STCS 6702: Foundations of Graphical Models: Homework 0

Maxime TCHIBOZO (MT3390)

September 2020

This homework will give you a sense of the background needed to take this course and the type of thinking and mathematics that you will use.

Please prepare your answers using LATEX with the template provided on the course website. Submit the PDF of your completed assignment on Courseworks.

1 Problem 1 (via David Duvenaud)

Consider a probability density $p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

a) For some $x \in \mathbb{R}$, can $p(x) < 0$?

Solution:

Let $x \in \mathbb{R}$, with μ and σ defined as above.

$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$.

The constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is always positive, and the exponential function always outputs positive values. $p(x)$ is the product of two positive values, so it is always positive.

b) For some $x \in \mathbb{R}$, can $p(x) > 1$?

Solution:

For $p(x)$ to be a probability density function, it must satisfy Kolmogorov's Axioms. In particular, $p(x)$ must integrate to 1.

If we also take into account the result from the above function, it holds that $p(x)$ can take any value in $\mathbb{R}^+ \setminus \{0\}$ so long as its integral over $\mathbb{R}$ is equal to one.

Example: Let $\mu = 0$ and $\sigma^2 = \frac{1}{8\pi}$. For $x = 0$, we have $p(x) = 2$.

2 Problem 2 (via Joe Blitzstein)

You have a jar of 1,000 coins. 999 are fair coins, and the remaining coin will always land heads. You take a single coin out of the jar and flip it 10 times in a row, all of which land heads. What is the probability your next toss with the same coin will land heads? Explain your answer.

Solution:

To simplify notations, we use the following event names:

$F = \{\text{Coin is Fair}\}$

$U = \overline{F} = \{\text{Coin is Unfair}\}$

$H^{10}_{past} = \{\text{Past 10 tosses were heads}\}$

$H^{1}_{next} = \{\text{Next toss will yield heads}\}$

We use the Law of Total Probability by conditioning on $\{U, F\}$ when using the probability distribution $P(\cdot \mid H^{10}_{past})$:

$P(H^1_{next} \mid H^{10}_{past}) = P(H^1_{next} \mid H^{10}_{past}, F)P(F \mid H^{10}_{past}) + P(H^1_{next} \mid H^{10}_{past}, U)P(U \mid H^{10}_{past})$

$= \frac{1}{2} \cdot P(F \mid H^{10}_{past}) + 1 \cdot P(U \mid H^{10}_{past})$

$= 1 - \frac{1}{2} \cdot P(F \mid H^{10}_{past})$

We must now infer $P(F \mid H^{10}_{past})$ using Bayes' Rule:

$P(F \mid H^{10}_{past}) \cdot P(H^{10}_{past}) = P(F, H^{10}_{past}) = P(H^{10}_{past} \mid F)P(F)$

$P(F \mid H^{10}_{past}) = \frac{P(H^{10}_{past}|F)P(F)}{P(H^{10}_{past})} = \frac{P(H^{10}_{past}|F)P(F)}{P(H^{10}_{past}|F)P(F)+P(H^{10}_{past}|U)P(U)}$

$P(F \mid H^{10}_{past}) = \frac{(\frac{1}{2})^{10} \cdot \frac{999}{1000}}{(\frac{1}{2})^{10} \cdot \frac{999}{1000} + 1 \cdot \frac{1}{1000}} = \frac{999}{999+2^{10}} \approx 0.4938.$

We have inferred that there is about a 49 % chance that the coin we have selected is the Unfair coin. Knowing this, we can deduce:

$P(H^1_{next} \mid H^{10}_{past}) = 1 - \frac{1}{2} \cdot P(F \mid H^{10}_{past}) \approx 0.753$

There is approximately a 75% chance that we will obtain a heads at the next toss.

3 Problem 3 (via David Duvenaud)

In the exponential family of distributions, $p(x \mid \theta) = h(x)exp\{\eta(\theta)^T t(x) - a(\theta)\}$, for $x \in \mathbb{R}^n, \theta \in \mathbb{R}^d, h(x) : \mathbb{R}^n \to \mathbb{R}, \eta(\theta) : \mathbb{R}^d \to \mathbb{R}^p, t(x) : \mathbb{R}^n \to \mathbb{R}^p$ and $a(\theta) : \mathbb{R}^d \to \mathbb{R}$. What must $a(\theta)$ be for $p(x \mid \theta)$ to be a valid probability distribution? Why?

