K Nearest Neighbors Self-Check

The purpose of this self-check is to make sure you understand key concepts for the algorithms presented during the module and to prepare you for the programming assignment. As you work through problems, you should always be thinking "how would I do this in code? What basic data structures would I need? What operations on those basic data structures?"

Euclidean Distance

The basis of most kNN systems is Euclidean Distance. The purpose of this self-check to make sure you can calculate it.

Consider the following six points of "tidy data" with x1 and x2 as X and y as y:

#	x1	x2	y
1	0.23	0.81	0.18
2	0.42	0.78	0.33
3	0.64	0.23	0.14
4	0.87	0.19	0.17
5	0.76	0.43	0.32
6	0.39	0.63	?

Here is the definition of Euclidean Distance:

$$d(p_1, p_2) = \left(\Sigma_d^D \left(p_1^d - p_2^d\right)^2\right)^{\frac{1}{2}}$$

where D is the number of dimensions (the number of features in X).

1. What are the 3 nearest neighbors to Point 6? What is the predicted value of y for Point 6?

2. Do you need to take the square root to compare distances?

Performance Metrics

Use the following Confusion Matrix when answering the questions below:

		Actual	
		Positive	Negative
Predicted	Positive	329	35
	Negative	87	357

1. What is the accuracy?
2. What is the error?
3. What is the precision?
4. What is the recall?

Mean Squared Error

Consider the following five predictions for a regression model:

y	y_hat	(y-y_hat)	(y-y_hat)^2	
3.78	3.32			
4.82	5.21			
2.83	2.97			
2.76	2.37			
3.48	3.29			

The formula for MSE (mean squared error) is MSE = $\frac{1}{n}\Sigma(y_i - \hat{y}_i)^2$

1. What is the MSE of the model?
2. What is the mean of y?
3. If you used the mean of y as a predictor of y (y_hat), what would *its* MSE be?
4. Look up for the formula for variance of a population. Surprised?

Learning Curves

For each of the questions below, simply draw the axes and curves, given them the shape you expect.

1. Give an example of a set of learning curves where collecting more data may help.
2. Give an example of a set of learning curves where collecting more data probably will not help.