

# CS305 Exercise 2

## Task 1: Companies use of machine learning to target advertisements

Read [this article](#) about how data and machine learning may be used to target product advertisements at different groups of consumers. Please answer (minimum 100 words) the following questions.

In your opinion, is it problematic (e.g., violating someone's privacy) for companies to gather data on their customer's purchasing habits and use these data to send targeted product advertisements? Assuming a company doesn't share or sell the data, are there other ways a company might use data on customers' shopping habits that you would consider to be problematic?

## **Task 2: $k$ nearest neighbors and distance measures**

What is the *training* accuracy (i.e., accuracy on the training data) of a  $k$  nearest neighbor classifier when  $k=1$ ?

Suppose when using a  $k$  nearest neighbor classifier on  $n$  training examples, you set  $k$  equal to  $n$ . What will be true about the classifier's predictions on the *test* data?

Suppose it takes a  $k$  nearest neighbor classifier  $N$  seconds to classify a *test* data point. If you double the number of *training* examples, about how long would you expect the classifier to take when classifying a *test* data point?

Suppose it takes a decision tree classifier  $N$  seconds to classify a *test* data point. If you double the number of *training* examples and re-build the decision tree, about how long would you expect the classifier to take when classifying a *test* data point?

Using the  $L^2$  norm, what is the distance between the following two four-dimensional points?  
(4, 7, 1, 10) and (8, 3, 3, 5)

Using the  $L^1$  norm, what is the distance between the following two four-dimensional points?  
(4, 7, 1, 10) and (8, 3, 3, 5)

Give an example of two distinct (non-equal) points in two-dimensional space such that the distance between the points is the same when computed using the  $L^1$  norm or the  $L^2$  norm.

### **Task 3: Feature scaling**

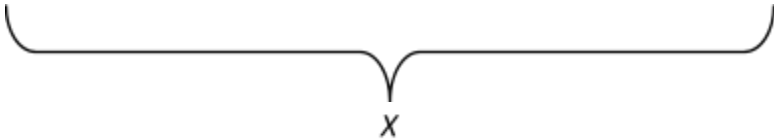

Feature scaling is useful when features have what property?

Feature scaling is unlikely to be useful when features have what property?

When using a decision tree classifier, is using feature scaling helpful? Why? When using a  $k$  nearest neighbor classifier, is using feature scaling helpful? Why?

Consider the following training data consisting of four examples with three features, where each training example belongs to one of two classes (0 or 1).

Feature 1	Feature 2	Feature 3	Class
-9	100	18	0
-1	85	11	0
-6	120	16	1
-4	95	19	1

Compute the 4x3 training matrix  $X$  after subtracting each feature's mean.

Compute the 4x3 training matrix  $X$  after subtracting each feature's mean and dividing by each feature's standard deviation.

Suppose we use a  $k$  nearest neighbor classifier with  $k=1$ . What is the predicted class for the test data point  $(-3, 88, 16)$  if feature scaling is not used?

Suppose we use a  $k$  nearest neighbor classifier with  $k=1$ . What is the predicted class for the test data point  $(-3, 88, 16)$  if feature scaling is used?

#### **Task 4: Using $k$ nearest neighbors for classifying images of handwritten digits**

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.

How many example images of handwritten digits are there? How many features does each example have?

How many images are in the training set? How many images are in the test set?

What is the accuracy of the kNN classifier in correctly identifying handwritten digits?

What is the accuracy of the kNN classifier using the centered (i.e., mean subtracted) data? Did centering the data improve the accuracy? Why or why not?

What is the accuracy of the kNN classifier using the simpler, binarized data? Did the accuracy increase or decrease? By a lot or a little? What is one advantage of using simpler features?

What is the accuracy of the Random Forest classifier on this handwritten digits dataset? Is it better or worse than the accuracy of the kNN classifier?

Which takes longer to make predictions on new (e.g., test) data, the kNN classifier or the random forest classifier? Are you surprised by this?

What is the accuracy of the kNN classifier in correctly identifying fashion items in the testing data? What is the accuracy of the kNN classifier when run on the simpler binarized data? What is the accuracy of the Random Forest classifier on this fashion dataset?

# CS305 Exercise 2 Final Page

Name(s): \_\_\_\_\_

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.

<b>PART</b>	<b>TIME</b>	<b>SCORE</b>
Exercise		