MAT 345 - Homework 4

Due Monday, October 15, 2018, in class

1. (3 points) Suppose you want to help your friend and use your knowledge from MAT 345 to build a simple spam filter. You decide to use Naive Bayes and start with three words: *free*, *trip*, and *insurance*. You do NOT use smoothing and work with the sample of 900 emails, out of which you counted 351 to be spam. You count how many times each of the three words appears in each set of messages:

	spam	not spam
	351	549
free	101	3
trip	76	52
insurance	87	21

   (a) Find the probability the message is spam, given the word *free* appears in the message.

   (b) Find the probability the message is spam, given the word *trip* appears in the message.

   (c) Find the probability the message is spam, given the word *insurance* does NOT appear in the message.

   (d) Suppose a new message contains the words *free* and *trip*, but NOT the word *insurance*. Find the probability it is a spam message.

2. (2 points) Suppose we consider the labeled data points given below:

x_1	x_2	Label
1	1	-1
3	0	-1
1	3	1
3	3	1
2	2	1

   Use the 3-nearest neighbors algorithm to label the data point $(x_1, x_2) = (2, 1)$. Show your steps.

3. (5 points) Consider the data in the file HW4-data.xls regarding videogame ratings (from a specific publisher). The data is already split for you in the *training set A*, the *test set B* and the *unclassified set C*. Perform the k-nearest neighbors algorithm to predict the rating of a videogame based on release year, user score and critic score, by following the steps:

   (a) Use the Euclidean distance to measure closeness between data points. For fixed $1 \leq k \leq 6$,

(i) For every data point in the *test set* B, **predict** its rating by the most common rating among its k nearest neighbors.

(ii) Count how many data points from B were predicted correctly.

(iii) Compute and **output** the accuracy $= \dfrac{\#\text{correct predictions}}{\text{total } \# \text{ points in } B}$.

(b) Evaluate the model using the accuracy metric, that is, pick k that maximizes how often the correct label of the testing data was predicted by the k-NN algorithm. **Output** optimal k.

(c) Use the k you found in (b) to label the set of points in the *unclassified set* C. **Output** newly classified set C.