Chris Piech
CS 109

Section #7

Section 7: Algorithmic Analysis, MLE, MAP, & ML

1. **Binary Tree**: Consider the following function for constructing binary trees:

```
def random_binary_tree(p):
    """
    Returns a dictionary representing a random binary tree structure.
    The dictionary can have two keys, "left" and "right".
    """
    if random_bernoulli(p):   # returns true with probability p
        new_node = {}
        new_node["left"] = random_binary_tree(p)
        new_node["right"] = random_binary_tree(p)
        return new_node
    else:
        return None
```

The `if` branch is taken with probability p (and the `else` branch with probability $1 - p$). A tree with no nodes is represented by `None`; so a tree node with no left child has `None` for the `left` field (and the same for the right child).

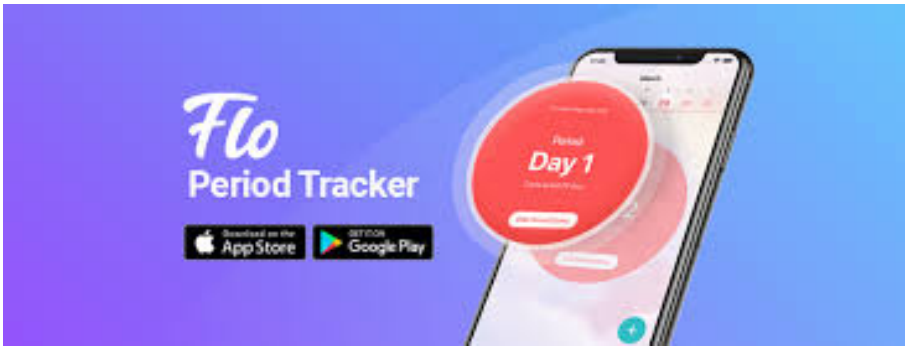
Let X be the number of nodes in a tree returned by `random_binary_tree`. You can assume $0 < p < 0.5$. What is $E[X]$, in terms of p?

Let X_1 and X_2 be number of nodes returned by the left and right calls to `random_binary_tree`. Because the recursive call is identical to the original function call, $E[X_1] = E[X_2] = E[X]$.

$$\begin{aligned} E[X] &= p \cdot E[X \mid \texttt{if}] + (1-p)E[X \mid \texttt{else}] \\ &= p \cdot E[1 + X_1 + X_2] + (1-p) \cdot 0 \\ &= p \cdot (1 + E[X] + E[X]) \\ &= p + 2pE[X] \\ (1-2p)E[X] &= p \\ E[X] &= \frac{p}{1-2p} \end{aligned}$$

Extra Challenge Q: Why did we need to assume that p is less than 0.5?

2. **Flo. Tracking Menstrual Cycles**



Let X represent the length of a menstrual cycle: the number of days, as a continuous value, between the first moment of one period to the first moment of the next, for a given person. X is parameterized by α and β with probability density function:

$$f(X = x) = \beta \cdot (x - \alpha)^{\beta - 1} \cdot e^{-(x-\alpha)^2}$$

a. For a particular person, $\alpha = 27$ and $\beta = 2$. Write an expression for the probability that they have their period on day 29. In other words, what is the $P(29.0 < X < 30.0)$?

$$P(29.0 < X < 30.0) = \int_{29.0}^{30.0} 2 * (x - 27) * e^{-(x-27)^2}$$

b. For a particular person, $\alpha = 27$ and $\beta = 2$. How many times more likely is their cycle to last **exactly** 28.0 days than exactly 29.0 days? You do not need to give a numeric answer. Simplify your expression.

$$\frac{f(X = 28)}{f(X = 29)} = \frac{2 * (28 - 27) * e^{-(28-27)^2}}{2 * (29 - 27) * e^{-(29-27)^2}} = \frac{e^3}{2}$$

c. A person has recorded their cycle length for 12 cycles stored in a list:

$$m = [29.0, 28.5, \ldots, 30.1]$$

where m_i is the recorded cycle length for cycle i. Use MLE to estimate the parameter values α and β. Assume that cycle lengths are IID.

You don't need a closed form solution. Derive any necessary partial derivatives and briefly describe how a program can use these derivatives to choose the most likely parameter values.

