Introduction to Machine Learning
Problems: Logistic Regression

Prof. Sundeep Rangan

1. Suggest possible response variables and predictors for the following classification problems. For each problem, indicate how many classes there are. There is no single correct answer.

   (a) Given an audio sample, to detect the gender of the voice.

   (b) A electronic writing pad records motion of a stylus and it is desired to determine which letter or number was written. Assume a segmentation algorithm is already run which indicates very reliably the beginning and end time of the writing of each character.

2. Suppose that a logistic regression model for a binary class label $y = 0, 1$ is given by

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

   where $\boldsymbol{\beta} = [1, 2, 3]^\mathsf{T}$. Describe the following sets:

   (a) The set of $\mathbf{x}$ such that $P(y = 1|\mathbf{x}) > P(y = 0|\mathbf{x})$.

   (b) The set of $\mathbf{x}$ such that $P(y = 1|\mathbf{x}) > 0.8$.

   (c) The set of x_1 such that $P(y = 1|\mathbf{x}) > 0.8$ and $x_2 = 0.5$.

3. A data scientist is hired by a political candidate to predict who will donate money. The data scientist decides to use two predictors for each possible donor:

   - x_1 = the income of the person (in thousands of dollars), and
   - x_2 = the number of websites with similar political views as the candidate the person follow on Facebook.

   To train the model, the scientist tries to solicit donations from a randomly selected subset of people and records who donates or not. She obtains the following data:

Income (thousands $), x_{i1}	30	50	70	80	100
Num websites, x_{i2}	0	1	1	2	1
Donate (1=yes or 0=no), y_i	0	1	0	1	1

   (a) Draw a scatter plot of the data labeling the two classes with different markers.

(b) Find a linear classifier that makes at most one error on the training data. The classifier should be of the form,

$$\hat{y}_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0, \end{cases} \qquad z_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i + b.$$

What is the weight vector $\mathbf{w}$ and bias b in your classifier?

(c) Now consider a logistic model of the form,

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + e^{-z_i}}, \qquad z_i = \mathbf{w}^\mathsf{T}\mathbf{x}_i + b.$$

Using $\mathbf{w}$ and b from the previous part, which sample i is the *least* likely (i.e. $P(y_i|\mathbf{x}_i)$ is the smallest). If you do the calculations correctly, you should not need a calculator.

(d) Now consider a new set of parameters

$$\mathbf{w}' = \alpha\mathbf{w}, \quad b' = \alpha b,$$

where $\alpha > 0$ is a positive scalar. Would using the new parameters change the values $\hat{y}$ in part (b)? Would they change the likelihoods $P(y_i|\mathbf{x}_i)$ in part (c)? If they do not change, state why. If they do change, qualitatively describe the change as a function of α.

4. Suppose we collect data for a group of students in a machine learning class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6,\ \beta_1 = 0.05,\ \beta_2 = 1$.

   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

5. The loss function for logistic regression for binary classification is the binary cross entropy defined as

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ln(1 + e^{z_i}) - y_i z_i$$

where $z_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ for two features $x_{1,i}$ and $x_{2,i}$.

   (a) What are the partial derivatives of z_i with respect to β_0, β_1, and β_2.

   (b) Compute the partial derivatives of $J(\boldsymbol{\beta})$ with respect to β_0, β_1, and β_2. You should use the chain rule of differentiation.

   (c) Can you find the close form expressions for the optimal parameters $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by putting the derivatives of $J(\boldsymbol{\beta})$ to 0? What methods can be used to optimize the loss function $J(\boldsymbol{\beta})$?