Rahul Ghosh
Student ID: 5476965
Teammate: Shreyasi Pal
Student ID: 5483657

Assignment#2
CSci 5525

Solution of Question 1

a. Primal Objective function,

$$\min_{w,\{\xi_i\}} \frac{1}{2}||w||^2 \ + C\sum_i \xi_i \quad such\ that \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0,\ \forall i \tag{1}$$

Lagrangian form,

$$\frac{1}{2}||w||^2 \ + C\sum_i \xi_i \ - \sum_i \alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) - \sum_i \mu_i \xi_i \tag{2}$$

Lagrangian Dual,

$$\max L^*(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{3}$$

KKT Conditions for the above Lagrangian Dual
Primal Feasibility: $y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$
Dual Feasibility: $\alpha_i \geq 0, \mu_i \geq 0$
Complementary Slackness: $\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0, \mu_i \xi_i = 0$
Gradient Condition: $w - \sum_i \alpha_i y_i x_i = 0, \sum \alpha_i y_i = 0, \alpha_i + \mu_i - C = 0$

Solving the above optimization problem we get the values of w and b as follows,

$$w = \sum_i \alpha_i y_i x_i \tag{4}$$

$$b = \frac{1}{N_s}\sum_n (t_n - \sum_m a_m y_m x_m) \tag{5}$$

b. i.
- C = 0.01
  Average train accuracy = 0.975, Standard deviation = 0.0051
- C = 0.1
  Average train accuracy = 0.985, Standard deviation = 0.0069
- C = 1
  Average train accuracy = 0.99, Standard deviation = 0.0073
- C = 10
  Average train accuracy = 0.998, Standard deviation = 0.0091
- C = 100
  Average train accuracy = 1.0, Standard deviation = 0.0085

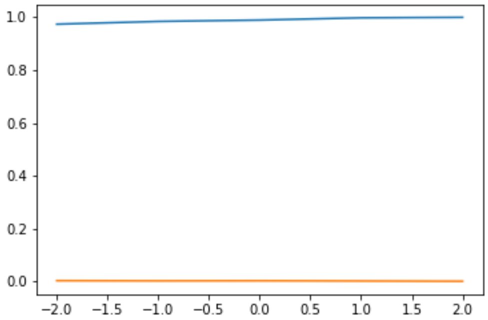


Figure 1: Average Train Accuracy And Standard Deviation

ii. - C = 0.01
 Average test accuracy = 0.970075, Standard deviation = 0.00669148
 Average Support Vectors = 1138, Standard deviation = 6.42262
- C = 0.1
 Average test accuracy = 0.98005, Standard deviation = 0.0064066
 Average Support Vectors = 340, Standard deviation = 6.21611
- C = 1
 Average test accuracy = 0.978304, Standard deviation = 0.00812294
 Average Support Vectors = 137, Standard deviation = 5.00899
- C = 10
 Average test accuracy = 0.982045, Standard deviation = 0.00746465
 Average Support Vectors = 96, Standard deviation = 5.53624
- C = 100
 Average test accuracy = 0.981796, Standard deviation = 0.00669612
 Average Support Vectors = 96, Standard deviation = 7.45989

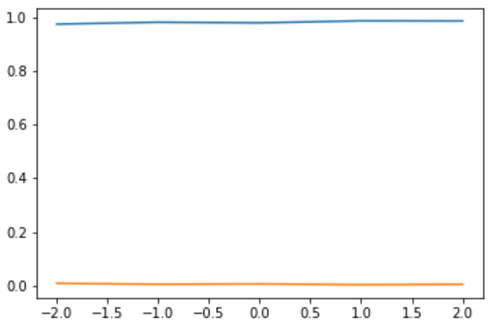


Figure 2: Average Test Accuracy and Standard Deviation for each log(C) value

iii. - C = 0.01, Average value of $\frac{1}{||w||} = 0.35378$
- C = 0.1, Average value of $\frac{1}{||w||} = 0.21748$
- C = 1, Average value of $\frac{1}{||w||} = 0.12268$

- C = 10, Average value of $\frac{1}{||w||} = 0.05431$
- C = 100, Average value of $\frac{1}{||w||} = 0.04066$

iv. - C = 0.01
  Average Support Vectors = 1138, Standard deviation = 6.42262
- C = 0.1
  Average Support Vectors = 340, Standard deviation = 6.21611
- C = 1
  Average Support Vectors = 137, Standard deviation = 5.00899
- C = 10
  Average Support Vectors = 96, Standard deviation = 5.53624
- C = 100
  Average Support Vectors = 96, Standard deviation = 7.45989

