Problem Set 3

1 Problem 1

1.1 Load dataset

```
mnist = input_data.read_data_sets("MNIST_data/")
X = mnist.test.images
Y = mnist.test.labels
```

Because the size of mnist.train is too large, the it will spend too much time to train the model. I choose the mnist.test whose size is smaller as the total set.

1.2 Split the dataset

```
train_image, test_image, train_label, test_label = train_test_split(X, Y, test_size=0.1, random_state=0)
```

1.3 Train and test

```
svc = svm.SVC(kernel='linear', C=2)
svc.fit(train_image, train_label)
scores = cross_validation.cross_val_score(svc, train_image, train_label, cv=5)
print(scores)
print('Accuracy: %.3f (+/- %.3f)' % (np.mean(scores), np.std(scores) * 2))
print('Test Accuracy: %.3f'%(svc.score(test_image,test_label)))
```

1.4 Result

```
[ 0.92627494  0.92068774  0.92825362  0.9243604   0.90929327]
Accuracy: 0.922 (+/- 0.013)
Test Accuracy: 0.936
```

2 Problem 2

Identify the Lagrange dual problem of the following primal problem

Given feature $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$, where $y_1, \dots y_N \in \{-1,1\}$,

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

Primal formulation:

$$\min \frac{1}{2} \| w \|^2 + \mathrm{C} \sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } , y_i(w^T x_i) \geq 1 - \xi_i \text{ , i} = 1, \dots, \text{n}$$

$$\xi_i \geq 0, \text{i} = 1, \dots, \text{n}$$

$$\mathcal{L}(w, \xi, \alpha, r) = \frac{1}{2} \| w \|^2 + + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} r_i \xi_i$$

Set partial derivative of Lagrange function

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \implies w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \implies C - \alpha_i - r_i = 0, i = 1, \dots, n$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \implies \sum_{i=1}^{N} \alpha_i y_i = 0$$

Substituting $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ into Lagrange function, we get dual problem of maxizing:

$$\mathcal{L} = \frac{1}{2} \| w \|^2 + + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} r_i \xi_i$$

$$= \frac{1}{2} w^T w + + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i y_i w^T x_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

$$= \frac{1}{2} w^T \sum_{i=1}^{N} \alpha_i y_i x_i + \mathrm{C} \sum_{i=1}^{N} \xi_i - w^T \sum_{i=1}^{N} \alpha_i y_i x_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

$$= -\frac{1}{2} w^T \sum_{i=1}^{N} \alpha_i y_i x_i + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

$$= -\frac{1}{2} (\sum_{i=1}^{N} \alpha_i y_i x_i)^T \sum_{i=1}^{N} \alpha_i y_i x_i + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \alpha_i y_i (x_i)^T \sum_{i=1}^{N} \alpha_i y_i x_i + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

$$= -\frac{1}{2} \sum_{i=1,j=1}^{N} \alpha_i y_i (x_i)^T \alpha_j y_j x_j + \mathrm{C} \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i y_i \, b + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} r_i \xi_i$$

b = 0

$$\mathcal{L} = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{N} \alpha_i y_i (x_i)^T \alpha_j y_j x_j$$

C-$\alpha_i - r_i = 0 \; and \; r_i \geq 0 \; and \; \alpha_i \geq 0 \quad \Rightarrow \quad 0 \leq \alpha_i \leq C$

So the dual problem is:

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{N} \alpha_i \alpha_j y_i y_j < x_i x_j >$$

$$\text{s.t.}, 0 \leq \alpha_i \leq C, i = 1, \dots, n$$

$$\sum_{i=1}^{N} \alpha_i \, y_i = 0$$

2.1 Point out what is the "margin" in both the primal formulation and the dual formulation

Margin in primal formulation:

$$\text{Margin} = \frac{1}{\sqrt{w^T w}}$$

Margin in dual formulation

$$\text{Margin} = \frac{1}{\sqrt{\sum_{i=1,j=1}^{N} \alpha_i y_i (x_i)^T \alpha_j y_j x_j}}$$

2.2 what are the benefits of maximizing the margin

One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized.

2.3 Characterize the support vectors

$$y_i(w^T x_i + b) = 1$$

Only the vectors x_i which can satisfy $y_i(w^T x_i + b) = 1$ are support vectors. It means the points on the margin are the support vectors.

2.4 Point out the benefit of solving the dual problem instead of the primal

problem

$$\theta(w) = \max_{a_i \geq 0} \mathcal{L}(w, \xi, \alpha, r)$$

$$\min_{w,b} \theta(w) = \min_{w,b} \max_{a_i \geq 0} \mathcal{L}(w, \xi, \alpha, r) = p^*$$

exchange max and min, we can get the dual problem

$$\max_{a_i \geq 0} \min_{w,b} \mathcal{L}(w, \xi, \alpha, r) = d^*$$

$$d^* \leq p^* \quad (\text{weak duality})$$

The benefit:

The primal problem is to minimize the $\max_{a_i \geq 0} \mathcal{L}(w, \xi, \alpha, r)$. After being converted to a dual problem, the problem is to maximize the $\min_{w,b} \mathcal{L}(w, \xi, \alpha, r)$. There are less variables in dual problem and the dual problem is easier to solve than the primal problem.