Revision exercises

Foundations of Data Science—DJW—2017/2018

This sample of questions contains some that are harder or longer than an exam question, and some that are shorter or easier. Each of these questions, like an exam question, draws on ideas from several parts of the course.

Question 1. (a) Let $X_1, X_2, \ldots, X_n$ be a random sample taken from the Uniform$[-\theta, \theta]$ distribution, where θ is some unknown parameter. Derive a formula for the maximum likelihood estimator for θ.

(b) Explain how to use the resampling method to compute an approximate confidence interval for θ. Give pseudocode.

(c) Many datasets contain a few extreme outliers, for example because of a glitch in data collection. To allow for this, suppose that each X_i is Uniform$[-\theta, \theta]$ with probability p, and Normal$(0, \sigma^2)$ with probability $1 - p$. Here $p \in [0, 1]$ and $\sigma > 0$ are unknown parameters. Derive an expression for the log likelihood of (θ, p, σ).

(d) Numerical optimization is often easier when parameters are unconstrained, i.e. allowed to take any value $-\infty$ to $+\infty$. Rewrite your log likelihood expression in terms of unconstrained parameters.

Note: In the exam, you would be told the density function for the Uniform *and* Normal *distributions.*

Question 2. A common task in data processing is counting the number of unique items in a collection. When the collection is too large to hold in memory, we may wish to use fast approximation methods, such as the following:

Given a collection of items $A_1, A_2, \ldots$, compute the hash of each item $X_1 = h(A_1)$, $X_2 = h(A_2)$, $\ldots$, and compute

$$T = \max_{1 \leq i \leq n} X_i.$$

If the hash function is well designed, then each X_i can be treated as uniformly distributed in $[0, 1]$, and unequal items will yield independent X_i.

(a) Show that $\mathbb{P}(T \leq t) = t^m$, where m is the number of unique items in the collection. Find the density function for T.

(b) Find the maximum likelihood estimator for m.

(c) Explain how to use the resampling method to find a confidence interval for m.

Note: You should explain the general procedure for resampling, and give pseudocode for this case. In questions like this one, where we want to study the distribution of a maximum, there are issues with the accuracy of the resampling procedure; in your answer you are expected to apply the procedure, not to worry about its accuracy.

Question 3. (a) Define the term *stationary*, as applied to Markov chains.

(b) Consider the noisy recurrence relation

$$X_{n+1} = \alpha X_n + \sigma \varepsilon_n$$

where $0 < \alpha < 1$, and $(\varepsilon_0, \varepsilon_1, \ldots)$ is a collection of independent Normal$(0, 1)$ random variables. Write down expressions for the mean and variance of X_{n+1} in terms of those for X_n. Assuming that the sequence $(X_0, X_1, \ldots)$ is stationary, calculate the mean and variance of X_n.

(c) By writing X_n in terms of $\varepsilon_0, \ldots, \varepsilon_{n-1}$, or otherwise, find the stationary distribution.

Note: In lectures you studied Markov chains with a finite state space, whereas here X_n is a real number. The question is asking you to apply your knowledge of stationarity to a new setting.

Question 4. A compulsive gampler has a choice of two machines to play. The first has probability α_1 of paying out, the second has probability α_2. The gambler doesn't know the values of the parameters α_1 and α_2, so treats them as unknown parameters, both with prior distribution Beta(δ, δ) where $\delta = 0.5$. Here are some strategies that the gambler might use to decide which machine to play next:

- Greedy: after each turn, compute the posterior distribution of α_1 and α_2. Play the machine with the larger posterior mean.

- ε-greedy: At each turn, with probability $1 - \varepsilon$ play the machine with the larger posterior mean, and with probability ε pick a machine uniformly at random.
- Probabilistic: compute the probability that $\alpha_1 > \alpha_2$ using the posterior distributions, and play machine 1 with this probability.
- "Thompson sampling": generate A_1 from the posterior distribution for α_1, and generate A_2 from the posterior distribution for α_2; play machine 1 if $A_1 > A_2$ and machine 2 otherwise.

(a) After w_1 wins and l_1 losses on the first machine, what is the posterior distribution of α_1? Explain your calculation.
(b) Give pseudocode for the probabilistic strategy. (You may assume there is a library routine `rbeta(x,y)` that generates a random value from the Beta(x, y) distribution.) Discuss the relationship between the probabilistic strategy and Thompson's sampling strategy.
(c) After n_1 plays of machine 1, give an approximate 95% confidence interval for the number of wins of that machine.
(d) Discuss how the strategies are likely to perform. In your answer, you might consider whether the gambler can end up playing only one machine, and whether that one machine is the one with the smaller payout probability.

Note: in the exam, you would be given the density function, mean and variance of the Beta *distribution. Part (d) is more open ended than you would be asked in the exam, but it is nonetheless possible to give a crisp answer using the techniques you've learnt in the course. Hint: use your answer to part (c) in your answer to (d).*

Question 5.

(a) Let $Y_1, \ldots, Y_n$ be a random sample taken from the distribution $\mathbb{P}(Y_i = 1) = e^{\xi}/(1 + e^{\xi})$, where ξ is an unknown parameter. Find the maximum likelihood estimator for ξ.

Three chess players play each other. In a tournament, A won 7 matches against B and lost 3, A won 9 matches against C and lost 1, and B won 6 matches against C and lost 4. We wish to ascribe a skill level for each player, such that the higher the skill difference the more likely it is that the higher-skilled player will win a match. Let μ_A, μ_B, and μ_C be skill levels, and consider this model: if match i is between players $p1(i)$ and $p2(i)$ then the probability that $p1(i)$ wins is $e^{\xi_i}/(1 + e^{\xi_i})$ where $\xi_i = \mu_{p1(i)} - \mu_{p2(i)}$.

(a) Find the log likelihood of (μ_A, μ_B, μ_C).
(b) Explain what is meant by a *linear model*. Write down a linear model for the vector $(\xi_1, \ldots, \xi_{30})$ in which the unknown parameters are μ_A, μ_B, and μ_C. Explain what the features are in your model.
(c) Explain the term *linear independence*. Show that the features you identified are not linearly independent.
(d) Give pseudocode for computing the maximum likelihood estimator of (μ_A, μ_B, μ_C). You may assume that there is a general purpose library routine `fmin(`f`,`x_0`)` which finds x to minimize an arbitrary function $f(x)$ with initial guess $x = x_0$. Explain your code.

Hint: what does part (c) tell you about part (d)?