HOMEWORK 3 - V1

STA414/2104 Winter 2021

University of Toronto

Version history: V0 → V1: starter code 2 typos fixed: import utils, train return value

- **Deadline:** March. 22, at 13:59.
- **Submission:** You need to submit your solutions through Crowdmark, including all your derivations, plots, and your code. You can produce the file however you like (e.g. LaTeX, Microsoft Word, etc), as long as it is readable. Points will be deducted if we have a hard time reading your solutions or understanding the structure of your code. For this assignment, you should submit the filled out starter code right after your answers.

1. Support Vector Machines Dual Problem - (40 pts). Assume that you are given a data set $\mathcal{D} = \{(t_i, \mathbf{x}_i) \ : \text{for } i = 1, ..., N\}$ with $t_i \in \{\pm 1\}$.

1.1. *Hard margin - (20 pts).* Recall that the hard margin SVM problem can be written in the following primal form

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 \\ &\text{s.t. } t_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, \ldots, N \end{aligned}$$

(a) Write down the Lagrangian for this problem with Lagrangian parameters denoted with α_i's.

(b) Show that the equivalent dual problem can be written as

$$\begin{aligned} &\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\ &\text{s.t. } 0 \leq \alpha_i \qquad i = 1, 2, ..., N. \\ &\qquad \sum_{i=1}^{N} \alpha_i t_i = 0. \end{aligned}$$

(c) Assume that we solved the above dual formulation and obtained the optimal α. For a given test data point $\mathbf{x}$, how can we predict its class?

1.2. *Soft margin - (20 pts).* Recall that the soft margin SVM problem can be written in the following primal form

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|_2^2 + \gamma \sum_{i=1}^{N} \xi_i \\ &\text{s.t. } t_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \qquad i = 1, \ldots, N \\ &\qquad \xi_i \geq 0 \qquad\qquad\qquad\qquad i = 1, \ldots, N \end{aligned}$$

(a) Use the Lagrangian provided in the lecture to show that the equivalent dual problem can be written as

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j$$
$$\text{s.t. } 0 \leq \alpha_i \leq \gamma \qquad i = 1, \ldots, N$$
$$\sum_{i=1}^{N} \alpha_i t_i = 0.$$

(b) Assume that we solved the above dual formulation and obtained the optimal α. For a given test data point $\mathbf{x}$, how can we predict its class?

What to submit?

1.1-a) Lagrangian.
b) Your derivation of equivalent optimization problem.
c) The prediction rule for a new test point.
1.2-a) Your derivation of equivalent optimization problem..
b) The prediction rule for a new test point.

2. Neural Networks (60 points). In this problem, you will experiment on a subset of the Toronto Faces Dataset (TFD). You will complete the starter code provided to you, and experiment with the completed code. You should understand the code instead of using it as a black box.

We subsample 3374, 419 and 385 grayscale images from TFD as the training, validation and testing set respectively. Each image is of size 48×48 and contains a face that has been extracted from a variety of sources. The faces have been rotated, scaled and aligned to make the task easier. The faces have been labeled by experts and research assistants based on their expression. These expressions fall into one of seven categories: 1-Anger, 2-Disgust, 3-Fear, 4-Happy, 5-Sad, 6-Surprise, 7-Neutral. We show one example face per class in Figure 1.

Fig 1: Example faces. From left to right, the the corresponding class is from 1 to 7.

Code for training a neural network (fully connected) is partially provided in `nn.py`.

2.1. *Complete the code [20 points].* Follow the instructions in nn.py to implement the missing functions that perform the backward pass of the network.

2.2. *Generalization [10 points].* Train the neural network with the default set of hyperparameters. Report training, and validation errors and a plot of error curves (training and validation). Examine the statistics and plots of training error and validation error (generalization). How does the network's performance differ on the training set vs. the validation set during learning?

2.3. *Optimization [10 points].* Try different values of the learning rate (step size) η ("eta") ranging from $\eta \in \{0.001, 0.01, 0.5\}$. What happens to the convergence properties of the algorithm (looking at both cross-entropy and percent-correct)? Try 3 different mini-batch sizes ranging from $\{10, 100, 1000\}$. How does mini-batch size affect convergence? How would you choose the best value of these parameters? In each of these hold the other parameters constant while you vary the one you are studying.

2.4. *Model architecture [10 points].* Try 3 different values of the number of hidden units for each layer of the fully connected network (range from $\{2, 20, 80\}$). You might need to adjust the learning rate and the number of epochs (iterations). Comment on the effect of this modification on the convergence properties, and the generalization of the network.

2.5. *Network Uncertainty [10 points].* Plot five examples where the neural network is not confident of the classification output (the top score is below some threshold), and comment on them. Will the classifier be correct if it outputs the top scoring class anyways?

