

## Section 9

---

Formula reference for Logistic Regression:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$LL(\theta) = \sum_{i=0}^n y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log [1 - \sigma(\theta^T \mathbf{x}^{(i)})]$$

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^n [y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)})] x_j^{(i)}$$

### 1. Vision Test

You decide that the vision tests given by eye doctors would be more precise if we used an approach inspired by logistic regression. In a vision test a user looks at a letter with a particular font size and either correctly guesses the letter or incorrectly guesses the letter.

You assume that the probability that a particular patient is able to guess a letter correctly is:

$$p = \sigma(\theta + f)$$

Where  $\theta$  is the user's vision score and  $f$  is the font size of the letter.

Explain how you could estimate a user's vision score ( $\theta$ ) based on their 20 responses  $(f^{(1)}, y^{(1)}) \dots (f^{(20)}, y^{(20)})$ , where  $y^{(i)}$  is an indicator variable for whether the user correctly identified the  $i$ th letter and  $f^{(i)}$  is the font size of the  $i$ th letter. Solve for any and all partial derivatives required by the approach you describe in your answer.

### 2. Multiclass Bayes

In this problem we are going to explore how to write Naive Bayes for multiple output classes. We want to predict a single output variable  $Y$  which represents how a user feels about a book. Unlike in your homework, the output variable  $Y$  can take on one of the *four* values in the set  $\{\text{Like, Love, Haha, Sad}\}$ . We will base our predictions off of three binary feature variables  $X_1, X_2$ , and  $X_3$  which are indicators of the user's taste. All values  $X_i \in \{0, 1\}$ .

We have access to a dataset with 10,000 users. Each user in the dataset has a value for  $X_1, X_2, X_3$  and  $Y$ . You can use a special query method **count** that returns the number of users in the dataset with the given *equality* constraints (and only equality constraints). Here are some example usages of **count**:

- count**( $X_1 = 1, Y = \text{Haha}$ ) returns the number of users where  $X_1 = 1$  and  $Y = \text{Haha}$ .
- count**( $Y = \text{Love}$ ) returns the number of users where  $Y = \text{Love}$ .
- count**( $X_1 = 0, X_3 = 0$ ) returns the number of users where  $X_1 = 0$ , and  $X_3 = 0$ .

You are given a new user with  $X_1 = 1$ ,  $X_2 = 1$ ,  $X_3 = 0$ . What is the best prediction for how the user will feel about the book ( $Y$ )? You may leave your answer in terms of an argmax function. You should explain how you would calculate all probabilities used in your expression. Use **Laplace estimation** when calculating probabilities.

### 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  $i$ th 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  $\theta$ . 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).

- What is the Log Likelihood of a single training example  $(\mathbf{x}, y)$  for a logistic regression classifier?
- Calculate is the saliency of a single feature  $(x_i)$  in a training example  $(\mathbf{x}, y)$ .
- Show that the ratio of saliency for features  $i$  and  $j$  is the ratio of the absolute value of their weights  $\frac{|\theta_i|}{|\theta_j|}$ .