清華大學

Tsinghua University

Department of Computer Science and Technology

Machine Learning

Homework 1

Sahand Sabour

2020280401

1 Mathematics Basics

1.1 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$\begin{aligned} \min_{x_1,x_2} \quad & x_1^2 + x_2^2 - 1 \\ s.t. \quad & x_1 + x_2 - 1 = 0 \\ & x_1 - 2x_2 \geq 0 \end{aligned}$$

Solution:

The Lagrangian equation can be written as $L(x_1, x_2, \lambda_1, \lambda_2)$, where λ_1 and λ_2 are the Lagrange multipliers. Therefore, Lagrangian function would be written as:

$$L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + x_2^2 - 1 - \lambda_1(x_1 + x_2 - 1) - \lambda_2(x_1 - 2x_2)$$

Hence, we can have that:

$$\begin{aligned} \frac{\partial L}{\partial x_1} = 0 \quad => \quad 2x_1 - \lambda_1 - \lambda_2 = 0 \quad => \quad x_1^* = \frac{\lambda_1 + \lambda_2}{2} \\ \frac{\partial L}{\partial x_2} = 0 \quad => \quad 2x_2 + \lambda_2 + 2\lambda_2 = 0 \quad => \quad x_2^* = \frac{\lambda_1 - 2\lambda_2}{2} \end{aligned}$$

Accordingly, using the above equations, we can expand and rewrite the Lagrangian equation as follows:

$$\begin{aligned} L(\lambda_1, \lambda_2) = & \frac{1}{4}\lambda_1^2 + \frac{1}{2}\lambda_1\lambda_2 + \frac{1}{4}\lambda_2^2 + \frac{1}{4}\lambda_1^2 - \lambda_1\lambda_2 + \lambda_2^2 \\ & -1 - \lambda_1^2 + \frac{1}{2}\lambda_1\lambda_2 + \lambda_1 + \frac{1}{2}\lambda_1\lambda_2 - \frac{5}{2}\lambda_2^2 \\ = & -\frac{1}{2}\lambda_1^2 - \frac{5}{4}\lambda_2^2 + \frac{1}{2}\lambda_1\lambda_2 + \lambda_1 - 1 \end{aligned}$$

Therefore, we can have that:

$$\begin{aligned} \frac{\partial L}{\partial \lambda_1} = 0 \quad => -\lambda_1 + \frac{1}{2}\lambda_2 + 1 = 0 \\ \frac{\partial L}{\partial \lambda_2} = 0 \quad => -\frac{5}{2}\lambda_2 + \frac{1}{2}\lambda_1 = 0 \end{aligned}$$

Solving the above two equations gives $\lambda_1 = \frac{10}{9}$ and $\lambda_2 = \frac{2}{9}$. Accordingly, we use these values to obtain $x_1 = \frac{2}{3}$ and $x_2 = \frac{1}{3}$ based on the initial two derivations. Since (1) both Lagrange multipliers satisfy $\lambda \geq 0$; (2) $g(x) = x_1 + x_2 - 1 = \frac{2}{3} + \frac{1}{3} - 1 = 0$ and $h(x) = x_1 - 2x_2 = \frac{2}{3} - 2(\frac{1}{3}) = 0 \geq 0$; (3) $\lambda_2 h(x) = \frac{2}{9}(\frac{2}{3} - 2(\frac{1}{3})) = 0$; KKT conditions are satisfied and therefore, the solutions $x_1 = \frac{2}{3}$ and $x_2 = \frac{1}{3}$ are valid.

1.2 Stochastic Process

We toss a fair coin for a number of times and use H(head) and T(tail) to denote the two sides of the coin. Please compute the expected number of tosses we need to observe a first time occurrence of the following consecutive pattern

H, T, T, ..., T

Solution:

Let x denote the number of tosses we need to get n consecutive turns of the same side (i.e n heads or n tails). For the first toss, if we get the side we want immediately, then the probability would be $\frac{1}{2}$. Otherwise, this turn is useless and the number of turns would be x+1. In the second toss, the probability of getting the order we want is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ and the case that we don't get what we want, the number of turns would be x+2. This gives us the following sequence for the expected number of tosses before observing our desired pattern for a coin side:

$$x = \frac{1}{2}(x+1) + \frac{1}{4}(x+2) + \frac{1}{8}(x+3) + ...$$

Accordingly, solving the above equation gives $x_{S(n)} = 2(2^n - 1)$, where S is the side we want to observe n times. Hence, if we think of this problem as number of tosses to see one consecutive head and k consecutive tails, we would need $x_{H(1)} + x_{T(k)} = 2(2^1 - 1) + 2(2^k - 1) = 2 + 2^{k+1} - 2 = 2^{k+1}$ tosses to observe this pattern.