What to submit?

2.1) Completed code.
2.2) Final training and validation errors, and a plot of these errors across all iterations. Your comments on network's performance.
2.3) The curves you obtained in the previous part for the given step size, mini-batch size choices. Your comments/answers to the two questions.
2.4) The curves you obtained in the previous part for the given number of hidden units. Your comments on the convergence/generalization.
2.5) Five example images and your comments, and answer to the question. Your comments on the convergence/generalization.

1. Support Vector Machines Dual Problem - (40 pts). Assume that you are given a data set $\mathcal{D} = \{(t_i, \mathbf{x}_i) \ : \text{for } i = 1, ..., N\}$ with $t_i \in \{\pm 1\}$.

1.1. *Hard margin - (20 pts).* Recall that the hard margin SVM problem can be written in the following primal form

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t. } t_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \qquad i = 1, \ldots, N$$

(a) Write down the Lagrangian for this problem with Lagrangian parameters denoted with α_i's.
(b) Show that the equivalent dual problem can be written as

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j$$
$$\text{s.t. } 0 \leq \alpha_i \qquad i = 1, 2, ..., N.$$
$$\sum_{i=1}^{N} \alpha_i t_i = 0.$$

(c) Assume that we solved the above dual formulation and obtained the optimal α. For a given test data point $\mathbf{x}$, how can we predict its class?

(a).

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{N} \alpha_i \left(t_i (w^T x_i + b) - 1 \right)$$

$$= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \alpha_i \left(1 - t_i (w^T x_i + b) \right)$$

$\therefore$ the lagrangian is $\frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \alpha_i \left(1 - t_i (w^T x_i + b) \right)$

(b).

$$\frac{\partial L(w, b, \alpha)}{\partial w} = \frac{1}{2} \cdot 2w + \sum_{i=1}^{N} \alpha_i \left(0 - (t_i)(x_i) - 0\right) = 0$$

$$w - \sum_{i=1}^{N} \alpha_i t_i x_i = 0$$

$$\boxed{w^* = \sum_{i=1}^{N} \alpha_i t_i x_i}$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0 + \sum_{i=1}^{N} \alpha_i (0 - 0 - t_i) = 0$$

$$-\sum_{i=1}^{N} \alpha_i t_i = 0$$

$$\boxed{\sum_{i=1}^{N} \alpha_i t_i = 0}$$

$$W(\alpha) = L(w^*, b, \alpha)$$

$$= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{N} \alpha_i \left(1 - t_i(w^T x_i + b)\right)$$

$$= \frac{1}{2}\left\|\sum_{i=1}^{N} \alpha_i t_i x_i\right\|_2^2 + \sum_{i=1}^{N} \alpha_i \left[1 - t_i\left(\left(\sum_{j=1}^{N} \alpha_j t_j x_j\right)^T x_i + b\right)\right]$$

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i t_i x_i\right)^T \left(\sum_{j=1}^{N} \alpha_j t_j x_j\right) + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i t_i \left(\sum_{j=1}^{N} \alpha_j t_j x_j\right)^T x_i - \sum_{i=1}^{N} \alpha_i t_i b$$

$$b \sum_{i=1}^{N} \alpha_i t_i = b \cdot 0 = 0$$

$$= \boxed{\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j x_i^T x_j}$$

$$\therefore \quad W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j x_i^T x_j$$

$$s.t \quad \alpha_i \geq 0 \quad , \quad i = 1, \dots, N$$

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

if $\alpha_i < 0$, $\min_{w,b} L(w, b, \alpha)$ will always be smaller than the previous iteration. we will not have a solution.

This is a dual problem,

$\min_{w,b} [\max_{\alpha} L(w, b, \alpha)]$ is equivalent to $\max_{\alpha} [\min_{w,b} L(w, b, \alpha)]$

$\therefore$ the equivalent dual problem can be written as:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j x_i^T x_j$$

$$s.t \quad \alpha_i \geq 0 \quad , \quad i = 1, \dots, N$$

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

(c).

For a given test point x, and optimal λ.

$$\hat{y} = w^{*T} x + b^*$$

$$= \left(\sum_{i=1}^{N} \lambda_i t_i x_i \right)^T x + b^*$$

if $\hat{y} > 0 \Rightarrow$ we will in the class of $t = +1$

if $\hat{y} < 0 \Rightarrow$ we will in the class of $t = -1$

1.2. *Soft margin - (20 pts).* Recall that the soft margin SVM problem can be written in the following primal form

$$\begin{aligned}
&\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|_2^2 + \gamma \sum_{i=1}^{N} \xi_i \\
&\text{s.t. } t_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \qquad i = 1, \ldots, N \\
&\quad \xi_i \geq 0 \qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, N
\end{aligned}$$

(a) Use the Lagrangian provided in the lecture to show that the equivalent dual problem can be written as

$$\begin{aligned}
&\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j \\
&\text{s.t. } 0 \leq \alpha_i \leq \gamma \qquad i = 1, \ldots, N \\
&\quad \sum_{i=1}^{N} \alpha_i t_i = 0.
\end{aligned}$$

(b) Assume that we solved the above dual formulation and obtained the optimal α. For a given test data point $\mathbf{x}$, how can we predict its class?

