EPFL | MGT-418 : Convex Optimization | Project 2

Questions – Fall 2019

Classification with Smooth Hinge Loss (graded)

This project is due on **November 20, 2019, at 23:59**. You may form teams of up to three people. Each team should upload a single zip-file containing their report and Matlab code to Moodle. Make sure to clearly state the team members in your report.

Description

Given training samples (x_i, y_i), $i = 1, \ldots, m$, where $x_i \in \mathbb{R}^d$ (age, blood pressure...) are features and $y_i \in \{+1, -1\}$ (healthy vs. not healthy) are labels, the goal of classification is to predict the label of a new point $x \in \mathbb{R}^d$. As we have seen in the lecture, this is usually achieved by solving an empirical loss minimization problem of the form

$$\underset{w \in \mathbb{R}^d,\, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^{m} L(y_i(w^\top x_i - b)),$$

where L is a loss function that quantifies classification errors. The support vector machine (SVM) typically uses the hinge loss $L(z) = \max\{0, 1 - z\}$. The hinge loss has two key properties. On the one hand, it evaluates to zero for $z \geq 1$. Thus, an SVM model using the hinge loss does not benefit by pushing the training samples far away from the decision boundary. On the other hand, the hinge loss is piecewise linear and therefore not overly sensitive to outliers.

Unfortunately, the hinge loss is non-smooth and therefore susceptible to numerical difficulties. The smooth hinge loss inherits all desirable properties of the hinge loss but is smooth, which makes it numerically appealing. The smooth hinge loss function is defined as

$$L(z) = \begin{cases} \frac{1}{2} - z & \text{if } z \leq 0 \\ \frac{1}{2}(1 - z)^2 & \text{if } 0 < z < 1 \\ 0 & \text{if } z \geq 1. \end{cases}$$

In this exercise, we will solve the following SVM problem, which minimizes the sum of the smooth hinge loss of the prediction errors and a Tikhonov regularization term $\frac{\rho}{2} \|w\|_2^2$, where $\rho > 0$ is the regularization weight. The corresponding optimization problem reads as follows.

$$\underset{w \in \mathbb{R}^d,\, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^{m} L(y_i(w^\top x_i - b)) + \frac{\rho}{2} \|w\|_2^2 \tag{1}$$

Questions

1. **QCQP Reformulation (30 points)**: The infimal convolution of two functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is the function $h : \mathbb{R} \to [-\infty, \infty)$ defined through

   $$h(z) = \inf_{t \in \mathbb{R}} f(t) + g(z - t).$$

   Show that the infimal convolution of $f(z) = \frac{1}{2}z^2$ and $g(z) = \max\{0, 1 - z\}$ is equal to the smooth hinge loss function $L(z)$.

   Using this result, verify that problem (1) is equivalent to

   $$\begin{aligned} \underset{w \in \mathbb{R}^d,\, b \in \mathbb{R},\, t, s \in \mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{m} \sum_{i=1}^{m} s_i + \frac{\rho}{2} \|w\|_2^2 \\ \text{subject to} \quad & \frac{1}{2} t_i^2 + 1 - y_i(w^\top x_i - b) + t_i \leq s_i \quad \forall i = 1, \ldots, m \\ & \frac{1}{2} t_i^2 \leq s_i \quad \forall i = 1, \ldots, m. \end{aligned} \tag{2}$$

2. **Linear SVM (20 points)**: Each of the data files `p2data1.mat` and `p2data2.mat` available from Moodle contains 50 training samples. Solve problem (2) for each of these datasets with $\rho = 10^{-4}$. A skeleton of the code you will have to implement is provided in the Matlab file `p2q2.m`. Plot the SVM decision boundary, and briefly comment on the results.

3. **Kernel Trick (50 points)**: As in the lecture, let $\phi : \mathbb{R}^d \to \mathbb{R}^D$ be a feature map that lifts the inputs to a higher-dimensional space $\mathbb{R}^D$, $D \geq d$. The resulting lifted regression problem is

$$\begin{array}{ll} \underset{w\in\mathbb{R}^D,\, b\in\mathbb{R},\, t,s\in\mathbb{R}^m}{\text{minimize}} & \frac{1}{m}\sum_{i=1}^m s_i + \frac{\rho}{2}\|w\|_2^2 \\ \text{subject to} & \frac{1}{2}t_i^2 + 1 - y_i(w^\top \phi(x_i) - b) + t_i \leq s_i \quad \forall i = 1,\ldots,m \\ & \frac{1}{2}t_i^2 \leq s_i \quad \forall i = 1,\ldots,m. \end{array} \tag{3}$$

   3.1. Denote by λ_i and γ_i the Lagrange multipliers corresponding to the constraints $\frac{1}{2}t_i^2 + 1 - y_i(w^\top x_i - b) + t_i \leq s_i$ and $\frac{1}{2}t_i^2 \leq s_i$, respectively, and construct the Lagrangian function for problem (3). **(5 points)**

   3.2. Show that the Lagrangian dual of problem (3) can be formulated as problem (4) below. **(15 points)**

$$\begin{array}{ll} \underset{\lambda\in\mathbb{R}^m}{\text{maximize}} & \sum_{i=1}^m \left(\lambda_i - \frac{m}{2}\lambda_i^2\right) - \frac{1}{2\rho}\sum_{i=1}^m\sum_{i'=1}^m \lambda_i\lambda_{i'}y_iy_{i'}\phi(x_i)^\top\phi(x_{i'}) \\ \text{subject to} & \sum_{i=1}^m \lambda_i y_i = 0, \quad 0 \leq \lambda_i \leq \frac{1}{m} \quad \forall i = 1,\ldots,m \end{array} \tag{4}$$

   3.3. Use the KKT conditions to show that

$$w_j^\star = \frac{1}{\rho}\sum_{i=1}^m \lambda_i^\star y_i \phi_j(x_i) \quad \forall j = 1,\ldots,D \quad \text{and} \quad t_i^\star = -m\lambda_i^\star \quad \forall i = 1,\ldots,m$$

   at optimality. **(5 points)**

   3.4. Use the KKT conditions to show that $b^\star = y_k(m\lambda_k^\star - 1) + \frac{1}{\rho}\sum_{i=1}^m \lambda_i^\star y_i \phi(x_i)^\top\phi(x_k)$ for any $k \in \{1,\ldots,m\}$ such that $\lambda_k^\star \in (0, \frac{1}{m})$. **(5 points)**

   3.5. In contrast to the primal problem (3), the dual problem (4) can be solved without knowledge of the feature map ϕ. Instead, it suffices to know the kernel function $K(x,x') = \phi(x)^\top\phi(x')$. For the dataset provided in `p2data2.mat`, solve the dual problem (4) using the Gaussian Kernel

$$K(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$$

   with $\rho = 10^{-4}$ and $\sigma = 3$. A skeleton of the code you will have to implement is provided in the Matlab file `p2q3.m`. Plot the SVM decision boundary, and briefly comment on the results with comparison to the results from Question 2. **(20 points)**