MAT 5314

Assignment 2

Soufiane Fadel

September 23, 2019

Exercise 1

Suppose we are in the Euclidean regression setting and examples come from $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} = \mathbb{R}^d$ is input and $\mathcal{Y} = \mathbb{R}$ output. Consider the usual Tikhonov regularization with standard inner product in $\mathbb{R}^d$, but assume that there is an unpenalized offset term b,

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i\rangle + b - y_i)^2 + \lambda \left\| w \right\|^2 \right\} \tag{1}$$

and let $(w_*; b_*)$ be the solution of this problem. For $i = 1 \cdots n$, denote by $x_i^c = x_i - \bar{x}$ and $y_i^c = y_i - \bar{y}$ the centered data. Show that w_* also solves

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i^c\rangle - y_i^c)^2 + \lambda \left\| w \right\|^2 \right\} \tag{2}$$

solution:

let's start showing that if b is a solution of:

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i\rangle + b - y_i)^2 + \lambda \left\| w \right\|^2 \right\}$$

then we have : $b = \bar{y} - \langle w_*, \bar{x}\rangle_{\mathbb{R}^d}$

$$\begin{aligned}
b \text{ is an optimal solution of (1)} &\Rightarrow \frac{\partial\left[\sum_{i=1}^{n}(\langle w, x_i\rangle_{\mathbb{R}^d} + b - y_i)^2 + \lambda ||w||^2_{\mathbb{R}^d}\right]}{\partial b} = 0 \\
&\Rightarrow b = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle w, x_i\rangle_{\mathbb{R}^d}) \\
&\Rightarrow b = \frac{1}{n}\sum_{i=1}^{n} y_i - \langle w, \frac{1}{n}\sum_{i=1}^{n} x_i\rangle_{\mathbb{R}^d} \\
&\Rightarrow b = \bar{y} - \langle w, \bar{x}\rangle_{\mathbb{R}^d}
\end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

by substituting the above value of b and using linearity of the inner product on the original minimization problem we have get the following result:

$$\begin{aligned}
w_* &= \underset{w\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i\rangle + b - y_i)^2 + \lambda\left\|w\right\|^2 \right\} \\
&= \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i\rangle + \bar{y} - \langle w, \bar{x}\rangle_{\mathbb{R}^d} - y_i)^2 + \lambda\left\|w\right\|^2 \right\} \\
&= \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i - \bar{x}\rangle_{\mathbb{R}^d} - (y_i - \bar{y}))^2 + \lambda||w||^2_{\mathbb{R}^d}) \right\} \\
&= \underset{w\in\mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i^c\rangle_{\mathbb{R}^d} - y_i^c)^2 + \lambda||w||^2_{\mathbb{R}^d}) \right\}
\end{aligned}$$

where $x_i^c = x_i - \bar{x}$ and $y_i^c = y_i - \bar{y}$

therefore w_* also solves

$$\min_{w\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n}(\langle w\, x_i^c\rangle - y_i^c)^2 + \lambda\left\|w\right\|^2 \right\}$$

Exercise 2

♣ **part a):**

(i) why is $\alpha_i \neq 0$ iff $|w.x_i + b| = 1$?
(ii) explain why b_* is as claimed in equation (11) of Ng's notes.

solution:

(i)

- **if** $\alpha_i \neq 0$:

  By the complementarity conditions Equation (5) (Ng's notes):

  $$\alpha_i g_i(w^*) = 0 \quad \text{for } i = 1, \cdots, m$$

  where

  $$g_i(w) = -y_i(w.x_i + b) + 1 = 0$$

  Thus, the support vectors lie on the marginal hyperplanes $y_i(w.x_i + b) = 1$ and therefore we got $|w.x_i + b| = 1$

(ii)

After Having found w^*, by considering the primal problem, we can find the optimal value for the intercept term b by solving b such that for $i = 1, \cdots, m$:

$$
\begin{aligned}
y_i(w^*.x_i+b) \geq 1 &\Leftrightarrow \begin{cases} (w^*.x_i+b) \geq 1 & \text{if } y_i = 1 \\ (w^*.x_i+b) \leq -1 & \text{if } y_i = -1 \end{cases} \\
&\Leftrightarrow \begin{cases} b \geq 1 - w^*.x_i & \text{if } y_i = 1 \\ b \leq -1 - w^*.x_i & \text{if } y_i = -1 \end{cases} \\
&\Leftrightarrow \max(1 - w^*.x_i | y_i = 1) \leq b \leq \min(-1 - w^*.x_i | y_i = -1) \\
&\Leftrightarrow 1 - \min(w^*.x_i | y_i = 1) \leq b \leq -\max(w^*.x_i | y_i = -1) - 1 \\
&\Rightarrow b^* = \frac{(1 - \min(w^*.x_i|y_i = 1)) + (-\max(w^*.x_i|y_i = -1) - 1)}{2} \\
&\Rightarrow b^* = -\frac{\min(w^*.x_i|y_i = 1)) + \max(w^*.x_i|y_i = -1)}{2}
\end{aligned}
$$

