Problem Set 1

*Start Date: Sep. 14, 2020
*Due Date: Sep. 28, 2020

Notes:

- Each homework problem must be done individually.
- Please make sure your full name, student ID, and assignment number appear at the top of each file you submit.
- Please answer questions in the order they are assigned.
- Please submit a **PDF** version. Homework may be typed or handwritten, but handwritten submissions need to be legible.
- Please submit on the Web Learning platform. Do NOT Email or Wechat your assignment to the instructor or TAs.
- No late homework is accepted. No exceptions.
- You can answer the question using English or Chinese.

1. Try to choose one **Classification** Model from scikit-learn Models and explain its **theory and meaning of parameters**. Loading the breast cancer dataset and apply the model you choose on the dataset, show your code and your precision on the task.

2. Classfication with Nearest Neighbours. In this question, you will use the scikit-learn's KNN classifer to classify real vs. fake news headlines. The aim of this question is for you to read the scikit-learn API and get comfortable with training/validation splits. We will use a dataset of 1298 "fake news" headlines (which mostly include headlines of articles classfied as biased, etc.) and 1968 "real" news headlines, where the "fake news" headlines are from https://www.kaggle.com/mrisdal/fake-news/data and "real news" headlines are from https://www.kaggle.com/therohk/million-headlines.
   Write a function load_data which loads the data, preprocesses it using a CountVectorizer (http://scikit-learn.org/stable/modules/classes.html#module-sklearn.feature_extraction.text), and splits the entire dataset randomly into 80% training, 10% validation, and 10% test examples.

3. Read the paper "Efficient Person Search: An Anchor-Free Approach" and give your review. (Hint: you can refer to "Tips and advice when you review a scientific paper" and browser the OpenReview website)

4. Write your Project proposal and your plan.(Please choose one competition from "AI innovation and application competition" ,"QQ Browser 2021 AI algorithm competition" or "ML Reproducibility Challenge 2021")