Homework #3 Solution

Due Mar 13 (Monday) at the beginning of class
Please show all work and intermediate steps. Late submission will lead to 0 credit.

Problem 1: Support Vector Machine

Support vector machine (SVM) is a popular model in machine learning used for classification. Mathematically, given a training dataset of $m$ points

$$(x_1, y_1), \ldots, (x_m, y_m)$$

where $x_i \in \mathbb{R}^n$ stands for the feature vector and $y_i \in \{1, -1\}$ stands for two classes. The goal is to find two parallel hyperplanes represented by $(w, b)$ with maximal margin that separates the two classes of data, such that for class with $y_i = 1$, we have $w^T x_i + b \geq 1$ and for class with $y_i = -1$, we have $w^T x_i + b \leq -1$. Hence, we wish to satisfy $y_i(w^T x_i + b) \geq 1$ for $i = 1, \ldots, m$.

![Illustration of SVM](image)

If the data is not fully separable, we allow for small margin errors $\epsilon_i > 0$, $i = 1, \ldots, m$, and we wish to also minimize these errors. This leads to solving the following optimization problem:

$$\min_{w, b, \epsilon} \frac{1}{2} \|w\|^2_2 + C \cdot \sum_{i=1}^{m} \epsilon_i$$

s.t. $y_i(w^T x_i + b) \geq 1 - \epsilon_i$, $i = 1, \ldots, m$  
$\epsilon_i \geq 0$, $i = 1, \ldots, m$  

where the parameter $C > 0$ plays a role of controlling the relative importance of minimizing the norm of $w$ (i.e., maximizing the margin) and minimize the errors. Note that this problem is indeed a convex optimization problem.
Exercise 1.1 (Lagrange Duality) Let $\alpha \geq 0$ and $\beta \geq 0$ be the Lagrange multipliers associated with the two constraints. Show that the Lagrange dual problem of $(P)$ is given by the quadratic program:

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

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

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \quad (D)$$

Moreover, show that the primal and dual optimal solutions satisfy that

$$\alpha_i = 0 \Rightarrow y_i (w^T x_i + b) \geq 1$$

$$\alpha_i = C \Rightarrow y_i (w^T x_i + b) \leq 1$$

$$0 < \alpha_i < C \Rightarrow y_i (w^T x_i + b) = 1$$

We call the data points with non-zero Lagrangian multipliers the support vectors.

Solution The Lagrange function is

$$L(w, b, \epsilon, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \cdot \sum_{i=1}^{m} \epsilon_i - \sum_{i=1}^{m} \alpha_i (y_i (w^T x_i + b) - (1 - \epsilon_i)) - \sum_{i=1}^{m} \beta_i \epsilon_i$$

The Lagrange dual function is

$$L(\alpha, \beta) = \inf_{w, b, \epsilon} L(w, b, \epsilon, \alpha, \beta)$$

The infimum is achieved when

$$\nabla_w L = \nabla_b L = \nabla_\epsilon L = 0$$

which implies that

$$w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0$$

$$- \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$C - \alpha_i - \beta_i = 0, \forall i = 1, \ldots, m$$

Hence, the Lagrange dual function is

$$L(\alpha, \beta) = \begin{cases} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j), & \text{if } \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i + \beta_i = C \forall i = 1, \ldots, m \\ -\infty, & \text{otherwise} \end{cases}$$

Therefore, Lagrange dual problem is given by

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

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

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m \quad (D)$$

The complementary slackness of KKT conditions says that for $i = 1, \ldots, m$

$$\alpha_i (y_i (w^T x_i + b) - 1 + \epsilon_i) = 0 \quad (1)$$

$$\beta_i \epsilon_i = 0 \quad (2)$$

Hence, we have
If \( \alpha_i = 0 \), we have \( \beta_i = C \) and \( \epsilon_i = 0 \), this implies that \( y_i(w^T x + b) \geq 1 \).

If \( \alpha_i = C \), we have \( y_i(w^T x_i + b) - 1 + \epsilon_i = 0 \) and \( \epsilon_i \geq 0 \), this implies that \( y_i(w^T x + b) \leq 1 \).

If \( \alpha_i \in (0, C) \), we have \( y_i(w^T x_i + b) - 1 + \epsilon_i = 0 \) and \( \epsilon_i = 0 \), this implies that \( y_i(w^T x + b) = 1 \).

**Exercise 1.2 (Reformulation)** Show that (P) can be equivalently written as an unconstrained convex problem

\[
\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b), 0) + \lambda \|w\|^2
\]

where \( \lambda > 0 \) is some parameter.

**Solution** The constraints in (P) implies that

\[
\epsilon_i \geq 1 - y_i(w^T x_i + b) \text{ and } \epsilon_i \geq 0
\]

which is equivalent to

\[
\epsilon_i \geq \max(1 - y_i(w^T x_i + b), 0)
\]

Hence, (P) can be rewritten as

\[
\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \cdot \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b), 0)
\]

Let \( C = \frac{1}{2xm} \) for some \( \lambda > 0 \), then it can be further reformulated as

\[
\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} \max(1 - y_i(w^T x_i + b), 0) + \lambda \|w\|_2^2
\]

**Exercise 1.3 (Programming)** Implement the Ellipsoid method to solve the problem \( (P') \) in Matlab or Python whichever you prefer. Your input should be the data matrix \( X \), \( y \) and the parameter \( \lambda \), and the maximum number of iterations \( T \). Your output should be the best solution and objective function value obtained after running the algorithm within \( T \) iterations.

**Solution** Sample code provided.

**Exercise 1.4 (Test on Real Dataset)** Apply your algorithm with \( T = 100 \) iterations on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset \((n = 30, m = 569)\) provided (read here for detailed description of the dataset) with \( \lambda = 1 \).

- Plot the objective function values at current solution, i.e. \( f(w_t) \) vs the number of iteration \( t \);
- On the same figure, plot the objective function values at best solution, i.e. \( \min_{1 \leq \tau \leq t} f(w_\tau) \) vs the number of iteration \( t \);
- Compute the classification error: the ratio of misclassified points (i.e. \( y_i(w^T x_i + b) < 1 \)).
Solution  Sample result:

Figure 2: Ellipsoid Method for SVM on WBDC dataset