♣ **part b):**

In the non-separable case, we have the following optimization problem:

$$
\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \right\}
$$
$$
\text{s.t} \quad y_i(w.x_i+b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \qquad \forall i
$$

(i) what are the KKT dual-complementary conditions for this problem?
(ii) why are the "support vectors", i.e. those for which $\alpha_i^* \neq 0$: now of two kinds: vectors on the marginal hyperplanes or outliers (respectively such points satisfy $|w.x_i + b| = 1$ or $\xi_i > 0$).
(iii) what value does α_i have in case of outliers?

solution:

(i)

The Lagrangian can then be defined for all $w \in \mathcal{R}^m$, $b \in \mathcal{R}$, and $\alpha \in \mathcal{R}_+^m$ by:

$$
\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i[y_i(w.x_i+b) - 1 + \xi_i)] - \sum_{i=1}^{m}\beta_i\xi_i
$$

The KKT complementarity conditions are :

$$
\alpha_i[y_i(w.x_i+b) - 1 + \xi_i] = 0 \qquad \forall i \tag{3}
$$
$$
\beta_i\xi_i = 0 \qquad \forall i \tag{4}
$$

(ii)

By the first complementarity condition (3), if $\alpha_i \neq 0$ (the "support vectors"), then $y_i(w.x_i + b) = 1 - \xi_i$.

- **If $\xi_i = 0$:** If $\xi_i = 0$, then $|w.x_i + b| = 1$ and then x_i lies on a marginal hyperplane, as in the separable case.
- **If $\xi_i > 0$:** Otherwise, if $\xi_i > 0$ and x_i is an outlier. In this case, the support vectors x_i are either outliers.

(iii)

By setting the gradient of the Lagrangian with respect to the primal variables ξ_i's to zero, we obtain :

$$\nabla_{\xi_i}\mathcal{L} = C - \alpha_i - \beta_i = 0 \quad \Rightarrow \quad \alpha_i + \beta_i = C \tag{5}$$

on the other hand if x_i is an outlier then $\xi_i > 0$ and and by the second complementarity condition (4) and then we have : $\beta_i = 0$. Finally by the equation (5) we have : $\alpha_i = C$

♣ **part c):**

Given training data

$$(x_1; y_1) = (-1; -1); \quad (x_2; y_2) = (-0.8; 1); \quad (x_3; y_3) = (1; 1)$$

is there a separating hyperplane? Suppose we nevertheless run the variant of SVM with slack variables discussed above for the non-separable case. Are there some choices of C which would result in the algorithm picking a non-separating hyperplane at $x = 0$?

solution:

Yes there is a clear hyper-plan $x = -0.9$ that separate the data. here is a figure that illustrate that :

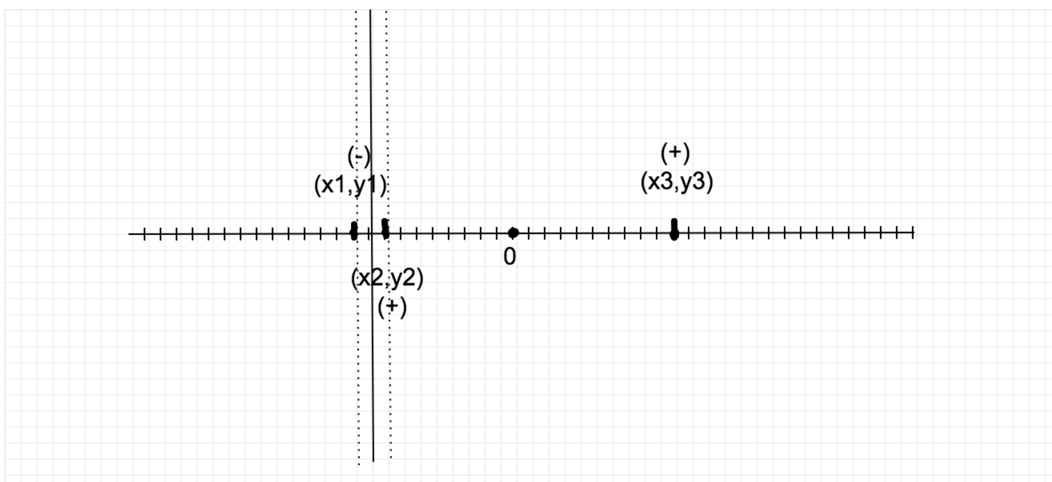

Figure 1: separating hyperplane

if we use the variant of SVM with slackvariables for the non-separable cas, an hyperplane at $x = 0$ shloud be like :

