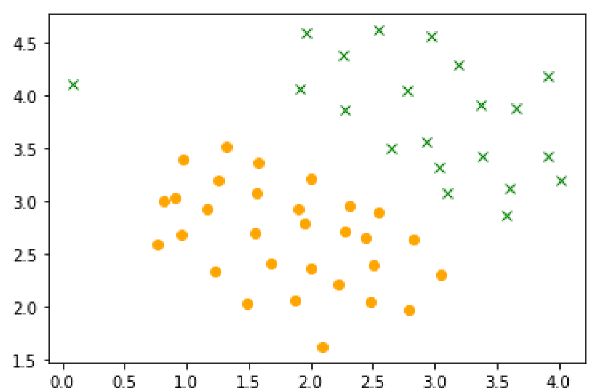
**Optimization in Machine Learning (2020 Winter)**
**Assignment 2**

**Instructions:** For the code parts, submit the completed files on Canvas (both .py file and .ipynb file are accepted). For the free response parts, type your solution into a seperate electronic document (.pdf file). Physical submissions will **NOT** be accepted.
To submit, compress all your files into a single compressed file (.zip file). E-mail a softcopy of your code and answers to rkwon@mie.utoronto.ca.
If you have any questions about this assignment, please e-mail yhe@mie.utoronto.ca.

### 1. Linear Support Vector Machine

Download the dataset 'prob1data.csv'. The dataset consists of two features as first two columns and its classification as the third column (0 and 1 refer to two distinct types.) A plot of the dataset is given below:




In this problem, you will solve both primal and dual optimization problem of the linear soft-margin support vector machine using CVXPY and analyze the results.

**(1a)** (Code + Free Response) Complete the function '`LinearSVM_Primal`' that solve the **primal** optimization problem of the linear SVM. Using the entire dataset, $C = 1$, solve the optimization problem and report:
(1) The optimal decision boundary.
(2) The optimal support vectors.
(3) The solution time.

**(1b)** (Code + Free Response) Complete the function '`LinearSVM_Dual`' that solve the **dual** optimization problem of the linear SVM. Using the entire dataset, $C = 1$, solve the optimization problem and report:
(1) The optimal dual solution.
(2) The optimal decision boundary.
(3) The optimal support vectors.
(4) The solution time.

**(1c)** (Free Response) Discuss if the decision boundary of the linear SVM will change with increased and decreased $C$ value. If the decision boundary changes, briefly discuss how it changes with $C$ and why. If the decision boundary does not change with $C$, discuss the reason.

**(1d)** (Code + Free Response) Complete the function 'Linearly_separable' that output 1 if the dataset is linearly separable and 0 otherwise. Determine if the given dataset is linearly separable. For any given dataset with multiple features, how can one conclude if the dataset is linearly separable based on the optimal solution (optimal decision boundary) and optimal objective function value solved? (Hint: consider varying $C$ values.)

In this following problems, we will consider an alternative soft-margin method, known as the $l_2$ norm soft margin SVM. This new algorithm is given by the following primal optimization problem (notice that the slack penalties are now squared, $n$ is the total number of datapoints, $(x^{(i)}, y^{(i)})$ is the $i^{th}$ datapoint):

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|_2^2 + \frac{C}{2}\sum_{i=1}^{n} \xi_i^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad \forall i = 1, ..., n. \end{aligned}$$

**(1e)** (Free Response) Notice that for the usual soft margin SVM implemented in (1a) $\xi_i \geq 0$ is a important constraints. Discuss why the constraint $\xi_i \geq 0$ is not needed in the primal $l_2$ norm soft margin SVM problem.

**(1f)** (Code + Free Response) Complete the function 'LinearSVM_l2_Primal' that solve the given **primal** optimization problem of the $l_2$ norm soft margin SVM. Using the entire dataset, $C = 1$, solve the optimization problem and report:
(1) The optimal decision boundary.
(2) The optimal support vectors.

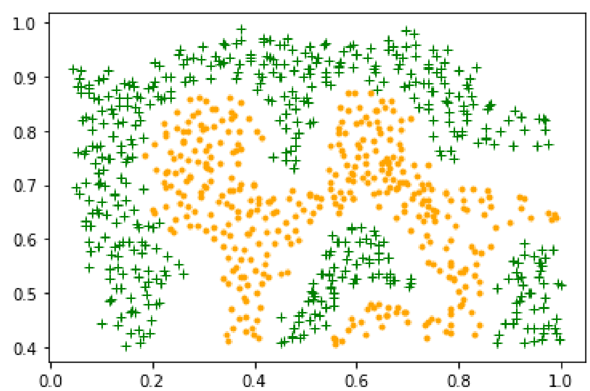
**(1g)** (Code + Free Response) Write down the dual of the primal $l_2$ norm soft margin SVM optimization problem. Complete the function 'LinearSVM_l2_Dual' that solve the given **primal** optimization problem of the $l_2$ norm soft margin SVM. Using the entire dataset, $C = 1$, solve the optimization problem and report:
(1) The optimal dual solution.
(2) The optimal decision boundary.
(3) The optimal support vectors.

**(1h)** (Code + Free Response) Plot the decision boundaries of the $l_1$ and $l_2$ soft margin SVM decision along with the dataset. Compare the two methods.

**(1i)** (Free Response) Discuss how one could choose proper $C$ in implementation of linear SVM.

**2. Kernel Support Vector Machine and Application**

Download the dataset 'prob2data.csv'. The dataset consists of two features as first two columns and its classification as the third column (0 and 1 refer to two distinct types.) A plot of the dataset is given below:




In the following problems, you will implement a kernel support vector machine with Gaussian kernel. The Gaussian kernel is given as:

$$K(x^{(i)}, x^{(j)}) = \exp(-\frac{1}{2\sigma^2}\|x^{(i)} - x^{(j)}\|_2^2)$$

**(2a)** (Code) Complete the function 'gaussian_kernel_sigma' that returns a function '`gaussian_kernel`' with the specified $\sigma$ value.

**(2b)** (Code + Free Response) Use 'SVC' from 'sklearn.svm' and 'gaussian_kernel_sigma' coded in (2a) to build a kernel SVM to classify the train data `X_train`. Use $C = 1$ and $\sigma = 0.1$. Report:
(1) Number of support vectors.
(2) Prediction error (ratio) in test set `X_test`.
(3) Plot decision boundary approximately.

Download the files '`votes.csv`'. This dataset votes consists of over 3000 counties in the United States along with their socioeconomic and demographic information and voting records in the 2016 US election; each row corresponds to a single county.

**(2c)** (Code) The response variable will be prefer trump, which is 0 or 1 indicating whether the percentage of people who voted for Trump in that county is greater than that who voted for Clinton. Compute the response variable `y`.

**(2d)** (Code + Free Response) Use 'SVC' from 'sklearn.svm' to implement polynomial kernel SVM with $C = 10.0$, `max_iter=1e6`. Implement SVM with kernel degree set to 1, then 2, 3, 4, and 5. For each model, report:
(1) Number of support vectors.
(2) Prediction error (ratio) in test set `X_train`.
(3) Prediction error (ratio) in test set `X_test`.

**(2e)** (Free Response) Based on the 5 models trained in (2d), how does the predictive error change with the degree of the polynomial kernel? Explain why. How does the number of support vectors change with the degree of the polynomial kernel? Explain why.