CSE 151: Machine learning **Winter 2019**

Homework 2

Submission instructions:

- Please type up your solutions.
- If a problem asks for a numerical answer, you need only provide this answer. There is no need to show your work, unless you would like to.
- Upload the PDF file for your homework to `gradescope` by 6pm on Tuesday January 22.

Part A: Short-answer questions

1. For the point $x = (1, 2, 3, 4, \ldots, d)$ in $\mathbb{R}^d$, what are $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$? Give your answers as functions of d.
2. How many points in $\mathbb{R}^d$ have $\|x\|_1 = \|x\|_2 = 1$? Give your answer in terms of d.
3. Which of these distance functions is a *metric*? If it is a metric, you need only say so. If it isn't, state which property of metrics it violates and give a specific counterexample showing this violation.
   (a) Let Σ be a finite set and $\mathcal{X} = \Sigma^m$. The *Hamming distance* on $\mathcal{X}$ is $d(x, y) =$ # of positions on which x and y differ.
   (b) Squared Euclidean distance on $\mathbb{R}^m$, that is, $d(x, y) = \sum_{i=1}^m (x_i - y_i)^2$. (It might be easiest to consider the case $m = 1$.)
4. Suppose d_1 and d_2 are two metrics on a space $\mathcal{X}$. Define d to be their sum: $d(x, y) = d_1(x, y) + d_2(x, y)$. Is d necessarily a metric? Either show that it is or give a counterexample.
5. For each of the following prediction tasks, state whether it is best thought of as a *classification* problem or a *regression* problem.
   (a) Based on sensors in a person's cell phone, predict whether they are walking, sitting, or running.
   (b) Based on sensors in a moving car, predict the speed of the car directly in front.
   (c) Based on a student's high-school SAT score, predict their GPA during freshman year of college.
   (d) Based on a student's high-school SAT score, predict whether or not they will complete college.
6. Define the *probability space* for each of the following experiments.
   (a) A fair coin is tossed.
   (b) A fair die is rolled.
   (c) A fair coin is tossed ten times in a row.

7. Consider a sample space $\Omega = \{a, b, c, d\}$ in which the probability of outcome a is $1/2$, the probability of outcome b is $1/8$, and the probability of outcome c is $1/4$.
   (a) What is the probability of outcome d?
   (b) Define *event* $A = \{a, b, c\}$. What is the probability of event A?
   (c) Define event $B = \{a, c, d\}$. What is the probability of event $A \cap B$?
8. Two fair dice are rolled. What is the probability that:
   (a) Their sum is 10, given that the first roll is a 6?
   (b) Their sum is 10, given that the first roll is an even number?
   (c) They have the same value?
9. A certain genetic disease occurs in 5% of men but just 1% of women. Let's say there are an equal number of men and women in the world. A person is picked at random and found to possess the disease. What is the probability, given this information, that the person is male?

Part B: Programming assignment

In this problem, you will use nearest neighbor to classify patients' back injuries based on measurements of the shape and orientation of their pelvis and spine.

The data set contains information from 310 patients. For each patient, there are: six numeric features (the x) and a label (the y): 'NO' (normal), 'DH' (herniated disk), or 'SL' (spondilolysthesis). We will divide this data into a training set with 250 points and a separate test set of 60 points.

- Download the data set `spine-data.txt`. You can load it into Python using the following.

```
import numpy as np
# Load data set and code labels as 0 = 'NO', 1 = 'DH', 2 = 'SL'
labels = ['NO', 'DH', 'SL']
data = np.loadtxt('spine-data.txt', converters={6:  lambda s:  labels.index(s)})
```

  This converts the labels in the last column into 0 (for 'NO'), 1 (for 'DH'), and 2 (for 'SL').

- Split the data into a training set, consisting of the *first* 250 points, and a test set, consisting of the remaining 60 points.
- Code up a nearest neighbor classifier based on this training set. Try both ℓ_2 and ℓ_1 distance.

Now do the following exercises, to be turned in.

1. What error rates do you get on the test set for each of the two distance functions?
2. For each of the two distance functions, give the *confusion matrix* of the NN classifier. This is a 3×3 table of the form:

	NO	DH	SL
NO			
DH			
SL			

   The entry at row DH, column SL, for instance, contains the number of test points whose correct label was DH but which were classified as SL.