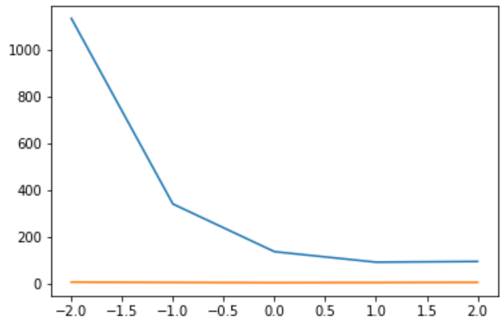


Figure 3: Average number of support vectors and Standard Deviation for each log(C) value

c. i. The parameter C controls the trade-off between the slack variable penalty and the margin. Because any point that is misclassified has ξ >1, it follows that $\sum_n \xi_n$ is an upper bound on the number of misclassified points. The parameter C is therefore analogous to (the inverse of) a regularization coefficient because it controls the trade-off between minimizing training errors and controlling model complexity. As C increases the margin decreases, which results in decrease in training error but leads to increase in test error due to overfitting.

ii. The main idea can be formulated as

$$\min \frac{1}{2}||w||^2 \quad \textit{such that} \quad y_i(w^T x_i + b) \geq 1, \forall i \tag{6}$$

For the non separable case, slack variables are introduced.

$$y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \tag{7}$$

$\sum_i \xi_i$ is the upper bound on the training error.

$$\min_{w,\{\xi_i\}} \frac{1}{2}||w||^2 + C\sum_i \xi_i \quad \textit{such that} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0,\ \forall i \tag{8}$$

Alternative version of equation (3) with N data points,

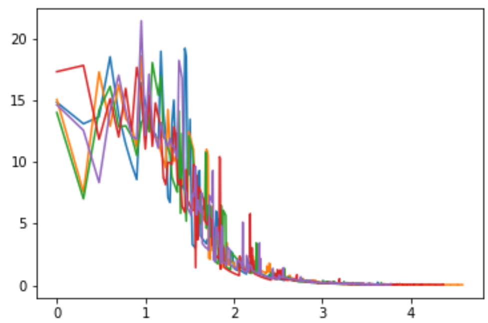
$$\min_{w} \frac{1}{N}\sum_{i=1}^{N} \lambda||w||^2 \ + \ max\{0, 1 - y_i f(x_i)\} \quad where\ f(x_i) = w^T x_i + b \tag{9}$$

In the above equation, first (regularization) term biases the solution towards zero in the absence of any data and the remaining terms give rise to the loss functions, one loss function per training point, encouraging correct classification. All the constraints of the equation(3) are also satisfied in equation(4)

Solution of Question 2

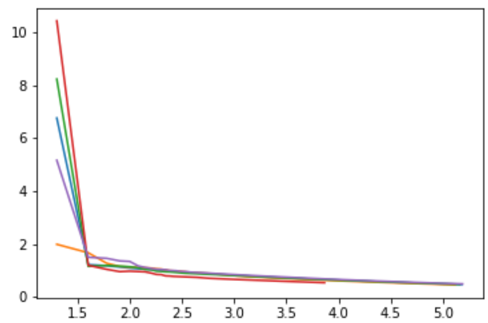
1. The Pegasos objective function values for the five runs for each K value are plotted with respect to $log(iteration_number)$ and the average runtime is mentioned below.

   - k=1,



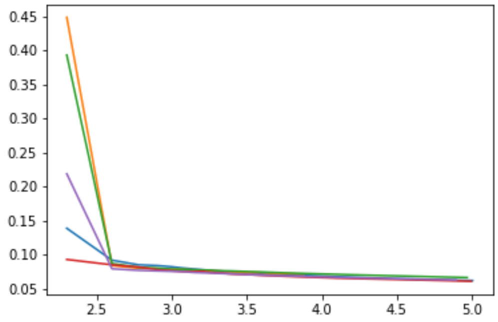
Average Runtime = 36.39650124279724, Standard Deviation = 9.565004753745608

   - k=20,



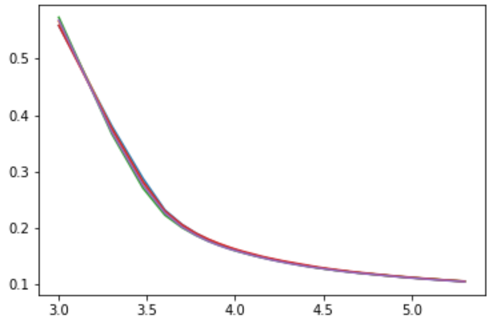
Average Runtime = 14.70385385400441, Standard Deviation = 6.9877448240205595

   - k=200,



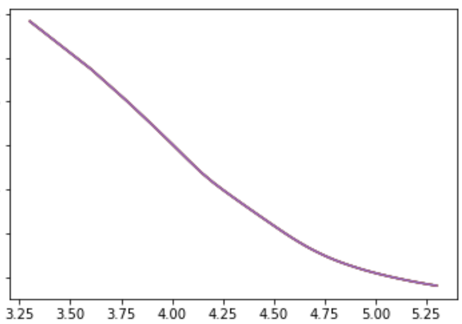
Average Runtime = 1.8656132497999351, Standard Deviation = 0.8689601290748988

- k=1000,



Average Runtime = 3.3322551333985757, Standard Deviation = 0.12662853439880215

- k=2000,



Average Runtime = 3.483599252998829, Standard Deviation = 0.07691296469641992

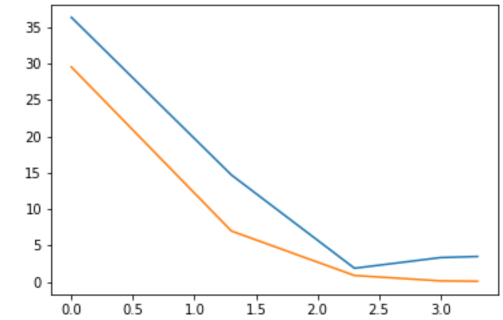


Figure 4: Average run-times and standard deviation with respect to log(k) values.

