CS 4641 - Homework 2

Karl Gemayel

- Submit your answers as an electronic copy on Canvas.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit. After submitting, make sure you **quadruple** check your submission on Canvas to make sure you submitted the right file.
- Typing with Latex is **highly recommended**. Typing with MS Word is also okay (**Submitted file should be in PDF format. Please do not submit .docx - it will NOT be graded**).
- Handwritten submissions will be accepted this time. **BUT BEWARE.** If the TA has trouble reading any part of the assignment, or if it is not presentable in a nice, professional format, then this part will receive a mark of 0. This is a hard rule, and you will not be given a second chance to explain what you have written. We are only doing this to save you time on writing out the math part. But this should not hinder the grading process.
- Do not submit a zip file. Your final submission should consist of three separate files: HTML or PDF version of document answering questions 1-3, completed Jupyter notebook, HTML/PDF version of completed (and executed Jupyter notebook).
- Explicitly mention your collaborators if any. Remember what I said about collaboration though!

1 Maximum Likelihood Estimation

Suppose we have N i.i.d (independent and identically distributed) data samples, i.e. $X = \{x^n\}_{n=1}^N$ from the following probability distributions. This problem asks you to **build a log-likelihood function** for each distribution ($i.e. \log P(X)$), and **find the maximum likelihood estimator of the parameter(s)**.

(a) Poisson distribution

The Poisson distribution is defined as

$$P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}(k = 0, 1, 2, ...).$$

What is the maximum likelihood estimator of λ?

(b) Exponential distribution

The probability density function of Exponential distribution is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the maximum likelihood estimator of λ?

(c) Gaussian normal distribution

A univariate Gaussian normal distribution, $\mathcal{N}(\mu, \sigma^2)$, is given by

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

What is the maximum likelihood estimator of μ and σ^2?

2 Chance of Karl Die

... No, this isn't a sinister Yoda-ish statement regarding my state of health. It refers to a type of dice famously (in my mind) named after me (by me). The dice is such that the value of the current roll depends on the value of the previous roll. More on that in a second - let's set up first.

Let $x^{(n)}$ be the n^{th} **sequence** of M dice rolls. We represent $x^{(n)}$ as a vector of individual rolls, i.e. $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_M^{(n)})$. Each Karl dice roll can take up to 12 values (that's my Messiah complex kicking in again), i.e. $x_i^{(n)} \in \{1, \ldots, 12\}$. Every **sequence** of rolls is independent of the other, but the rolls **within** a sequence are dependent on each other in the following way: first roll does not depend on any other roll, second roll depends on the first, third depends on the second, etc....

We also have a Regular 12-sided dice in which individual rolls are completely independent from each other. Every sequence of rolls uses a single dice (i.e. rolls within a sequence must be performed by the same dice). For the n^{th} sequence, we have a label $y^{(n)}$ such that

$$y^{(n)} = \begin{cases} 1 & \text{if Karl dice is used} \\ 0 & \text{if Regular dice is used} \end{cases}$$

We collect all the data in X and y variables which collectively constitute our dataset D, i.e.

$$D = (X, Y) = \{(x^{(n)}, y^{(n)})\}_{n=1}^{N}.$$

Our game is simple: given a sequence of rolls, how can we identify which type of dice was used to generate the roll? We will solve this problem by estimating the parameters that maximize the likelihood of our data.

(a) Show that the likelihood function can be written as

$$p(D|h) = \prod_{n=1}^{N} \left[\pi_1 K(x^{(n)})\right]^{y^{(n)}} \left[\pi_0 R(x^{(n)})\right]^{1-y^{(n)}}$$

where

$$\begin{aligned} \pi_i &= p(y = i) \\ K(x^{(n)}) &= p(x^{(n)}|y^{(n)} = 1) \\ R(x^{(n)}) &= p(x^{(n)}|y^{(n)} = 0) \end{aligned}$$

Note: if, during your derivation, you feel a step needs explanation, please feel free to describe it in English (as opposed to?) as well.

(b) Show that the log-likelihood can be written as

$$\log p(D|h) = \sum_{n=1}^{N} y^{(n)} \left[\pi_1 + \log K(x_1^{(n)}) + \sum_{i=2}^{M} \log K(x_i^{(n)}|x_{i-1}^{(n)}) \right] + \sum_{n=1}^{N} (1 - y^{(n)}) \left[\log \pi_0 + \sum_{i=1}^{M} \log R(x_i^{(n)}) \right] \tag{1}$$

where

$$\begin{aligned} K(x_1^{(n)}) &= p(x_1^{(n)}|y^{(n)} = 1) \\ K(x_i^{(n)}|x_{i-1}^{(n)}) &= p(x_i^{(n)}|x_{i-1}^{(n)}, y^{(n)} = 1) \\ R(x_i^{(n)}) &= p(x_i^{(n)}|y^{(n)} = 0) \end{aligned}$$

(c) Now that we have that done, we can see that we have 4 sets of parameters. Specifically,

$$\begin{aligned} \pi_i \quad & \textbf{for } i = 1, 2 \\ K(v) \quad & \textbf{for } v = 1, \ldots, 12 \\ K(v|v') \quad & \textbf{for } v = 1, \ldots, 12, v' = 1, \ldots, 12 \\ R(v) \quad & \textbf{for } v = 1, \ldots, 12 \end{aligned}$$

Write the optimization constraints for each of the above sets of parameters.

(d) Write the Lagrangian form of the maximum (log) likelihood estimation problem.

(e) Show that the log-likelihood can be further expanded to the form

$$\begin{aligned} \log p(D|h) = \sum_{n=1}^{N} y^{(n)} \Bigg[& \pi_1 + \sum_{v} I(x_1^{(n)} = v) \log K(x_1^{(n)} = v) \\ & + \sum_{i=2}^{M} \sum_{v} \sum_{v'} I(x_i^{(n)} = v, x_{i-1}^{(n)} = v') \log K(x_i^{(n)} = v | x_{i-1}^{(n)} = v') \Bigg] \\ + \sum_{n=1}^{N} (1 - y^{(n)}) & \left[\log \pi_0 + \sum_{i=1}^{M} \sum_{v} I(x_i^{(n)} = v) \log R(x_i^{(n)} = v) \right] \end{aligned}$$

where

$$I(a = b) = \begin{cases} 1 & \textbf{if } a = b \\ 0 & \textbf{otherwise} \end{cases}$$

$$I(a = b, c = d) = \begin{cases} 1 & \textbf{if } a = b \textbf{ AND } c = d \\ 0 & \textbf{otherwise} \end{cases}$$

and $\sum_v$ **is simply** $\sum_{v \in \{1,\ldots,12\}}$ **(similarly for** v'**).**

(f) Show that the MLE for π_1 **has the form**

$$\pi_1 = \frac{1}{N} \sum_{n=1}^{N} y^{(n)}$$

where N is the number of data points.

(g) Show that the MLE estimate for $K(v)$ has the form

$$K(v) = \frac{1}{N_1}\,\mathrm{count}_1(v, y = 1)$$

where N_1 is the number of data points where $y = 1$, and $\mathrm{count}_1(v, y = 1)$ is the number of data points with label $y = 1$ where feature 1 has the value v.

(h) Derive the MLE estimate for $K(v|v')$. You can't expect me to give you *all* the answers... but I'll be nice - you may (will) find the following measure helpful

$$\mathrm{count}_i(v, v', y^{(n)} = 1)$$

which is simply the number of data points with label $y = 1$, where the i^{th} feature has value v, and the $i - 1^{th}$ feature has value v'.

(i) Derive the MLE estimate for $R(v)$.