2)$ does not return σ^2)

3 marks for

- Using the results for iid random variables
- Using the definition of variance to compute the expectation of the quadratic terms
- Correct argument, clearly explained

c) Derive the MAP solution for μ in George's model. Do your analysis replacing 10 with μ_0 and 25 with σ_0^2 (this makes it easier to read). Write your answer in terms of Fred's maximum likelihood solution μ_{ML}.

- Statement of log joint
- Correct derivative
- Correct answer
- Written in terms of σ_{ML}

d) Explaining your reasoning, calculate the posterior for George's model. Show that the MAP point you calculated in the previous exercise is also the mean, and give a reason why this is true in this example but not true in general. Again use μ_0 and σ_0^2 rather than use the actual numbers.

- Statement of why only the exponent is necessary
- Completing the square
- Only relevant terms kept, with explanation
- Correct answer
- Explanation why μ_{post} is the same as μ_{MAP}

e) Derive the MAP estimate for Harry's model. Use a and b for the shape and scale respectively instead of the numbers. Write the your answer for σ^2_{MAP} in terms of Fred's maximum likelihood result σ^2_{ML}. NB you might find it easier to work with (and differentiate with respect to) σ^2 rather than σ.

- Log joint written down with irrelevant terms dropped
- $\log \sigma^2$ and $\frac{1}{\sigma^2}$ terms collected
- Derivative taken and maximum found
- written in terms of σ^2_{ML}

f) Derive Harry's posterior distribution. You may reuse some of your working from your previous answer. State also the posterior mean and explain why it is not equal to the MAP estimate you found in the previous part. You may use standard results for the mean of the Inverse Gamma distribution.

- Joint distribution stated as a function of σ^2 with all irrelevant terms dropped. NB you may as well exponentiate the log joint from the previous answer
- Explanation of what this implies about the posterior
- Posterior distribution stated
- Posterior mean stated
- Explanation of disparity of MAP and posterior mean

g) Explain what would happen to the result of inference in the three models if Elizabeth was to take a very large sample from the random number generator

- Statement of what happens to the three estimates
- Brief explanation for Fred's model
- Brief explanation of taking the limit for George's model. You may use standard results for the mean and variance of the Inverse Gamma