Example sheet 1

Probability and random variables

Foundations of Data Science—DJW—2018/2019

This example sheet covers material up to Lecture 4 on 15 October. The more challenging questions are 3(b), 5(b), 8, 10(b).

Question 1. If $X_1, \dots, X_n$ are independent samples from Uniform$[0, \theta]$, find the maximum likelihood estimator for θ. *Hint. Write the density as*

$$\Pr_X(x \mid \theta) = \frac{1}{\theta} 1_{x \geq 0} 1_{x \leq \theta} \quad \textit{for } x \in \mathbb{R}$$

where $1_{\{\cdot\}}$ *stands for the indicator function,* $1_{true} = 1$ *and* $1_{false} = 0$.

Question 2. Let x_i be the population of city i, and let y_i be the number of crimes reported. Fit the model $Y_i \sim \text{Poisson}(\lambda x_i)$, where λ is an unknown parameter.

Question 3. A 0/1 signal is being sent over a noisy wire. If x_n is the true signal at timestep $n \in \{1, 2, \dots\}$, then the received message is $R_n = x_n + \text{Normal}(0, \varepsilon^2)$, where ε is known. Suppose that the signal being sent has a single changepoint, i.e.

$$x_n = \begin{cases} 0 & \text{for } n \leq \theta \\ 1 & \text{for } n > \theta \end{cases}$$

(a) Give pseudocode for a function $\mathsf{changepoint}([r_1, \dots, r_N])$ to estimate θ from N received messages.

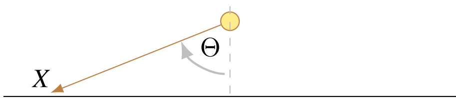
We'd like to run this as a 'streaming' procedure, looking for a changepoint in $[r_1]$ then in $[r_1, r_2]$ then in $[r_1, r_2, r_3]$ and so on.

(b) The naive $\mathsf{changepoint}$ function might produce lots of false alarms, detecting a changepoint as soon as it sees a high r_n and then realising its mistake when it sees r_{n+1}. Suggest an evidence-based approach to fix this problem.

Question 4.

(a) Let U be a uniform random variable on $[0, 1]$. Let $Y = U(1 - U)$. Calculate $\mathbb{P}(Y \leq y)$, and hence find the density of Y.

(b) A point lightsource at coordinates $(0, 1)$ sends out a ray of light at an angle Θ chosen uniformly in $[-\pi/2, \pi/2]$. Let X be the point where the ray intersects the horizontal line through the origin. What is the density of X? *This random variable is known as the Cauchy distribution. It is unusual in that it has no mean.*



Question 5. The Gumbel distribution is used in econometrics, for modelling how people make choices. If $X \sim \text{Gumbel}(\lambda)$ then

$$\mathbb{P}(X \leq x) = \exp\Big[-\exp\big(-(x - \lambda)\big)\Big], \quad x \in \mathbb{R}.$$

Let $X_1 \sim \text{Gumbel}(\lambda_1)$ and $X_2 \sim \text{Gumbel}(\lambda_2)$ be independent. Show the following:

$$\max(X_1, X_2) \sim \text{Gumbel}\Big(\log(e^{\lambda_1} + e^{\lambda_2})\Big) \tag{a}$$

and

$$\mathbb{P}(X_1 \geq X_2) = \frac{e^{\lambda_1}}{e^{\lambda_1} + e^{\lambda_2}}. \tag{b}$$

Hint. For the first equation, $\mathbb{P}(\max(X_1, X_2) \leq x) = \mathbb{P}(X_1 \leq x \textit{ and } X_2 \leq x)$. *This trick saves you a fiddly integration.*

Question 6. Let $X \sim \text{Normal}(\mu, \sigma^2)$. We wish to sample from $(X \mid X \geq 0)$. Give pseudocode, based on the inversion method. You should use the scipy functions `ppf` and `cdf` functions, described in the lecture notes appendix.

Question 7. Consider a pair of random variables with joint density

$$\Pr_{X,Y}(x, y) = \frac{3}{16} x y^2, \qquad 0 \leq x \leq 2, \quad 0 \leq y \leq 2.$$

Find $\Pr_X(x)$ and $\Pr_Y(y)$, the marginal densities. *Hint. It may be easier to first prove Exercise 1.7 from lecture notes.*

Question 8. In a hash table with n buckets and a load factor of α (i.e. with αn items hashed), what is the expected number of empty buckets? *Hint:* $\mathbb{E}\, 1_A = \mathbb{P}(A)$.

Question 9. Let $X_i \sim \text{Uniform}(2x/3, 4x/3)$, where x is given. Find a 95% confidence interval for $X_1 + \cdots + X_n$. *This arises in the context of statistical multiplexing of TCP flows on the Internet. See Exercise 2.6 in lecture notes for the background.*

Question 10. Let $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$, where the X_i are independent $\text{Exp}(\lambda)$ random variables.

(a) Find the mean and standard deviation of $\bar{X}_n$.

(b) Let $N \sim \text{Poisson}(\nu)$, independent of the X_i. Find the mean and standard deviation of $\bar{X}_{N+1}$. *Hint. Use the law of total expectation, and condition on N.*

Question 11. Suppose we're given a function $f(x) \geq 0$ and we want to evaluate

$$\int_{x=a}^{b} f(x)\, dx.$$

Here's an approximation method: (i) draw a box that contains $f(x)$ over the range $x \in [a, b]$, (ii) scatter points uniformly at random in this box, (iii) return $A \times p$ where A is the area of the box and p is the fraction of points that are under the curve.

Explain why this is a special case of Monte Carlo integration. In your answer, you should identify the random variable and the function to which Monte Carlo is being applied.