2. The primal objective function has the form,

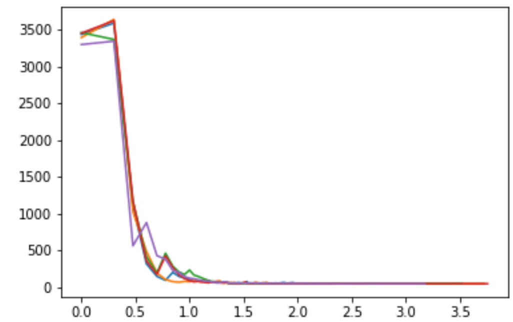
$$L(w) = \frac{1}{N}\sum_{i=1}^{N} \lambda||w||^2 \ + \ max\{0, 1 - y_i(w^T x_i)\} \tag{10}$$

$$\nabla L(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla[\lambda||w||^2 \ + \ alog(1 + e^{\frac{1 - y_i(w^T x_i)}{a}})] \tag{11}$$

$$= 2\lambda W - \frac{1}{N}\sum_{i=1}^{N} \frac{x_i y_i}{1 + e^{\frac{y_i(w^T x_i) - 1}{a}}}$$

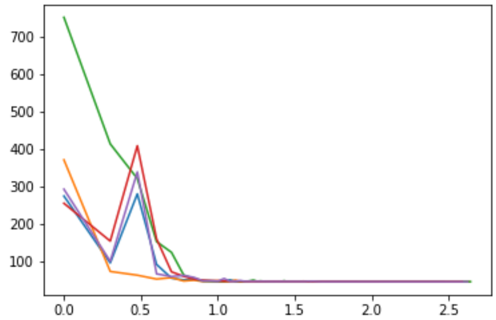
The Softplus objective function values for the five runs for each K value are plotted with respect to $log(iteration_number)$ and the average runtime is mentioned below.

- k=1,



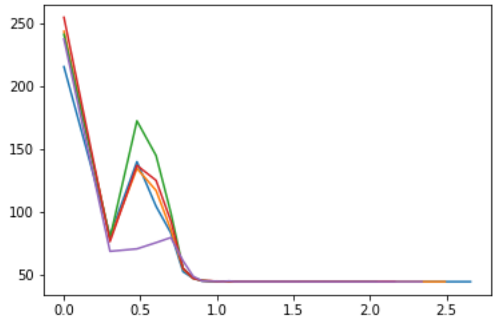
Average Runtime = 8.357152669600328, Standard Deviation = 5.144878815543728

- k=20,



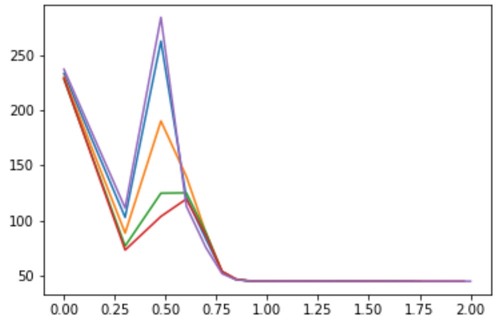
Average Runtime = 1.2854633577982895, Standard Deviation = 0.18256312217949544

- k=200,



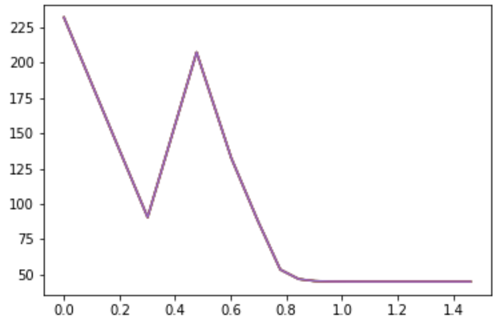
Average Runtime = 2.133322844599024, Standard Deviation = 1.1635392687934163

- k=1000,



Average Runtime = 2.2417441355995833, Standard Deviation = 0.6747164183768805

- k=2000,



Average Runtime = 1.8584549207997043, Standard Deviation = 0.14296474711850965

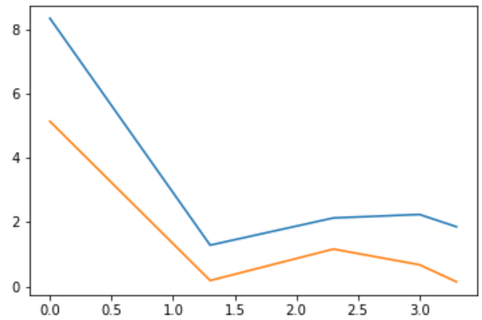


Figure 5: Average run-times and standard deviation with respect to log(k) values.