CS344 Exercise 2

Task 1: Logistic regression

Answer the following TRUE/FALSE or multiple choice questions.

Machine learning algorithms are more likely to perform better on testing data than on training data

TRUE FALSE

Logistic regression solves regression problems

TRUE FALSE

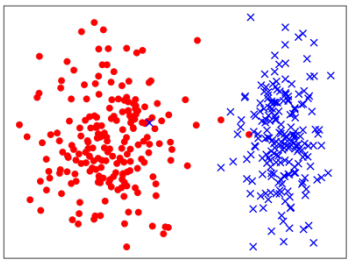
Logistic regression can be used on data even if data are not linearly separable

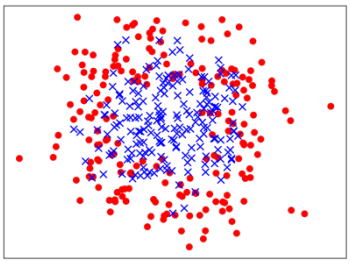
TRUE FALSE

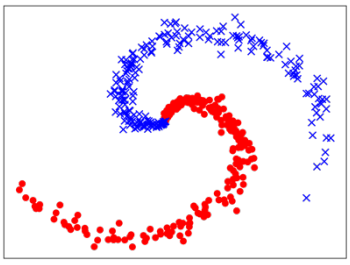
The sigmoid function returns either 0 or 1

TRUE FALSE

Circle the dataset for which you expect logistic regression to perform the worst (to view the points more clearly, check the PDF online)







Task 2: Calculating predictions, loss, and cost for a logistic regression model

Suppose a model has parameters $\boldsymbol{w}$ = (-1.3) and b = 2.5.
For a data point $\boldsymbol{x}$ = (4.7), what value $\hat{y}$ would be predicted by forward propagation?

Suppose a model has parameters $\boldsymbol{w}$ = (0.8, -2.4, -1.9) and b = -0.3.
For a data point $\boldsymbol{x}$ = (5.3, -0.6, 2.7), what value $\hat{y}$ would be predicted by forward propagation?

For a data point in class **0**, suppose a model predicts $\hat{y}$ = 0.27.
What is the *loss* accociated with the prediction for this point?

Suppose a model has parameters $\boldsymbol{w}$ = (-3.3, 1.8, -4.2) and b = 6.8.
For a data point $\boldsymbol{x}$ = (-2.1, 8.6, 6.7) with label y = (1), what is the *loss* associated with the prediction for this point?

For three data points with class labels $y_1 = 0$, $y_2 = 1$, and $y_3 = 0$, suppose a model makes predictions $\hat{y}_1 = 0.4$, $\hat{y}_2 = 0.8$, and $\hat{y}_3 = 0.7$.
What is the *cost* accociated with the predictions for these three points?

Task 3: Logistic regression with `sklearn`

Download the Jupyter Notebook for Exercise 2 from the course website. Open the Notebook in your web browser and work through it. As you work through the Notebook, answer the following questions.

Looking at the plot for the new dataset, do you expect $\boldsymbol{w}[0]$ or $\boldsymbol{w}[1]$ to be larger in magnitude? Do you expect $\boldsymbol{w}[1]$ to have a positive or negative sign?

How many *testing* data points are misclassified by the model?

What is the accuracy of the model on the *test* data?

Which four features have the highest magnitude weights (ignoring sign, i.e., absolute value)?

What two features are least predicative of heart disease, i.e., have weights closest to 0?

Are highly educated people more likely or less likely to have heart disease?

How many features does the breast cancer data have after *ordinal* encoding?

How many features does the breast cancer data have after *one-hot* encoding?

What is the *testing* accuracy when the features are *ordinal* encoded?

What is the *testing* accuracy when the features are *one-hot* encoded?

How many features does the mushroom data have after *ordinal* encoding?

How many features does the mushroom data have after *one-hot* encoding?

What is the *testing* accuracy when the features are *ordinal* encoded?

What is the *testing* accuracy when the features are *one-hot* encoded?

CS344 Exercise 2 Final Page

In the *TIME* column, please estimate the time you spent on this exercise. Please try to be as accurate as possible; this information will help us to design future exercises.

PART	TIME
Exercise	