Course › Week 4... › Progra... › Progra...

Programming Assignment 4

Click this link to download the Diabetes Regression notebook and then complete problems 1-4.

Click this link to download the mystery.dat file which will help you complete problem 5.

Click this link to download the Sentiment Logistic Regression notebook.

Problem 1

1/1 point (graded)

This problem is based on the *Diabetes Regression notebook*. You should work through that notebook before entering your answers here.

If a single feature is to be used to predict y, the best choice (the one that yields the smallest MSE) is feature 2 ('body mass index'). What is the second-best choice? Your answer should be the feature number $(0 - 9)$.

8 ✓

8

Submit

Problem 2

2/2 points (graded)

Use the `split_data` procedure to create training/test splits of various sizes. In particular, try training set sizes of 20, 50, 100, and 200. In each case, record the training error and test error *when using all features for prediction*.

For a training set size of 100, what are the training MSE and test MSE (just round to the nearest integer)?

Training MSE =

2883.7785 ✓

2883.7785

Test MSE =

3583.008 ✓

3583.008

Submit

Problem 3

1/1 point (graded)

What *rough* trends do you observe as the training set size increases (from, say, 20 to 400)? Select all that apply.

☑ The training error increases

☑ The test error decreases

☑ The gap between the training and test error decreases

✓

Submit

Problem 4

1/1 point (graded)

What is the single best explanation for these trends? Choose one of the following.

○ With more training data, we get better estimates of training error.

● With more training data, we learn a more accurate model.

○ The error is proportional to the amount of data.

✓

Submit

Problem 5 relates to finding relevant features.

Problem 5

1/1 point (graded)

The file `mystery.dat` contains pairs (x, y), where $x \in \mathbb{R}^{100}$ and $y \in \mathbb{R}$. There is one data point per line, with comma-separated values; the very last number in each line is the y-value.

In this data set, y is a linear function of just *ten* of the features in x, plus some noise. Your job is to identify those ten features.

Which of the following contain only relevant features?

(Think of the feature numbers as being in the range 1 to 100, but be aware that Python indexes arrays starting at zero.)

- ☐ 1,5,7,19,44
- ☑ 2,3,13,17,29
- ☐ 3,7,13,19,44
- ☐ 5,23,24,51,61

✓

Submit