# Maximum-likelihood and Maximum-a-posteriori practice problems

John Briguglio

July 28, 2019

## 1 Understanding simple parameter estimation procedures

### 1.a

Show that computing the maximum likelihood estimator is equivalent to computing the maximum a posteriori estimator when the prior is flat (i.e. $P(\theta) = P_0$).

### 1.b

Sometimes, the parameter estimate $\hat{\theta}$ may be chosen to minimize an objective function. Suppose that we have some cost $C(\hat{\theta}, \theta)$, associated with using/estimating a parameter $\hat{\theta}$ when the real, underlying parameter was $\theta$. A common choice for the objective function to minimize is the expectation value of $C$,

$$< C(\hat{\theta}) >= \int P(\theta|\vec{x}) C(\hat{\theta}, \theta) \mathrm{d}\theta$$

where $\vec{x}$ is the observed data. If we take $C(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$, what parameter estimate minimizes the objective function? What is the significance of this quantity?

### 1.c

If we take $C(\hat{\theta}, \theta) = |\theta - \hat{\theta}|$, what parameter estimate minimizes the objective function? What is the significance of this quantity?

### 1.d

Which cost function, $C(\hat{\theta}, \theta)$, yields an estimate corresponding to the MAP estimator?

## 2 Firing rate estimation of a Poisson neuron

Suppose a neuron spikes with an underlying rate, $r$, in response to some stimulus, and we want to estimate this firing rate. If the stimulus presentation is of length $\Delta t$, and the neuron exhibits Poisson-like firing (a reasonable approximation for many neurons), then the total number of spikes observed during the stimulus presentation should follow a Poisson distribution:

$$P(n|r) = e^{-r\Delta t}\frac{(r\Delta t)^n}{n!}$$

Let N be the number of total measurements. It is convenient for much of this analysis to let $\mu = r\Delta t$ be the average number of spikes during a stimulus presentation.

### 2.a

Compute (analytically) the maximum likelihood estimator, $\hat{r}_{MLE}$. What is the significance of this value?

### 2.b

Simulate this process using a Poisson random number generation a mean spike number $\mu = 1$. Compute and plot the full distribution, $P(r|\vec{n})$ under the maximum-likelihood using $N = 5, 20, 50$ trials. How close are the ML estimates to the true value? Consider running this with multiple seeds for the random number generator to explore the variability across 'experiments'.

### 2.c

The width of the distribution demonstrates the degree of uncertainty in the parameter estimation (assuming the underlying model is correct). An interesting quantity to examine is

$$I = -\left.\frac{\partial^2}{\partial r^2} f(r;\vec{n})\right|_{\hat{r}_{MLE}}$$

(In the case of a Gaussian likelihood function, convince yourself that this is simply $1/\sigma^2$). If the likelihood is well-approximated by a Gaussian, then the 95% confidence interval for $r$ is $[\hat{r}_{MLE} - 2/\sqrt{I}, \hat{r}_{MLE} + 2/\sqrt{I}]$. How does $I$ scale with $N$? You may want to use more sample sizes than just the three above to make this scaling relationship convincing. Asymptotically, the slope of this relationship is called the "Fisher Information" and may be interpreted as the expected amount of information each observation of the cell's response provides about its firing rate. BONUS: How does the Fisher Information depend on $\mu$?

### 2.d

One frequent model of neural population activity describes the distribution of firing rates using a log-normal distribution,

$$P(r; m, s) = \frac{1}{\sqrt{2\pi} s r} \exp\left(-\frac{(\log(r) - m)^2}{2s^2}\right)$$

where log refers to the natural logarithm (base $e$), and $m$ and $s$ respectively describe the mean and standard deviation of the Gaussian distribution $P(\log r)$.

Take $\Delta t = 1s$, $m = .3$ and $s = .9$. How do MAP estimates using this prior distribution compare to the ML estimates? How do the differences change with the number of observations, N? Why? How does the prior's influence depend on the underlying firing rate of the neuron?