Define our likelihood function:

$$L(\alpha, \beta) = \prod_{i=1}^{12} f(m_i)$$

Now log likelihood to make the math easier later:

$$LL(\alpha, \beta) = \sum_{i=1}^{12} \log f(m_i)$$

$$\alpha = \arg\max_{\alpha} LL(\alpha, \beta)$$

$$\beta = \arg\max_{\beta} LL(\alpha, \beta)$$

Log of the pdf simplifies:

$$\log f(m) = \log \beta + (\beta - 1) \log(m - \alpha) - (m - \alpha)^2$$

Now take partial derivative w.r.t α and β:

$$\frac{\partial}{\partial \alpha} LL(\alpha, \beta) = \sum_{i=1}^{12} 2(m_i - \alpha) - \frac{\beta - 1}{m_i - \alpha}$$

$$\frac{\partial}{\partial \beta} LL(\alpha, \beta) = \sum_{i=1}^{12} \frac{1}{\beta} + \log(m_i - \alpha)$$

we can use gradient ascent to maximize LL. This computes gradient w.r.t each parameter α, β then moves the parameters a small step in the direction of the gradient.

We also accept valid closed-form solutions. For example, can perform gradient descent on α, then update β by computing closed-form optimal value (given some value of α:

$$\beta = -\frac{12}{\sum_{i=1}^{12} \log(m_i - \alpha)}$$

3. **Why Boba Cares About MAP**

A new boba place is coming to campus! They've hired you to help predict revenue by estimating how many orders you will receive per hour. After taking CS109, you are pretty confident that incoming orders can be considered independent events, and the process can be modeled with a Poisson.

Now the question is - what is the best λ parameter to use with the Poisson? To gather data, the boba place holds a soft opening for one hour and is visited by 4 curious students, each of whom made an order.

(a) What is the MLE estimate for λ, based only on this single data point?

(b) Because one observation isn't a lot of data, you want to leverage your belief that this boba place will be way more popular than 4 orders per hour. In particular, you want to incorporate the prior $\Lambda \sim N(\mu = 7, \sigma^2 = 10)$. What is the Maximum-A-Posteriori (MAP) estimate of λ?

To find the MLE, we start from finding the likelihood function (i.e. joint probability of observed events) and find the λ that maximizes the likelihood function.

$$\begin{aligned} L(\lambda) &= \frac{\lambda^4 \cdot e^{-\lambda}}{4!} \\ LL(\lambda) &= 4\log(\lambda) - \lambda - \log(4!) \\ \frac{\partial LL}{\partial \lambda} &= \frac{4}{\lambda} - 1 \end{aligned}$$

Set $\frac{\partial LL}{\partial \lambda}$ to 0 and solve for λ.

$$\lambda = 4$$

MAP estimate of λ: we find the λ that maximizes the inference expression given the observation, i.e. we want to solve:

$$\begin{aligned} \arg\max_{\lambda} f(\lambda|X=4) &= \arg\max_{\lambda} \frac{P(X=4|\lambda) \cdot f(\lambda)}{P(X=4)} \\ &= \arg\max_{\lambda} P(X=4|\lambda) \cdot f(\lambda) \\ &= \arg\max_{\lambda} log(P(X=4|\lambda)) + log(f(\lambda)) \\ &= \arg\max_{\lambda} log\left(\frac{\lambda^4 \cdot e^{-\lambda}}{4!}\right) + log\left(\frac{1}{\sqrt{10 \cdot 2\pi}} e^{-\frac{(\lambda-7)^2}{2 \cdot 10}}\right) \\ &= \arg\max_{\lambda} 4log\lambda - \lambda - log4! - log(\sqrt{20\pi}) - \frac{(\lambda-7)^2}{20} \\ &= \arg\max_{\lambda} 4log\lambda - \lambda - \frac{(\lambda-7)^2}{20} \end{aligned}$$

Differentiate with respect to λ, set to 0 and solve.

$$\frac{4}{\lambda} - 1 - \frac{(\lambda - 7)}{10} = 0$$
$$\lambda^2 + 3\lambda - 40 = 0$$
$$(\lambda - 5)(\lambda + 8) = 0$$

Since $\lambda > 0$ for any Poisson, we should choose the positive root: $\lambda = 5$.

4. **ML Short Answers**

   (a) When implementing logistic regression, a student decides to add a second intercept value. To do so they add an extra feature with value 0 to each datapoint. How will this impact training?

   There will be no impact on training since we have a value of 0, so when we compute $\theta^T x$, this new feature will have no contribution to our probability.

   (b) A Naive Bayes classifier is trained on a dataset with 100 examples. For 30 of those samples, $Y = 1$; for the rest, $Y = 0$. Instead of using a Laplace prior for $P(Y = 1)$, you use a Beta($a = 3$, $b = 4$) prior. What is your MAP estimate for the probability $Y = 1$?

   This implies that we have 5 imaginary trials with 2 successes and 3 failures. We can update our MAP probability to be:

$$P(Y = 1) = \frac{30 + 2}{100 + 5} = \frac{32}{105}$$

   (c) True or False: The log likelihood function that is used to estimate p of a Bernoulli, for a set of observations, must always be 0 or smaller.

   Yes. All probabilities are between 0 and 1, and log of a fraction is negative! For example, $\log(0.5) \approx -0.69$.