CS109 Problem Review

Casey Haaland – chaaland@stanford.edu – 05743086

June 1, 2016

Some CDFs

Compute the CDF of the following distributions

(a) The *uniform distribution*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & else \end{cases}$$

Solution: Recalling that the CDF $F_X(x)$ is just $\textbf{prob}(X \leq x)$ we use integration to compute this as

$$\begin{aligned} F_X(x) &= \int_a^x \frac{1}{b-a} dx \\ &= \frac{x-a}{b-a} \end{aligned}$$

Where of course the CDF is 0 if $x \leq 0$ and 1 if $x \geq 1$.

(b) The *Rayleigh Distribution*

$$f_X(x) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & else \end{cases}$$

This gives the probability density of the length of a vector in $\mathbb{R}^2$ which has components which are independent zero mean Gaussian random variables.

Solution: Assuming x is positive

$$F_X(x) = \int_0^x \frac{z}{\sigma^2}\ e^{-z^2/2\sigma^2} dz$$

Now let $u = z^2/2\sigma^2$ then $du = z/\sigma^2\ dx$ so

$$\begin{aligned} F_X(x) &= \int_0^{x^2/2\sigma^2} e^{-u} du \\ &= \left[-e^{-u}\right]_0^{x^2/2\sigma^2} \\ &= 1 - e^{-x^2/2\sigma^2} \end{aligned}$$

(c) The standard *Cauchy Distribution*

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

This gives the probability density of the ratio of two standard normal random variables.

Solution:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} \frac{1}{\pi(1+z^2)} \\ &= \frac{1}{\pi}\left[\arctan z\right]_{-\infty}^{x} \\ &= \frac{1}{\pi}\left(\arctan(x) + \pi/2\right) \end{aligned}$$

Functions of a Random Variable

Let U be the random variable which is uniformly distributed on (0,1]. Compute the distributions of the following random variables.

(a) $X = U^{1/2}$

Solution:

$$\begin{aligned} F_X(x) &= \mathbf{prob}(X \leq x) \\ &= \mathbf{prob}(U^{1/2} \leq x) \\ &= \mathbf{prob}(U \leq x^2) \end{aligned}$$

$$F_X(x) = \begin{cases} 1 & x \geq 1 \\ x^2 & x \in (0,1) \\ 0 & else \end{cases}$$

Where the last line follows from the fact that this is just the CDF of the uniform evaluated at x^2. We then get the PDF by differentiation.

$$f_X(x) = \begin{cases} 2x & x \in (0,1) \\ 0 & else \end{cases}$$

(b) $Y = -\ln U$

Solution:

$$\begin{aligned} F_Y(y) &= \mathbf{prob}(Y \leq y) \\ &= \mathbf{prob}(-\ln U \leq y) \\ &= \mathbf{prob}(\ln U \geq -y) \\ &= \mathbf{prob}(U \geq e^{-y}) \\ &= 1 - \mathbf{prob}(U < e^{-y}) \end{aligned}$$

Now plugging into the CDF of the standard uniform distribution we can compute the second term as

$$F_U(e^{-y}) = \begin{cases} 1 & e^{-y} \geq 1 \\ e^{-y} & e^{-y} \in (0,1) \\ 0 & else \end{cases}$$

$$F_U(e^{-y}) = \begin{cases} 1 & y \leq 0 \\ e^{-y} & y > 0 \end{cases}$$

Now we substitute this function back into our expression for the CDF of Y to get

$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ 1 - e^{-y} & y > 0 \end{cases}$$

Taking the derivative w.r.t y we find $Y \sim \text{Exp}(1)$

(c) $Z = aU + b \quad a < 0, b > 0$

$$\begin{aligned} F_Z(z) &= \textbf{prob}(Z \leq z) \\ &= \textbf{prob}(aU + b \leq z) \\ &= \textbf{prob}\left(U \geq \frac{z-b}{a}\right) \\ &= 1 - \textbf{prob}\left(U \leq \frac{z-b}{a}\right) \end{aligned}$$

Again just dealing with the second quantity we have

$$F_U\left(\frac{z-b}{a}\right) = \begin{cases} 1 & \frac{z-b}{a} \geq 1 \\ \frac{z-b}{a} & \frac{z-b}{a} \in (0,1) \\ 0 & else \end{cases}$$

$$F_U\left(\frac{z-b}{a}\right) = \begin{cases} 0 & z \leq a+b \\ \frac{z-b}{a} & z \in (a+b, b) \\ 0 & else \end{cases}$$

Now plugging into our expression for the CDF of Z we have

$$F_Z(Z \leq z) = \begin{cases} 1 & z \leq a+b \\ 1 - \frac{z-b}{a} & z \in (a+b, b) \\ 0 & else \end{cases}$$

Taking the derivative with respect to z we find $Z \sim \text{Uni}(a+b, b)$

Chi Squared Distribution

From years of extensive data collection, you and your colleagues determine the length of adult snakes in a nature preserve are well modeled as a Gaussian random variable. Specifically, suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ where X is the length of the snake. Since you have a good model for size of the snakes, you want to characterize the deviation of a snake's length from the mean to determine if a snake is abnormally large or small for the preserve. A large number of snakes that deviate "too much" from the mean

might indicate the model is no longer correct and perhaps an underlying enviromental change causing it.

