PSYCH-GA.2211/NEURL-GA.2201 – Fall 2020
Mathematical Tools for Neural and Cognitive Science

Homework 5

Due: 25 Nov 2020
(late homeworks penalized 10% per day)

See the course web site for submission details. For each problem, show your work - if you only provide the answer, and it is wrong, then there is no way to assign partial credit! And, please don't procrastinate until the day before the due date... *start now*!

1. **Comparing two estimators**. In World War II, Allied data analysts needed to estimate the number of German tanks based on daily surveillance reports of the serial numbers of observed tanks. Consider a simple version of this problem, where there are N tanks numbered from 1 to N (with N unknown), and K daily observations are drawn independently and uniformly from this set (with replacement - a tank might be observed more than once).

   (a) Examine the maximum likelihood estimator (MLE). The likelihood function of N is the probability of observing a given set of K values, drawn independently from the uniform distribution, $p(\{n_1, n_2, n_3, ...n_K\} \mid N) = (1/N)^K$, for $1 \leq n_k \leq N$. This expression decreases with increasing N, so it will be maximized by choosing N as small as possible. Thus, the MLE is simply the largest of the observed serial numbers: $\hat{N}_{\text{MLE}} = \max(\{n_1, n_2, n_3, ...n_K\})$. To examine the behavior of this estimator, we need to calculate the distribution of the maximum of K observed serial numbers. The probability that an observation is less than or equal to any given value n is just (n/N), so the probability that K (independent) observations are all less than or equal to n is $(n/N)^K$. If we place these values for different n into a vector, $c(n)$, this will represent the *cumulative* distribution of the max, and the difference, $c(n) - c(n-1)$, gives the probability that the maximum value is equal to n. Compute this distribution for $N = 40$ and $K = 12$, and plot it as a bar plot (use `bar()`). What is the probability that the maximum value is equal to N? What is are the mean and standard deviation of the distribution of maximum values?

   (b) For comparison, consider an intuitive estimator derived from the sample average. The mean of the distribution of observations (conditioned on N) is $(N + 1)/2$, we can estimate N by taking twice the sample average minus 1. Using the expressions for the mean and standard deviation of a uniform distribution, and the fact that the standard deviation of a sample average falls with the number of samples as $1/\sqrt{K}$, write an expression for the mean and standard deviation of this estimator as a function of K and N . For the example values of N and K in the previous part, plot a bar plot of a Gaussian with this mean and variance (plot it from 1 to $2N - 1$, since the values can exceed N). How does this distribution (and the mean and stdev values) compare to that of the previous estimator? Which estimator do you think is "better", and why?

   (c) To verify your math, simulate the results for the previous two parts. First, write a function `observe(K, N)`, to generate K random observations of N tanks (use the matlab function `randi`). Then write functions `est1(obs)` and `est2(obs)` that compute each of the two estimates. Simulate 10,000 days worth of observations (again, assume $N = 40$ and $K = 16$) and collect the results of applying each estimator in a vector (length 10,000 each).

Plot histograms of the estimates (use the function `hist`), and compare to the plots from parts (a) and (b) (note: normalize histograms to sum to one, and make the axes the same!). Compare the mean and standard deviations of the simulated data to the values you computed in parts (a) and (b).

2. **Bayesian estimation.** Teddy and Hope are looking for Aaron in a very large one-dimensional shopping mall. Location is specified by a coordinate X. They know that, all else being equal, Aaron likes to hang out near the center of the shopping mall. Specifically, the probability distribution of his location is Gaussian with mean X=50 and variance 40. The only clue they have is a coffee cup at location X=30, containing the residue of a coffee that only Math Tools TAs drink (a naturally dried, triple-caffeine Ethiopian heirloom varietal, hand-extracted). Given the location of the coffee cup, and the aroma and dryness of the coffee residue, they estimate the conditional distribution of his position to be a Gaussian with mean X=30 and variance 100.

   (a) Frame this problem as a problem in Bayesian estimation, using appropriate terminology. What is Aaron's posterior distribution? Draw his prior, likelihood and posterior distributions on a single plot. (Rather than `normpdf`, compute the probabilities from the formula for the Gaussian distribution.) What are the mean and variance of the posterior?

   (b) Hope and Teddy realize they over-estimated the sitting time of the coffee residue, and decide that Aaron's likelihood function has mean X=30 but with a much smaller variance of 20. Redo part (a), and describe what happened to the posterior distribution, in terms of mean and variance. Does the change make sense?

   (c) What would the posterior distribution in (a) be if the prior had nearly flat (e.g., variance 10^6). Compare this variance to that of the posterior in (a). How does the inclusion of prior information affect the variance?

