Data 100, Spring 2022	**April 15, 2022**

Discussion #12

Logistic Regression

1. Suppose we are given the following dataset, with two features ($\mathbb{X}_{:,0}$ and $\mathbb{X}_{:,1}$) and one response variable (y).

$\mathbb{X}_{:,0}$	$\mathbb{X}_{:,1}$	y
2	2	0
1	-1	1

Here, $\vec{x}$ corresponds to a single row of our data matrix, not including the y column. We write all vectors for this class as column vectors. Thus, we can write, $\vec{x}_1$, as $\vec{x}_1 = [2 \quad 2]^T$. Note that there is no intercept term!

You run an algorithm to fit a model for the probability of $Y = 1$ given $\vec{x}$:

$$\mathbb{P}\left(Y = 1 \mid \vec{x}\right) = \sigma(\vec{x}^T\theta)$$

where

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

Your algorithm returns $\hat{\theta} = \left[-\frac{1}{2} \quad -\frac{1}{2}\right]^T$

(a) Calculate $\hat{\mathbb{P}}\left(Y = 1 \mid \vec{x} = [1 \quad 0]^T\right)$.

(b) The empirical risk using log loss (a.k.a., cross-entropy loss) is given by the following expression. Remember, whenever you see $\log$ in this course, you must assume the natural logarithm (base-e), unless explicitly told otherwise.

$$R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{\mathbb{P}}\left(Y = 1 \mid \vec{x}_i\right) + (1 - y_i)\log \hat{\mathbb{P}}\left(Y = 0 \mid \vec{x}_i\right)\right)$$

Suppose we run a different algorithm and obtain $\hat{\theta} = \begin{bmatrix}0 & 0\end{bmatrix}^T$. Calculate the empirical risk for this new $\hat{\theta}$ on our dataset.

(c) Is our dataset linearly separable? If so, write the equation of a hyperplane that separates the two classes.

(d) Does either of our fitted models $\hat{\theta} = \begin{bmatrix}0 & 0\end{bmatrix}^T$ or $\hat{\theta} = \begin{bmatrix}-\frac{1}{2} & -\frac{1}{2}\end{bmatrix}^T$ minimize cross-entropy loss?

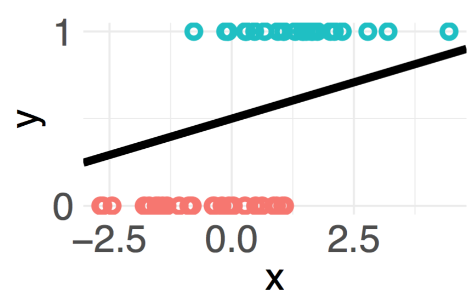
(e) Consider using a linear regression model such that $\mathbb{X}\hat{\theta} = \hat{\mathbb{Y}}$. What is the MSE of the optimal linear regression model given the dataset we used above for logistic regression?

(f) Assume we add the data point $[3, -2]$ to the design matrix such that our resulting design matrix is as follows.

$\mathbb{X}_{:,0}$	$\mathbb{X}_{:,1}$	y
2	2	0
1	-1	1
3	-2	0

Comment on whether it is possible to achieve perfect accuracy using a logistic regression model and whether it is possible to achieve zero MSE using a linear regression model trained on the new data. Assume that we don't use an intercept term.

2. Suppose your friend obtains a dataset with a single feature x. Your friend argues that the data are linearly separable by drawing the line on the following plot of the data.



(a) Argue whether or not your friend is correct. Note: this question refers to a binary classification problem with a single feature.

(b) Suppose you use gradient descent for a **fixed number of iterations** to train a logistic regression model on two design matrices $\mathbb{X}_a$ and $\mathbb{X}_b$. After training, you find that the training accuracy for $\mathbb{X}_a$ is 100% and the training accuracy for $\mathbb{X}_b$ is 98%. What can you say about whether the data is linearly separable for the two design matrices?

3. Suppose we train a binary classifier on this dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) Confusion matrix.

Hint: The first row contains the true negatives and false positives, and the second row contains false negatives and true positives (in that order).

(b) The precision of our classifier. Write your answer as a simplified fraction.

(c) The recall of our classifier. Write your answer as a simplified fraction.