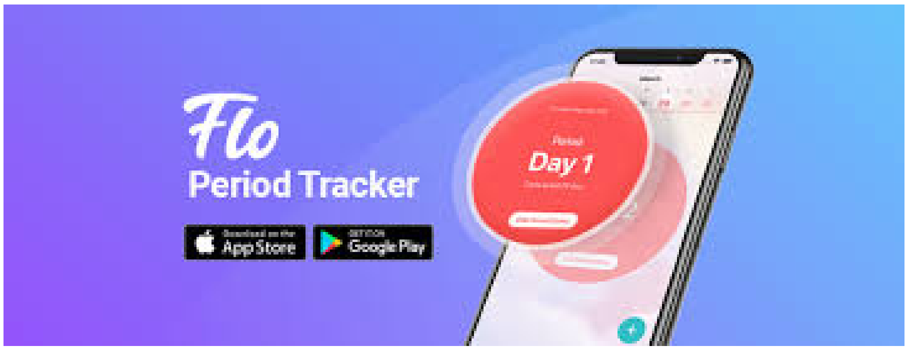
Chris Piech
CS 109

Section #9
Mar 9, 2022

Section Solution 9: Final Section

Problem 2 by David Varodayan

1. **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 a simplified version of the PDF of X.

$f(X = x) = 2 * (x - 27) * e^{-(x-27)^2}$

b. 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 $P(29.0 < X < 30.0)$?

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

Okay if expression inside integral is incorrect, as long as it's the same answer as part (a).

c. 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.

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

d. 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 write up to three sentences describing how a program can use the derivatives in order to chose 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:

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

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

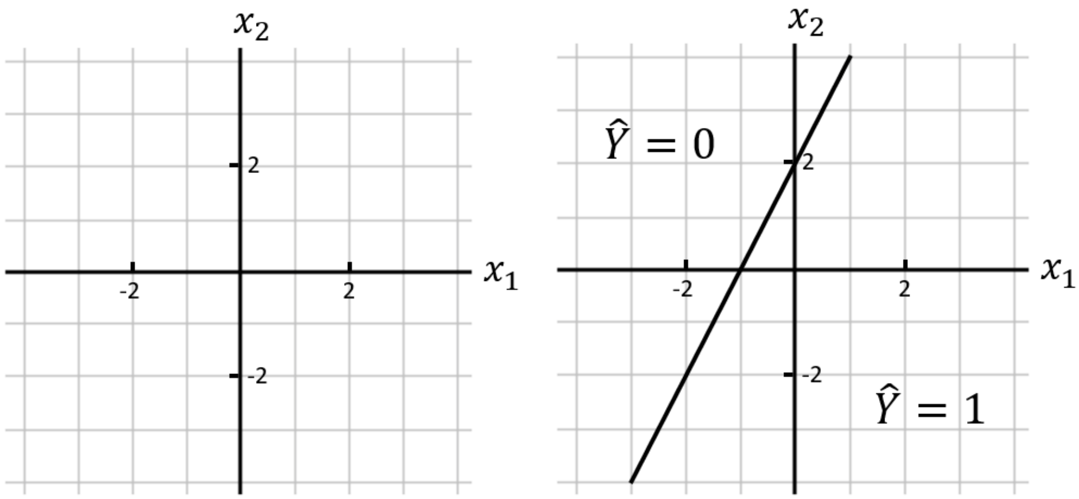
we can use gradient ascent to maximize LL. This computes gradient w.r.t each parameter α, β then moves the parameters a small step in the direction of the gradient.

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:

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

Note: Flo is a real “AI based" app that helps people track their period lengths. The real world distribution of periods is thought to be a mixture distribution between a normal and a weibell distribution [1]. This problem only has you estimate parameters for a simplified Weibull [2].

2. **Logistic regression**



Suppose you have trained a logistic regression classifier that accepts as input a data point (x_1, x_2) and predicts a class label $\hat{Y}$. The parameters of the model are $(\theta_0, \theta_1, \theta_2) = (2, 2, -1)$. On the axes, draw the decision boundary $\theta^T \mathbf{x} = 0$ and clearly mark which side of the boundary predicts $\hat{Y} = 0$ and which side predicts $\hat{Y} = 1$.

$\theta^T \mathbf{x}$ can be expanded as $2 + 2x_1 - x_2 = 0$ because $x_0 = 1$ by definition. The prediction is 1 when $\theta^T \mathbf{x} > 0$. For example, the origin $(x_1, x_2) = (0, 0)$ yields $\theta^T \mathbf{x} = 2$, which gives us the prediction $\hat{Y} = 1$.

See the graph above, to the right of the original.

3. **The Most Important Features**

Let's explore saliency, a measure of how important a feature is for classification. We define the saliency of the ith input feature for a given example $(\mathbf{x}, y)$ to be the absolute value of the partial derivative of the log likelihood of the sample prediction, with respect to that input feature $\left|\frac{\partial LL}{\partial x_i}\right|$. In the images below, we show both input images and the corresponding saliency of the input features (in this case, input features are pixels):

First consider a trained logistic regression classifier with weights θ. Like the logistic regression classifier that you wrote in your homework it predicts binary class labels. In this question we allow the values of $\mathbf{x}$ to be real numbers, which doesn't change the algorithm (neither training nor testing).

a. What is the Log Likelihood of a single training example $(\mathbf{x}, y)$ for a logistic regression classifier?

$$LL(\theta) = y \cdot \log \sigma\left(\theta^T \cdot \mathbf{x}\right) + \left(1 - y\right)\log\left[1 - \sigma\left(\theta^T \cdot \mathbf{x}\right)\right]$$

b. Calculate is the saliency of a single feature (x_i) in a training example $(\mathbf{x}, y)$.

We can calculate the saliency for a single feature as follows.

$$\begin{aligned}
LL(\theta) &= y \log z + \left(1 - y\right) \log \left(1 - z\right) && \text{where } z = \sigma\left(\theta^T \cdot \mathbf{x}\right) \\
\frac{\partial LL}{\partial x_i} &= \frac{\partial LL}{\partial z} \cdot \frac{\partial z}{\partial x_i} && \text{chain rule} \\
&= \left(\frac{y}{z} - \frac{1-y}{1-z}\right) \cdot \left(z(1-z)\theta_i\right) && \text{partial derivatives} \\
\text{saliency} &= \left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|
\end{aligned}$$

Show that the ratio of saliency for features i and j is the ratio of the absolute value of their weights $\frac{|\theta_i|}{|\theta_j|}$.

We can take the ratio as follows using our expression above.

$$\text{saliency for feature } i, S_i = \left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|, \text{ and same for } S_j$$

$$\frac{S_i}{S_j} = \frac{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|}{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|} = \frac{S_i}{S_j} = \frac{\left| \theta_i \right|}{\left| \theta_j \right|} \text{ by elimination}$$

[1]: Modeling menstrual cycle length using a mixture distribution.
https://academic.oup.com/biostatistics/article/7/1/100/243078

[2]: Weibull Distribution.
https://en.wikipedia.org/wiki/Weibull_distribution