Tutorial 5 - Quiz 1 review

Q1

In the equation P(h|y) = P(y|h)P(h)/ P(y) what is the name of the distribution P(y|h)?

(a) **the likelihood**
(b) the prior
(c) the posterior
(d) the posterior mode

Notes:

Basically everyone got this.

Q2

What is maximum a posteriori (MAP) estimation?

(a) we choose the parameters that maximize the likelihood
(b) **we choose the parameters that maximize p(theta|y)**
(c) we choose parameters by taking the posterior mean
(d) we choose parameters by consensus in an anarcho-syndicalist commune

Notes:

Almost everyone chose (b).

Q3

Chapter 2 of L&F described a model that can be used to predict accuracy and reaction time in speeded choice tasks. Which of the following statements are true of this model? Suppose we're using it to model behavior in the Smith and Vickers task, also discussed in F&L.

(a) The response criterion (“criterion” in listing 2.1) determines the proportion of “left” versus “right” responses.
(b) **The drift term biases responses towards one direction or another.**
(c) When the standard deviation of the noise is higher than the criterion, we should expect to see few errors.
(d) When the drift rate is 0.5 there is no evidence to drive the decision process and the decision will be entirely random.

Notes:

Not many mistakes on this one.

(a) false (70 percent correct)
(b) true (90 percent correct)
(c) false (80 percent correct)
(d) false (100 percent correct)

Q4

Imagine that we have fitted a shifted Weibull to some reaction time data, using MLE on individual judgments from several participants. Which of the following statements are true?

(a) The fitted shift parameter of the Weibull distribution will correspond to the most frequent reaction time time in our data.
(b) **The fitted shift parameter of the Weibull can never be larger than the smallest reaction time in our data set.**
(c) The shifted Weibull will not account for the data well, since the Weibull distribution cannot account for skewed data.
(d) **Using a Weibull distribution is compatible with the idea that there are two separate accumulators that are “racing” toward thresholds.**

Notes:

(a) false (30 percent correct) is not true in general, but there are some cases where it could be. There are some parameterizations of the Weibull distribution that have monotonically decreasing density. In these specific cases, it
(b) true (50 percent correct) is true because the support of the original Weibull distribution is positive; anything less than the threshold has probability zero. The answer might be different if we weren’t using individual judgments.
(c) false (70 percent correct) The Weibull distribution can account for skewed data.

(d) (100 percent correct) this is a fact about the Weibull distribution that is mentioned in the book. For more about this property of the Weibull distribution, see Berger, James O.; Sun, Dongchu (1993). "Bayesian Analysis for the Poly-Weibull Distribution" (page 2 of the pre-print, section 1.2).

Q5

We have fitted a linear regression using MSE to several data points. Which of the following statements are true?

(a) Errors contribute to MSE in direct proportion to their size, e.g., adding an error of magnitude 2 will influence the MSE twice as much as adding an error of magnitude 1.
(b) **The RMSE reflects how much our predictions deviate from the true values.**
(c) The model parameters found by minimizing the RMSE will always correspond to the MLE.
(d) If we obtain a low RMSE we can be confident that our model will generalize to new data.

Notes

(a) false (90)
(b) true (90)
(c) false (60) this was a tricky one, and part of the point in asking it was to make sure that an earlier point in lecture didn't lead anyone astray. RMSE-minimizing estimates will correspond to MLEs when the mean of the posterior and the mode of the likelihood are the same (as it is when we have an improper uniform prior on our coefficients and a Gaussian or other symmetric likelihood)
(d) false (80); consider the case of a 9th-order polynomial fitted to 10 data points via MLE.

Q6

To find parameters of our model that best describe our data we can:

(a) Maximize the RMSE between the data and the model's predictions.
(b) **Maximize the likelihood of the data given the model.**
(c) Minimize the log likelihood of the data given the model.
(d) Dephlogisticate the quaternions.

Notes:

(a) false (80)
(b) true (80)
(c) false (50) missing "negative"
(d) false (100) I don't know what this means

Q7

Wikipedia offers the following example to illustrate the base rate fallacy: Suppose a breathlyzer test (b=1 for positive result, b=0 for negative) to detect drunk drivers (d=1 for drunk, d=0 for sober) has a false positive rate of 0.05, that is, P(b=1|d=0) is 0.05 and never fails to detect a true case where d=1, that is, P(b=1|d=1) is 1. Further, we know that 1 in 1000 drivers subjected to the test is drunk. The fallacy involves neglecting the "base rate" of drunk drivers when making probability judgments.

What is the actual probability that a particular driver is drunk, if the breathalyzer returns a positive result? If we express this as u/z, what's u?

