1a) $\theta^* \epsilon^* = \underset{\theta, \epsilon}{\arg\min}\ C \sum_{n=1}^{N} (\epsilon_n^+ + \epsilon_n^-) + \|w\|_2^2$

$\forall n \quad w x_n^T + b + \epsilon_n^+ - y_n \geq 0$

$y_n - w x_n^T - b + \epsilon_n^- \geq 0$

$\epsilon_n^+ \geq 0$

$\epsilon_n^- \geq 0$

Lagrangian is given by

$$\mathcal{L} = C \sum_{n=1}^{N} (\epsilon_n^+ + \epsilon_n^-) + \|w\|_2^2 - \sum_{n=1}^{N} \alpha_n (w x_n^T + b + \epsilon_n^+ - y_n) - \sum_{n=1}^{N} \beta_n (y_n - w x_n^T - b + \epsilon_n^-) - \sum_{n=1}^{N} \lambda_n \epsilon_n^+ - \sum_{n=1}^{N} \gamma_n \epsilon_n^-$$

such that $\alpha_n \geq 0$, $\beta_n \geq 0$, $\lambda_n \geq 0$, $\gamma_n \geq 0$

1b) ① $\frac{dL}{dw} \Rightarrow 2w - \sum_{n=1}^{N} \alpha_n x_n + \sum_{n=1}^{N} \beta_n x_n = 0$

$$\Rightarrow w = \frac{1}{2} \sum_{n=1}^{N} (\alpha_n - \beta_n) x_n$$

② $\frac{dL}{d\epsilon_n^+} \Rightarrow C - \alpha_n - \lambda_n = 0 \Rightarrow \alpha_n + \lambda_n = C$

③ $\frac{dL}{d\epsilon_n^-} \Rightarrow C - \beta_n - \gamma_n = 0 \Rightarrow \beta_n + \gamma_n = C$

④ $\frac{dL}{db} \Rightarrow -\sum_{n=1}^{N} \alpha_n + \sum_{n=1}^{N} \beta_n = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n = \sum_{n=1}^{N} \beta_n$

⑤ $\frac{dL}{d\alpha_n} \Rightarrow w x_n^T + b + \epsilon_n^+ - y_n = 0$

⑥ $\frac{dL}{d\beta_n} \Rightarrow \quad y_n - wx_n^T - b + \epsilon_n^- = 0$

⑦ $\frac{dL}{d\lambda_n} \Rightarrow \quad \epsilon_n^+ = 0$

⑧ $\frac{dL}{d\gamma_n} \Rightarrow \quad \epsilon_n^- = 0$

From ⑤ and ⑥ $\quad wx_n^T + b + \epsilon_n^+ - y_n = 0$

& ⑦ and ⑧ $\quad y_n - wx_n^T - b + \epsilon_n^- = 0$

$\Rightarrow \quad \epsilon_n^+ + \epsilon_n^- = 0$

$\Rightarrow$ Lagrangian dual

$$= \frac{1}{4} \sum_{n=1}^{N} \sum_{m=1}^{N} (\alpha_n - \beta_n)(\alpha_m - \beta_m) x_n x_m^T$$

such that $\quad 0 \le \alpha_n \le C \quad (\because \alpha_n + \lambda_n = C \;\&\; \alpha_n, \lambda_n \ge 0)$

$0 \le \beta_n \le C \quad (\because \beta_n + \gamma_n = C \;\&\; \beta_n, \gamma_n \ge 0)$

$$\sum_{n=1}^{N} \alpha_n = \sum_{n=1}^{N} \beta_n$$

$$\epsilon_n^+, \epsilon_n^- = 0$$

Maximizing the dual is equivalent to minimizing the original problem.

2a) $$w^* = \arg\min_w C \sum_{n=1}^{N} \max(0, 1 - y_n(wx_n^T + b)) + \|w\|_2^2$$

when $\quad 1 - y_n(wx_n^T + b) > 0$

$$w^* = \arg\min_w C \sum_{n=1}^{N} (1 - y_n(wx_n^T + b)) + \|w\|_2^2$$

Subgradient with respect to w

$$= -C \sum_{n=1}^{N} y_n x_n + 2w$$

Subgradient with respect to b

$$= 0$$

When $1 - y_n(w x_n^T + b) \leq 0$

$$w^* = \arg\min_w \|w\|_2^2$$

Subgradient with respect to $w = 2w$

Subgradient with respect to $b = 0$

1c)
Solving the constrained version of the primal problem may be computationally infeasible in some situations. For example, if we want to capture nonlinear trends in a linear model we typically introduce basis expansion. Mapping of existing features into the new space and computing the solution is prohibitively expensive for a space with a large number of dimensions. In such cases, solving the dual form of the problem is more efficient as the dual only depends on the sum of dot products of the data points which can be computed using kernels. Kernels operate in a high dimensional space without ever computing the coordinates of the data in that space but rather by computing the inner products between the images of all pairs of data in the feature space.

2b)
Optimal parameters for this model can be derived by minimizing the hinge loss of the SVM classifier under the RRM framework. As the data samples were small and the model has a global minima, simple batch gradient descent was used to converge to the optimal values. Step size alpha was initially set to 0.01 based on empirical guidelines. The model was then trained on a subset of the training data and prediction accuracy was evaluated on the other set (validation set) for different values of the step size. The step size was tuned to 0.001 to minimize classification error and the model was retrained on the entire training set to identify optimal weights and bias.

2c)
The value of the SVC objective function at the optimal model parameters: 48.389660
Classification error rate on the training data: 0.014000

2d)
Value of logistic regression objective function at Wlr and Blr: 60.772521
Value of SVC objective function at Wlr and Blr: 63.814570
Classification error rate: 0.001000

2e)
SVC and logistic regression fare similarly on the test data.
Classification error rate for SVC on test data: 0.028939
Classification error rate for logistic regression on test data: 0.020900
In the case of linearly separable data, SVC could provide lower generalization error on test data drawn from the same distribution as the training data than a logistic regression model due to the maximum margin properties of SVC.

3a)
The multi-output custom neural networks model was developed using Tensorflow which facilitates automatic differentiation. Parameter initialization was done using Xavier initialization and Adam optimizer was used on mini batches of fifty samples over ten epochs to converge to the minima. Adam uses per parameter adaptive learning rates based on both first and second order moments. Adam makes use of three parameters: alpha the step size and beta1 and beta2 which control the decay rates. Step size was tuned to 0.001 using cross validation and beta1 and beta2 were set to 0.9 and 0.999.

3b)
Alpha: 0.5
Classification error rate on training data: 0.0414528227398
MSE of location predictions on training data: 3.5063
Classification error rate on test data: 0.0425531914894
MSE of location predictions on test data: 7.72012

3c)
Alpha represents the tradeoff between classification error and localization error. It defines the degree to which we attempt to minimize each of these losses.
As alpha increases more weightage is given to minimizing the classification error as opposed to the localization loss. Therefore would see better classification performance and deteriorating localization performance.
Similarly, as alpha decreases we attempt to minimize the localization error to a greater degree and therefore observe better localization performance and poor classification.
The following figures represent the classification error and localization errors at different alpha values.