One common criticism of the Bayesian approach is that results of the analysis can sometimes depend quite exquisitely on the form of the prior distribution, which is an additional assumption of the analysis. If scientific conclusions depend on this inference process, one common control is to show that the results of the analysis hold across multiple different 'reasonable' prior distributions.

## 3 Decoding activity from a neural population

Neurons in some brain areas have firing rates that are modulated by some continuously varying stimulus property (e.g. orientation of a grating in early visual areas, frequency tuning in auditory areas, spatial location in hippocampus). To model this activity, we denote the firing rates $r_i(x)$ for neuron $i$ at location $x$.

### 3.a

If the neural activity is well described by a Poisson process and each neuron's activity is independent from one another (i.e. only conditionally dependent on $x$), what is the probability of observing a population activity $\vec{n} = (n_1, n_2, ..., n_N)$ given that the animal is at location $x$?

### 3.b

Write down the log-likelihood function.

### 3.c

The result of **b** provides a simple form with some useful properties. Since we are interested in extracting location information rather than a full probability distribution, terms with no $x$ dependence may be considered constants. The remaining terms are a linear function of the number of spikes. If a new linear neural unit were to respond in a manner such that its activity represents the log-likelihood of being in location $y$, how strongly should it be connected to each

neuron in this brain area? The simplicity of this form suggests the plausibility of ML decoding in real neural systems.

### 3.d

(Real decoding using real data? Friday programming session?)

## 4 Interpreting a common machine learning cost function

In many machine learning applications, there is a probability distribution of data, $p(x_i)$, that a model distribution $q(x_i)$ is going to try to match.

### 4.a

Consider the significance of the quantity

$$\log \frac{p(x_i)}{q(x_i)}$$

What does it mean if this value is positive, negative or equal to zero?

### 4.b

The expectation value of this quantity across the data is called the *Kullback-Leibler divergence*

$$\sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)} = \sum_i p(x_i) \log p(x_i) - \sum_i p(x_i) \log q(x_i)$$

It is most briefly expressed as on the left, but we will use the expression on the right to understand its pieces.

Are each of these terms greater than or less than zero?

### 4.c

The term on the left is $-H(p)$, where $H(p)$ is called the *entropy* of $p(x_i)$. The term on the right is called the *cross-entropy*, $H(p, q)$. Show that the cross-entropy is extremized (i.e. has a local maxima or minima) when $q(x_i) = p(x_i)$. Is this a maximum or a minimum? (Hint: since the model $q(x_i)$ is something we may vary, but the data is fixed for us to model, you may consider each $q(x_i)$ to be an independent variable and use Lagrange multipliers to enforce the normalization constraint).

### 4.d

Overall, we have shown that Kullback-Leibler divergence is generally a positive quantity which becomes zero exactly when $q(x_i) = p(x_i)$. In this sense, it is one useful measure of how distinguishable different models are from the data.

In a machine learning context, because the data is given and all degrees of freedom are in the model, minimizing KL-divergence is equivalent to minimizing cross-entropy. Show that minimizing the cross-entropy is equivalent to maximizing the likelihood of the data given the model.

(Hint: write down the full probability of observing the data, noting that if $N$ is the number of samples, the number of times $x_i$ was observed is $N \times p(x_i)$.)

## 5 German tank problem

In this problem, we will look into a scenario where the maximum likelihood estimate gives a less than optimal results. The 'German tank problem' historically related to the problem of estimating the size of the German cavalry during the second world war based on serial numbers from seized German tanks.

### 5.a

If there are $N$ tanks produced and the serial numbers range from $1, ..., N$, assuming each serial number is equally likely to be seen, what is the likelihood function, $P(k|N)$ after a single observation?

### 5.b

Which value of $N$ maximizes the likelihood function when only a single tank is observed?

Here, the mean number of tanks expected is not well defined, as the resulting likelihood function cannot even be normalized. With more than one tank, the problem does become more tractable. The full solution to this problem using Bayesian analysis involves significantly more combinatorics (with more than one observation, the likelihood function becomes more well-behaved). Analysis of this problem was able to give Allied forces a more accurate estimate of the number of tanks produced than intelligence estimates.