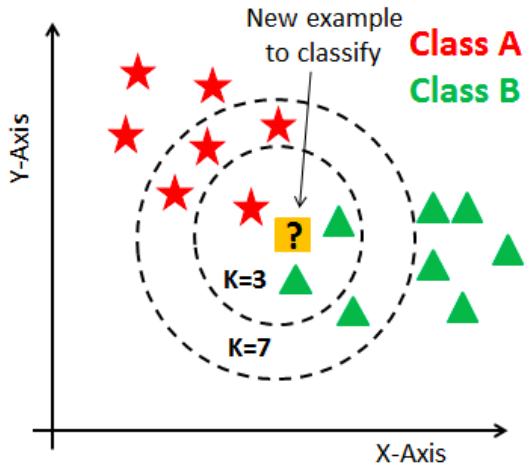
National University of Singapore
School of Computing
Tutorial 3:
K-NEAREST NEIGHBORS

K-Nearest Neighbours is a **supervised learning technique** that is used mostly for classification, but sometimes for regression as well. The 'K' in KNN is the number of nearest neighbors used to classify/predict a test sample.



Produce a Jupyter Notebook code to answer questions 1 to 7.

1. Import the Wisconsin Breast Cancer dataset from Sklearn datasets. What format is it in? Inspect the keys. What are the names of the keys?

2. Create your X variable (the features) and the y variable (the labels).

3. Create a train-test split in your data using the SKLearn Train-Test split library.

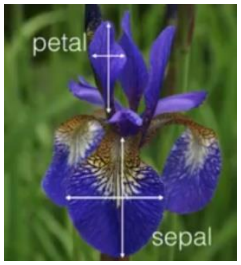
4. Fit the SKLearn KNeighborsClassifier with a n_neighbors value of 3. What is the accuracy score?

5. Create predictions on the test set and use the SKLearn Classification_report library to generate a classification report. Discuss your results.

6. Visualize the dataset you have as a histogram. Normalize your data using SKLearn's standard scaler and re-run the classifier on the data. Why do we need to normalize our data, and why does our result change? Discuss the results that you have obtained.

7. Use an SVM to conduct the same classification. What are the differences in result?

8. Produce a Jupyter Notebook code and use KNN classification on the IRIS dataset contained in the SKLearn datasets library (i.e. sklearn.datasets.load_iris).



   a) What are the features and species of flowers that are measured in this dataset?
   b) Print first 10 measurements taken in this dataset.
   c) Using only Sepal length and Sepal width to classify flowers, create a color-coded scatterplot.
   d) Using only Petal length and Petal width to classify flowers, create a color-coded scatterplot.
   e) Choose two features and classify using K nearest neighbor and plot the decision boundaries using np.meshgrid, np.ravel and plt.colormesh.

9. Produce a Jupyter Notebook code and use KNN classification on the wine dataset contained in the SKLearn datasets library (i.e. sklearn.datasets.load_wine). What is the optimal value for n_neighbors? What is the accuracy score? Produce a classification report and discuss your results.

1-7 by John Ang
8-9 by Yihong Lan