3. **Bayesian inference of binomial proportions.** Poldrack (2006) published an influential attack on the practice of "reverse inference" in fMRI studies, i.e., inferring that a cognitive process was engaged on the basis of activation in some area. For instance, if Broca's area was found to be activated using standard fMRI statistical-contrast techniques, researchers might infer that the subjects were using language. In a search of the literature, Poldrack found that Broca's area was reported activated in 103 out of 869 fMRI contrasts involving engagement of language, but this area was also active in 199 out of 2353 contrasts not involving language.

   (a) Assume that the conditional probability of activation given language, as well as that of activation given no language, each follow a Bernoulli distribution (i.e., like coin-flipping), with parameters x_l and x_{nl}. Compute the likelihoods of these parameters, given Poldrack's observed frequencies of activation. Compute these functions at the values `x=[0:.001:1]` and plot them as a bar chart.

   (b) Find the value of x that maximizes each discretized likelihood function. Compare these to the exact maximum likelihood estimates given by the formula for the ML estimator of a Bernoulli probability.

   (c) Using the likelihood functions computed for discrete x, compute and plot the discrete posterior distributions $P(x \mid data)$ and the associated cumulative distributions $P(X \leq x \mid data)$ for both processes. For this, assume a uniform prior $P(x) \propto 1$ and note that it will be necessary to compute (rather than ignore) the normalizing constant for Bayes' rule. Use the cumulative distributions to compute (discrete approximations to) upper and lower 95% confidence bounds on each proportion.

(d) *Are these frequencies different from one another?* Consider the joint posterior distribution over x_l and x_{nl}, the Bernoulli probability parameters for the language and non-language contrasts. Given that these two frequencies are independent, the (discrete) joint distribution is given by the outer product of the two marginals. Plot it (with `imagesc`). Compute (by summing the appropriate entries in the joint distribution) the posterior probabilities that $x_l > x_{nl}$ and, conversely, that $x_l \leq x_{nl}$.

(e) *Is this difference sufficient to support reverse inference?* Compute the probability $P(language \mid activation)$. This is the probability that observing activation in Broca's area implies engagement of language processes. To do this use the estimates from part (b) as the relevant conditional probabilities, and assuming the prior that a contrast engages language, $P(language) = 0.5$. Poldrack's critique said that we cannot simply conclude that activation in a given area indicates that a cognitive process was engaged without computing the posterior probability. Is this critique correct? To answer this, compare the Bayes factor ($\frac{p(\text{language}|\text{activation})}{p(\text{not language}|\text{activation})}$) using the maximum-likelihood estimates from Poldrack's data of activation probabilities, compared to the prior odds before running your experiment ($\frac{p(\text{language})}{p(\text{not language})}$).

4. **Signal Detection Theory.** Consider an experiment where a moving-dot visual stimulus is presented to a subject. The difficulty of detecting the motion is varied by changing the *coherence* of the moving dots, which is the fraction of dots moving to the right (at zero coherence, the dots move randomly, and at 100% coherence, all of the dots move to the right). Suppose we want to decide whether the stimulus is random or is moving to the right, based on the response of a single neuron that fires at a random rate, whose mean is 5 spikes/s in response to a 0% coherence noisy stimulus and 8 spikes/s for 10% coherence. Suppose also that the distribution of firing rates is Gaussian with a standard deviation of 1 spikes/s for both stimuli.

   (a) For the "no coherence"' stimulus, generate 1000 trials of the firing rate of the neuron in response to these stimuli (i.e., draw 1000 random samples from a Gaussian with $\mu = 5$ and $\sigma = 1$). Since we cannot have negative firing rates, set all rates that are below zero to zero. Now do the same thing for the 10% coherence stimulus. On the same figure, plot the histograms of the firing rates for each stimulus type.

   (b) The success of the decoder (assuming this model of Gaussian noise) is determined by two things, the separation of the mean firing rates and the standard deviation of the neuron. From class, we know that this is captured in the measure known as d'. Calculate d' for this task and pair of stimuli (ignoring the fact that you are clipping firing rates at zero).

   (c) Explain why the maximum likelihood decoder for this problem involves comparing the measurement to a threshold. For various thresholds t, calculate the hit and false-alarm rates using your sample data from (a), and plot these against each other (this is an ROC curve, defined in class). What threshold would you pick based on this curve to maximize the percentage-correct of the decoder, assuming that 0% and 10% coherence stimuli occur equally often. Plot this threshold as a point on the ROC curve and as a vertical line on your histogram from part (a). Next, suppose that 10% coherence stimuli occur 75% of the time. Determine and plot the threshold that maximizes percentage correct for this new prior.

   (d) Consider now a neuron with a more "noisy" response so that the mean firing rates are the same but the standard deviation is 2 spikes/s instead of 1 spike/s. What is the new value of d'. Recompute and plot the optimal (maximum accuracy) thresholds for this

noisy neuron for both the 50-50 and 75-25 priors. How do they differ from those in the previous part?