Solution :

$a(\theta)$ is the normalization function for the exponential family. For $p(x \mid \theta)$ to be a probability distribution, $a(\theta)$ must be such that $\int\limits_{\mathcal{X}} p(x \mid \theta)dx = 1$.

Mathematically speaking, $a(\theta)$ must verify $a(\theta) = log\left(\int\limits_{\mathcal{X}} h(x)exp(\eta(\theta)^T t(x))dx\right)$

4 Problem 4

Consider n i.i.d. random variables $X_1, ..., X_n$ with probability density:

$f(x \mid k) = kx^{k-1}e^{-x^k}$.

The density is only defined for positive random variables, so assume all observations are positive. The parameter k is also positive.

(a) Write down the log-likelihood of the data.

Solution :

$\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; k) = \sum\limits_{i=1}^{n} log(f(X_i \mid k))$

$= \sum\limits_{i=1}^{n} log(k) + (k-1)log(x_i) - x_i^k$

$= n \cdot log(k) + (k-1)\sum\limits_{i=1}^{n} log(x_i) - \sum\limits_{i=1}^{n} x_i^k$

(b) Write down the derivative of the log-likelihood with respect to k.

Solution :

$\frac{d}{dk}\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; k) = \frac{n}{k} + \sum\limits_{i=1}^{n} log(x_i) - \sum\limits_{i=1}^{n} log(x_i)x_i^k$

(c) k is constrained to be positive, but often times it is convenient to work with parameters that live on an unconstrained space. To do this, we reparametrize the distribution. Let $\theta \in \mathbb{R}$ be such that $k = exp(\theta)$. Write the log-likelihood as a function of θ, and the derivative of this log-likelihood with respect to θ.

Solution :

$$\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; \theta) = n \cdot \theta + (exp(\theta) - 1) \sum_{i=1}^{n} log(x_i) - \sum_{i=1}^{n} x_i^{exp(\theta)}$$

$$\frac{d}{d\theta}\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; \theta) = n + exp(\theta) \sum_{i=1}^{n} log(x_i) - exp(\theta) \sum_{i=1}^{n} log(x_i) x_i^{exp(\theta)}$$

(d) Now consider the parameterization $k = log(1 + exp(\theta))$. Write the log-likelihood and derivative with respect to θ. At a high level, compare these two parameterizations. What are the benefits and drawbacks of each?

Solution :

$$\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; \theta) = n \cdot log(log(1+exp(\theta))) + (log(1+exp(\theta)) - 1) \sum_{i=1}^{n} log(x_i) - \sum_{i=1}^{n} log(x_i) x_i^{log(1+exp(\theta))}$$

$$\frac{d}{d\theta}\mathcal{L}(\{\mathbf{x_i}\}_{i=1}^n; \theta) = \frac{n \cdot exp(\theta)}{(1+exp(\theta)) log(1+exp(\theta))} + \frac{exp(\theta)}{1+exp(\theta)} \sum_{i=1}^{n} log(x_i) - \frac{exp(\theta)}{1+exp(\theta)} \sum_{i=1}^{n} log(x_i) x_i^{log(1+exp(\theta))}$$

We should first note that in all three cases (k, $exp(\theta)$, $log(1+\exp(\theta))$) solving optimization problems by utilizing the derivative of the log-likelihood with respect to k or θ will be a computationally difficult task.

Even in their wildest dreams, an imaginative mathematician would struggle to find a closed form analytic solution to any equation involving the derivative of these log-likelihoods (i.e Maximum Likelihood Estimation equations).

Next we must understand the role of each parametrization:

When $\theta \sim \pm\infty, log(1 + exp(\theta)) \sim \theta$. This transformation "smooths" over large values theta whereas the exponential transformation "explodes" for large values of theta.

When $\theta \to 0, exp(\theta) \sim \theta$. This transformation approximates θ for points surrounding 0 whereas the log-exponential transformation approximates to $log(2)$.

Depending on the prior knowledge of the range of values the parameter k could take, we would select one of the two parametrizations. Specifically, for large values of k, the log-exponential transformation would give more stability with regards to the variations of the log-likelihood and derivative-log-likelihood functions. It would also prevent gradients with respect to theta from exploding. This can be useful for numerical approximation algorithms.

For small values of k, the exponential transformation provides more stability for the log-likelihood and its derivative with respect to theta. This is also useful for numerical approximation algorithms.