(a) 1
(b) 0.95
(c) **.001**
(d) .999
(e) .001 * 0.95

Notes:

(c) (90) this is just the product of the prior probability and the likelihood.

Q8

What is z?

(a) 1
(b) (.001 * .95)+(.999 * .05)
(c) **(.999 * .05)+(.001 * 1)**
(d) .999 * .05
(e) .001 * .95

Notes

(c) (80) As question 7 above, but for both possible outcomes.

Q9

If we imagine that the true drunk/sober state (d=1 or d=0) of the driver is a binary parameter to be estimated from the data we have observed, which of the following estimates, applied in a simple way, would commit the base rate fallacy?

(a) MAP estimate
(b) **maximum-likelihood estimate**
(c) posterior expectation of d given on the test result (E[d|b=1])
(d) ignoring the data and relying on the prior

Notes:

(a) false (70 percent correct)
(b) true (70)
(c) false (70)
(d) false (80)

Q10

Which of the following models of planetary motion provides the best explanation of the observed trajectories of the planets, assuming by "best" we mean the simplest explanation?

(a) Ptolemaic model
(b) **Copernican model**
(c) Keplerian model

Notes:

(b) true (70) Epicycles etc. make the Ptolomaic model more complex (unless you think the concept of heliocentric orbits is inherently very complex.

Q11

Which of the following models of planetary motion provides the best fit with the observed trajectories?

(a) Ptolemaic model
(b) Copernican model
(c) **Keplerian model**

Notes:

(c) (70) Kepler's model isn't the simplest, but it gives great fits.

Q12

In which of the following situations might we need a data model?

(a) **We have a model that predicts the most likely judgment a person will make.**
(b) **We have a model that gives the probabilities that an agent relying on Bayesian inference would assign to different choices being correct.**
(c) We have a model that predicts the utilities that people will assign to options. It assumes they will make choices from these utilities via "soft maximization".
(d) We have a model that assigns likelihoods to different choices, conditional on parameters in the model.

Notes

(a) true (60) The single most-likely judgment doesn't give us a full distribution. We'll get zeroes, which will be a problem.
(b) true (80) Even though this is a distribution, it doesn't necessarily correspond to behavior. The minimial data model (or decision rule) in this case is probability matching, but we can generalize w/softmax, just as with utilities. In this case, however, the softmax "temperature" parameter is more interpretable when equal to 1.
(c) false (70).
(d) false (50).

Q13

If we learn that the maximum of a likelihood function p(d|theta) (i.e., L(theta|x)) is not unique, which of the following statements might we expect to be true?

(a) **theta is not identifiable**
(b) the model is highly sensitive to theta
(c) trick question! The maximum of the likelihood function must be unique
(d) it is impossible to make inferences about theta

Notes

(a) true (50) if our prior is uniform (as is fairly common) then we will never be able to distinguish between values of theta at the maxima of the likelihood function.
(b) false (90) we would expect the opposite to be true
(c) false (80) the likelihood doesn't need to have unique maxima. For example, imagine that theta defines the phase of a sinusoid; there will be infinite equally likely maxima in matches to data.
(d) false (80) we can learn something about theta even if we can't pin it down uniquely. For example, finding an equivalence class of phases (above) could be useful.

Q14

In the context of comparing models with the AIC and the likelihood-ratio test, how do we measure the complexity of a model?

(a) As the number of lines of code that implement the model
(b) As the number of data points used to fit the model
(c) As the number of concepts required to define the model
(d) **As the number of free parameters in the model**

Notes:

(d) (90)

Q15

What is the derivative of w_1*x_1 + w_2*x_2 with respect to x_1 ?

(a) w_1*x_1
(b) w_2*x_2
(c) **w_1**
(d) x_1
(e) other

Notes:

(c) (90)

Q16

What is $log(c * exp((x - u)^2/s)$ if all terms are positive?

(a) $c + (x - u)^2/s$
(b) $\log(c) + 2\log(x - u) - \log(s)$
(c) $\log(c) + (x - u)^2/s$
(d) other:
(e) Don't know

Notes:

(c) (90)

Q17

Which of these inequalities suggests we have made a mistake, if they apply to our results or inferences?

(a) $p(x) > 1$
(b) $P(x) > 1$
(c) $-\log(P(x)) < 0$
(d) $\int_{x \in X} p(x)dx < 1$
(e) $\int_{x \in X} xp(x)dx < 1$

Notes:

(a) false (50) densities can be arbitrarily high
(b) true (80) probability masses cannot exceed 1
(c) true (70) if probability masses cannot exceed 1, their logs cannot be positive, so negative logs cannot be negative
(d) true (50) this integral must be 1
(e) false (90) this integral can be any value; imagine adding or subtracting a value to all x