You and your colleagues are agnostic to whether the snakes are too far below or above the mean, since any deviation might indicate a failing of the model. You propose a good measurement to characterize this is the random variable $Y = (X - \mu)^2/\sigma^2$. In order to calculate the probability a snake deviates by a certain amount from the mean, we need a distribution for Y. Find the PDF of y.[1]

Solution: As per usual, we start with the CDF in order to derive the PDF of y.

$$\begin{aligned} F_Y(y) &= \mathbf{prob}(Y \leq y) \\ &= \mathbf{prob}((X-\mu)^2/\sigma^2 \leq y) \\ &= \mathbf{prob}((X-\mu)^2 \leq y\sigma^2) \\ &= \mathbf{prob}(-\sigma\sqrt{y} \leq X - \mu \leq \sigma\sqrt{y}) \\ &= \mathbf{prob}(-\sqrt{y} \leq Z \leq \sqrt{y}) \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) \end{aligned}$$

Taking the derivative we get the PDF. Note that this derivative is calculated using the chain rule. The derivative of the standard normal CDF (ϕ) is the standard normal PDF.

$$\begin{aligned} f_Y(y) &= 0.5y^{-1/2}\left(\frac{1}{\sqrt{2\pi}}e^{-(\sqrt{y})^2/2}\right) - \left[-0.5y^{-1/2}\left(\frac{1}{\sqrt{2\pi}}e^{-(\sqrt{y})^2/2}\right)\right] \\ &= \frac{1}{\sqrt{2y\pi}}e^{-y^2/2} \end{aligned}$$

The PDF is zero for $y < 0$

Discretizing a Continous RV

A large call center has determined that the number of minutes until an employee is able to serve the next customer on hold is distributed $X \sim \text{Exp}(\lambda)$. The amount of time it takes to answer a customer on hold directly affects the call center's revenue as well as their customer's satisfaction. As a result the call center naturally wants to do some analysis on the time taken to serve its customers. The company decides that the difference between waiting 5 minutes versus waiting, say 5 minutes and 23 seconds, is more or less irrelevant and they would instead like to count them both as having taken 5 minutes. Rather than model X which can take on a continous range of values, they want a model for $Y = \lfloor X \rfloor$ (the truncated value of X). Find the distribution of this random variable and its relevant parameters. Though wrong, an answer might be something like $\text{Poi}(\log \lambda)$

Solution: We note that Y is a discrete random variable that can take on non-negative integer values. We start with the following observation

$$\mathbf{prob}(Y = y) = \mathbf{prob}(y \leq X < y + 1)$$

We can then use the CDF of an exponential random variable to evaluate the quantity on the right.

$$\begin{aligned} \mathbf{prob}(Y = y) &= F_X(y+1) - F_X(y) \\ &= \left(1 - e^{-\lambda(y+1)}\right) - (1 - e^{-\lambda y}) \\ &= e^{-\lambda y}(1 - e^{-\lambda}) \\ &= (1-p)^y p \end{aligned}$$

[1] This random variable is distributed chi squared with one degree of freedom. Roughly, it gives the distribution of how far away from the mean a gaussian random variable will be (normalized by the variance)

Where $p = 1 - e^{-\lambda y}$ which shows that $Y \sim Geo(1 - e^{-\lambda})$. From this we can see that the geometric is a sort of discretized version of the exponential distribution.

Windfarm Modeling

Wind velocity can be expressed in terms of its north and east components denoted v_x and v_y respectively. On our wind farm from the midterm, the north and east components of wind velocity can be modeled as independent gaussian random variables with distribution $v_x, v_y \sim \mathcal{N}(0, \sigma^2)$ during the summer months. The overall magnitude of the wind is then distributed as a *Rayliegh Distribution* given by the PDF

$$f_X(x) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & else \end{cases}$$

Always interested in evaluating the effectiveness of our business, we wish to model the wind speed on our farm. To this end we collect N independent measurements of wind speeds (the magnitude of velocity) $w_1, w_2, \cdots, w_N$. Find a maximum likelihood estimate of σ^2 if we are modeling the wind speed as coming from a Rayleigh distribution

Solution: Denote our wind speed as a random variable $W \sim \text{Rayleigh}(\sigma)$. The likelihood is given by

$$L(\sigma^2) = \prod_{i=1}^{N} f_W(w_i; \sigma^2)$$

We then take the logarithm and do some simplification

$$\begin{aligned} \ell(\sigma^2) &= \log \prod_{i=1}^{N} f_W(w_i; \sigma^2) \\ &= \sum_{i=1}^{N} \log f_W(w_i; \sigma^2) \\ &= \sum_{i=1}^{N} \log \left(\frac{w_i}{\sigma^2} e^{-w_i^2/2\sigma^2} \right) \\ &= \sum_{i=1}^{N} \log w_i - \log \sigma^2 - w_i^2/(2\sigma^2) \end{aligned}$$

Maximizing over σ^2 we have

$$\begin{aligned} \arg\max_{\sigma^2} \ell(\sigma^2) &= \arg\max_{\sigma^2} \left[\sum_{i=1}^{N} \log w_i - \sum_{i=1}^{N} \log \sigma^2 - \sum_{i=1}^{N} w_i^2/(2\sigma^2) \right] \\ &= \arg\min_{\sigma^2} \left[N \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} w_i^2 \right] \end{aligned}$$

Taking the derivative with respect to σ^2 and equating to 0 we have

$$0 = N/\sigma^2 - \frac{1}{2\sigma^4}\sum_{i=1}^{N} w_i^2$$

$$N/\sigma^2 = \frac{1}{2\sigma^4}\sum_{i=1}^{N} w_i^2$$

$$\sigma^2 = \frac{1}{2N}\sum_{i=1}^{N} w_i^2$$

As it turns out, this is an unbiased estimate of the parameter though we didn't ask you to show it.