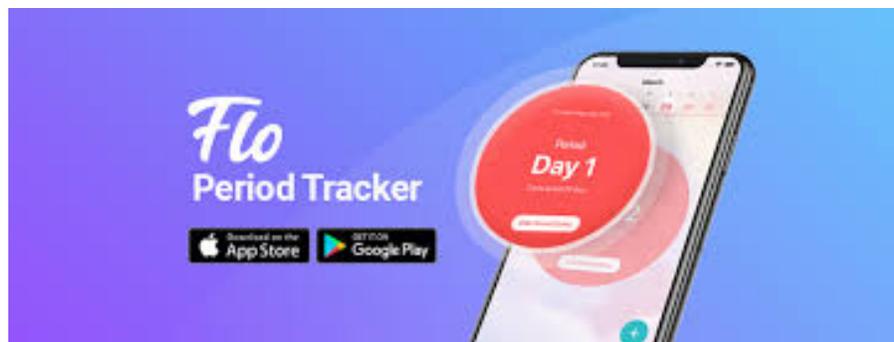Chris Piech
CS 109

Section #9
Mar 9, 2022

Section 9

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.

  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 $P(29.0 < X < 30.0)$?

  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.

  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.

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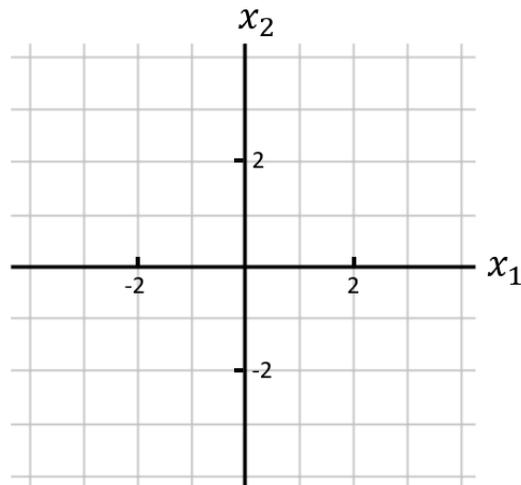
Suppose you have trained a logistic regression classifier that accepts as input a data point (x_1, x_2) and predicts a class label $\hat{Y}$. The parameters of the model are $(\theta_0, \theta_1, \theta_2) = (2, 2, -1)$. On the axes, draw the decision boundary $\theta^T \mathbf{x} = 0$ and clearly mark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

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?

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

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

[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`