Assignment -3

Problem Set 3: Linear Discriminant Functions and Support Vector Machines

In this problem set, we will ask you to train a support vector classifier with using the the MNIST training data set and report performance using MNIST testing data set. Then we will ask you to derive the primal-dual relationship of the 1-norm soft-margin classification problem, from which process we hope to help you understand several key concepts, such as maximal margin and support vector. Our emphasis is on getting hands dirty with SVM and understanding the theory. There are many tutorials already on this problem, and you can refer the solution to problem 30 in Chapter 5 for a similar mathematical derivation involving the method of Lagrange multipliers and KKT condition.

1. (2 points) Write code to train a multi-class support vector classifier with dot-product kernel and 1-norm soft margin using the MNIST training data set. Then reporting the performance using MNIST test data set. There is a hyper-parameter that sets the trade-off between the margin and the training error --- tune this hyper-parameter through cross-validation.

2. (8 points) Identify the Lagrange dual problem of the following primal problem:

Given features , where ,

Minimize , the weighted sum between the squared length of the separating vector and the errors, where w is the separating vector, is the dot product, and is the error made by separating vector w on feature .

Subject to and for . In other words, if the "normalized feature" has a margin less than 1, , we add a slackness term to make it 1.

Point out what is the "margin" in both the primal formulation and the dual formulation, what are the benefits of maximizing the margin. Characterize the support vectors. Point out the benefit of solving the dual problem instead of the primal problem.

2. (8 points) Identify the Lagrange dual problem of the following primal problem:
Given features $(x_1,y_1),\cdots,(x_N,y_N)$, where $y_1,\cdots,y_N \in \{-1,1\}$,

Minimize $w^T \cdot w + C\sum_{i=1}^{N} \xi_i$, the weighted sum between the squared length of the separating vector and the errors, where

w is the separating vector, $w^T \cdot w$ is the dot product, and ξ_i is the error made by separating vector w on feature (x_i,y_i).

Subject to $y_i \cdot (w^T \cdot x_i) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for $i = 1,\cdots,N$. In other words, if the "normalized feature" $y_i x_i$ has a margin less than 1, $w^T \cdot (y_i x_i) \leq 1$, we add a slackness term to make it 1.

Point out what is the "margin" in both the primal formulation and the dual formulation, what are the benefits of maximizing the margin. Characterize the support vectors. Point out the benefit of solving the dual problem instead of the primal problem.

3. (Optional) Formulate the primal problem and derive the dual problem if there are multiple classes.