2 SVM

Consider the regression problem with training data $\{(x_i, y_i)\}_{i=1}^N (x_i \in R^d, y_i \in R)$. $\epsilon < 0$ denotes a fixed small value. Derive the dual problem of the following primal problem of linear SVM:

$$\begin{aligned}
& \min_{w,b,\xi,\hat{\xi}} \quad \frac{1}{2}||w||^2 \quad + \quad C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\
s.t. \quad & y_i \leq w^T x_i + b + \epsilon + \xi_i, i = 1, ..., N \\
& y_i \geq w^T x_i + b - \epsilon - \xi_i, i = 1, ..., N \\
& \xi_i \geq 0 \quad \forall i = 1, ..., N \\
& \hat{\xi}_i \geq 0 \quad \forall i = 1, ..., N
\end{aligned}$$

Solution:

Let $a, c, d,$ and $e \geq 0$ be the Lagrange multipliers. Then, the Lagrangian function would be

$$L(w, b, \xi, \hat{\xi}, a, c, d, e) = \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \hat{\xi}_i) - \sum_i a_i(w^T x_i + b + \epsilon + \xi_i - y_i) \\ - \sum_i c_i(y_i - w^T x_i - b + \epsilon + \hat{\xi}_i) - \sum_i d_i\xi_i - \sum_i e_i\hat{\xi}_i$$

Therefore, we can have that

$$\frac{\partial L}{\partial w} = \hat{w} - \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N} c_i x_i = 0 \quad giving \quad \hat{w} = \sum_{i=1}^{N}(a_i - c_i)x_i \tag{1}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N} c_i x_i = 0 \quad giving \quad \sum_i a_i - c_i = 0 \tag{2}$$

$$\frac{\partial L}{\partial \xi} = C1 - \sum_i a_i - \sum_i d_i = 0 \quad giving \quad C = a + d \tag{3}$$

$$\frac{\partial L}{\partial \hat{\xi}} = C1 + \sum_i c_i - \sum_i e_i = 0 \quad giving \quad C = c + e \tag{4}$$

Therefore, we initially expand the Lagrangian function as below

$$\begin{aligned} L(w, b, \xi, \hat{\xi}, a, c, d, e) &= \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \hat{\xi}_i) - \sum_i a_i w^T x_i - \sum_i a_i b - \sum_i a_i\epsilon \\ &\quad - \sum_i a_i\xi_i + \sum_i a_i y_i - \sum_i c_i y_i + \sum_i c_i w^T x_i + \sum_i c_i b \\ &\quad - \sum_i c_i\epsilon - \sum_i c_i\hat{\xi}_i - \sum_i d_i\xi_i - \sum_i e_i\hat{\xi}_i \\ &= \frac{1}{2}||w||^2 + C\sum_i(\xi_i + \hat{\xi}_i) - \sum_i(a_i - c_i)w^T x_i \\ &\quad - \sum_i(a_i - c_i)b - \sum_i(a_i + c_i)\epsilon - \sum_i(a_i + d_i)\xi_i \\ &\quad + \sum_i(a_i - c_i)y_i - \sum_i(c_i + e_i)\hat{\xi}_i \end{aligned}$$

Then, using equations 5-8 ,we can write the Lagrangian function as

$$\begin{aligned} L(w,b,\xi,\hat{\xi},a,c,d,e) &= \frac{1}{2}||w||^2 - \hat{w}w^T - b(0) - \epsilon\sum_i(a_i+c_i) + \sum_i(a_i-c_i)y_i \\ &\quad + C\sum_i(\xi_i+\hat{\xi}_i) - \sum_i(a_i+d_i)(\xi_i+\hat{\xi}_i) \\ &= -\frac{1}{2}||w||^2 - \epsilon\sum_i(a_i+c_i) + \sum_i(a_i-c_i)y_i \end{aligned}$$

Hence, the dual optimization problem would be

$$\begin{aligned} \underset{a,c}{argmax} -\frac{1}{2}\sum_i\sum_j(a_i-c_i)(a_j-c_j)x_i^Tx_j - \epsilon\sum_i(a_i+c_i) + \sum_i(a_i-c_i)y_i \\ s.t \qquad \sum_i(a_i-c_i) = 0 \\ 0 \le a_i, c_i \le C \end{aligned}$$

3 Deep Neural Networks

To make neural networks work well in practice is not easy in general, since there are too many hyper-parameters to tune such as the choice of the number of hidden layers, the activation function, the learning rate and so on. Besides some general guidelines (some standard techniques which are useful at most cases such as dropout, data augmentation), experience is of great importance.
Though a beginner may often be confused with them, luckily, there are some software available on the internet to help you build up a good sense on tuning neural networks. In this problem, you need to train the neural networks with different choices of hyper-parameters from the following link - A Neural Network Playground (you may need a VPN) - and answer the following questions:

1. Identify the best configuration you find for different problems and datasets. Here you only need to list you configuration for the bottom-right dataset of the classification problem.

2. List your findings that how the learning rate, the activation function, the number of hidden layers and the regularization influence the performance and convergence rate.

Solution:

1. The values for the best configuration are provided respectively below:

Parameter	Value
Learning rate	0.01

2. The analyzed parameters and their influence for the performance and convergence rate are discussed respectively below:

   Learning Rate: as the name suggests, the learning rate relates to the rate at which the model learns; i.e. the amount of correction that is applied to weights with each training example. Hence, smaller values of the learning rate may result in a considerably slow convergence while larger values of learning rate may result in over-shooting as the convergence point may be missed due to large changes in the weights. Therefore, the learning rate should be set to a small enough value to ensure that it does not miss its convergence point, but a large enough value to reach convergence in a reasonable amount of time (i.e. avoid slow convergence). In practice, the learning set is initially rate is set to 0.01 and may be slightly modified depending on the given application.

   Activation Function:

   Number of Hidden Layers:

   Regularization:

4 IRLS for Logistic Regression