What to submit?

1.1-a) Lagrangian.
b) Your derivation of equivalent optimization problem.
c) The prediction rule for a new test point.
1.2-a) Your derivation of equivalent optimization problem..
b) The prediction rule for a new test point.

(a)

$$L(w, b, \varepsilon, \alpha, \gamma) = \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \alpha_i [t_i (w^T x_i + b) - 1 + \varepsilon_i] - \sum_{i=1}^{N} \beta_i \varepsilon_i$$

$$\frac{\partial L(w, b, \varepsilon, \alpha, \gamma)}{\partial w} = 2 \cdot \frac{1}{2} w - \sum_{i=1}^{N} \alpha_i t_i x_i = 0$$

$$\boxed{w^* = \sum_{i=1}^{N} \alpha_i t_i x_i}$$

$$\frac{\partial L(w, b, \varepsilon, \alpha, \gamma)}{\partial b} = - \sum_{i=1}^{N} \alpha_i t_i = 0$$

$$\boxed{\sum_{i=1}^{N} \alpha_i t_i = 0}$$

$$\frac{\partial L(w, b, \varepsilon, \alpha, \gamma)}{\partial \varepsilon} = \gamma - \alpha_i - \beta_i$$

$$\boxed{\gamma = \alpha_i + \beta_i}$$

$$w(\alpha) = L(w, b, \varepsilon, \alpha, \gamma)$$

$$= \frac{1}{2}\|w\|^2 + \gamma \sum_{i=1}^{N} \varepsilon_i - \sum_{i=1}^{N} \alpha_i [t_i (w^T x_i + b) - 1 + \varepsilon_i] - \sum_{i=1}^{N} \beta_i \varepsilon_i$$

$$= \frac{1}{2}\left\| \sum_{i=1}^{N} \alpha_i t_i x_i \right\|_2^2 + \cancel{\sum_{i=1}^{N} (\alpha_i + \beta_i) \varepsilon_i} - \sum_{i=1}^{N} \alpha_i \left[t_i \left(\left(\sum_{j=1}^{N} \alpha_j t_j x_j \right)^T x_i + b \right) \right]$$

$$+ \sum_{i=1}^{N} \alpha_i - \cancel{\sum_{i=1}^{N} \alpha_i \varepsilon_i} - \cancel{\sum_{i=1}^{N} \beta_i \varepsilon_i}$$

$$= \sum_{i=1}^{N} \alpha_i + \frac{1}{2} \left(\sum_{i=1}^{N} \alpha_i t_i x_i \right)^T \left(\sum_{j=1}^{N} \alpha_j t_j x_j \right) - \sum_{i=1}^{N} \alpha_i t_i \left(\sum_{j=1}^{N} \alpha_j t_j x_j \right)^T x_i - \underbrace{\sum_{i=1}^{N} \alpha_i t_i b}_{b \sum_{i=1}^{N} \alpha_i t_i = b \cdot 0 = 0}$$

$$= \boxed{\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j x_i^T x_j}$$

$$\text{s.t} \quad \alpha_i \geq 0 \quad , \quad i = 1, \dots, N$$

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

$$\boxed{\alpha_i \geq 0, \quad \beta_i \geq 0, \quad \alpha_i + \beta_i = \gamma}$$

$$\Rightarrow \quad \alpha_i \leq \gamma$$

$$\Rightarrow \quad 0 \leq \alpha_i \leq \gamma$$

This is a dual problem,

$\min_{w,b} [\max_{\alpha} L(w, b, \xi, \alpha, \gamma)]$ is equivalent to $\max_{\alpha} [\min_{w,b} L(w, b, \xi, \alpha, \gamma)]$

$\therefore$ the equivalent dual problem can be written as:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} t_i t_j \alpha_i \alpha_j x_i^T x_j$$

$$s.t \quad 0 \le \alpha_i \le \gamma \quad , \quad i = 1, \dots, N$$

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

(b).

For a given test point x, and optimal α.

$$\hat{y} = w^{*T} x + b^*$$

$$= \left(\sum_{i=1}^{N} \alpha_i t_i x_i \right)^T x + b^*$$

if $\hat{y} > 0 \Rightarrow$ we will in the class of $t = +1$

if $\hat{y} < 0 \Rightarrow$ we will in the class of $t = -1$