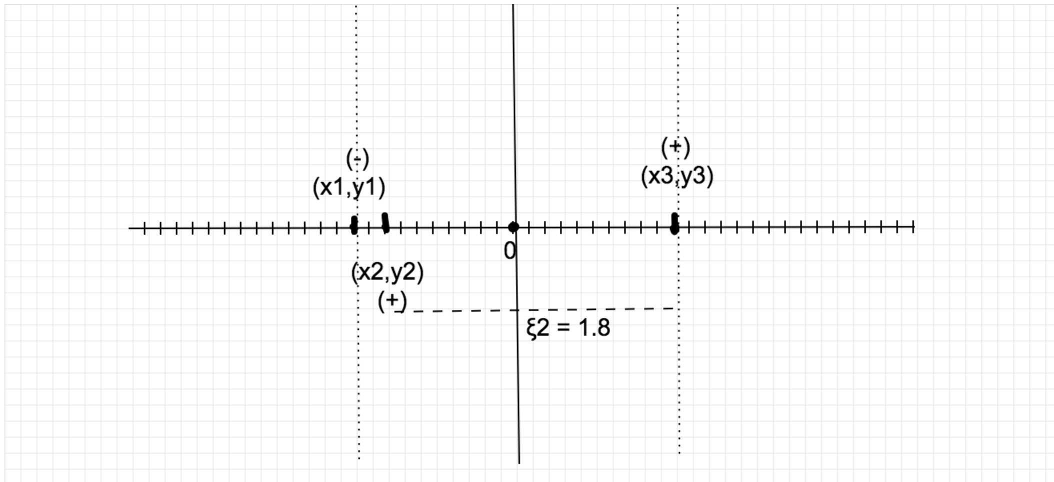

Figure 2: hyperplane at x=0

the value of C will be determined by KKT-conditions :

$$\begin{cases} w^* = y_1x_1\alpha_1 + y_2x_2\alpha_2 + y_3x_3\alpha_3 \\ y_1\alpha_1 + y_2\alpha_2 + y_3\alpha_3 = 0 \\ \alpha_2 = C \\ 0 \le \alpha_1 \le C \\ 0 \le \alpha_3 \le C \end{cases} \Leftrightarrow \begin{cases} 1 = \alpha_1 - 0.8\alpha_2 + \alpha_3 \\ -\alpha_1 + \alpha_2 + \alpha_3 = 0 \\ \alpha_2 = C \\ 0 \le \alpha_1 \le C \\ 0 \le \alpha_3 \le C \end{cases}$$

$$\Leftrightarrow \begin{cases} \alpha_1 = 0.5 + 0.9C \\ \alpha_2 = C \\ \alpha_3 = 0.5 - 0.1 \\ 0 \le \alpha_1 \le C \\ 0 \le \alpha_3 \le C \end{cases}$$

$$\Rightarrow \begin{cases} 5 \le C \\ 0.45 \le C \le 5 \end{cases}$$

$$\Rightarrow C = 5$$

finally the value that we choose for C is : $C = 5$

Exercise 3

part (b) Theory:

You are given a dataset of x, y pairs $\{(x_i; y_i)\}_{i=1}^N$ with $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. Assume that $n_+; n_-$ of the x_i have label +1,-1 respectively (so $n_+ + n_- = N$) and also assume you are given a kernel K and an associated feature map $\Phi : \mathcal{X} \mapsto \mathcal{F}$ to some Hilbert space $\mathcal{F}$ so:

$$K(x; x') = \langle\ \Phi(x), \Phi(x')\rangle_{\mathcal{F}}$$

Derive a classification rule, involving only kernel products (and the sign function), that assigns to a new test point the label of the class whose mean is closest in the feature space.

solution:

This problem is binary classification problem where our data is divide the training into two sets X_+ and X_- containing the positive and negative examples respectively. also we need to define $\mu_+ = \frac{1}{n_+}\sum_{x_+ \in X_+} \Phi(x_+)$

the feature-map average for the positive labels. similarly $\mu_- = \frac{1}{n_-}\sum_{x_-\in X_-}\Phi(x_-)$ the feature-map average for the negative labels.

A simple classification rule would be to assign x to the class corresponding to the smaller **distance**:

$$\begin{aligned} h(x) &= \begin{cases} +1 & \text{if } d_-(x) > d_+(x) \\ -1 & \text{otherwise} \end{cases} \\ &= \operatorname{sign}\left(d_-(x) - d_+(x)\right) \end{aligned}$$

where $d_+(x) = ||\Phi(x) - \mu_+||_{\mathcal{F}}$ and $d_-(x) = ||\Phi(x) - \mu_-||_{\mathcal{F}}$

$$\begin{aligned} d_+(x) &= ||\Phi(x) - \mu_+||_{\mathcal{F}} \\ &= \sqrt{\langle \Phi(x) - \frac{1}{n_+}\sum_{x_+\in X_+}\Phi(x_+), \Phi(x) - \frac{1}{n_+}\sum_{x_+\in X_+}\Phi(x_+)\rangle_{\mathcal{F}}} \\ &= \sqrt{\langle \Phi(x), \Phi(x)\rangle_{\mathcal{F}} - \frac{1}{n_+}\sum_{x_+\in X_+} 2\langle \Phi(x), \Phi(x_+)\rangle_{\mathcal{F}} - \frac{1}{n_+^2}\sum_{x_+\in X_+}\sum_{x'_+\in X_+}\langle \Phi(x'_+), \Phi(x_+)\rangle_{\mathcal{F}}} \\ &= \sqrt{K(x,x) - \frac{1}{n_+}\sum_{x_+\in X_+} 2K(x,x_+) - \frac{1}{n_+^2}\sum_{x_+\in X_+}\sum_{x'_+\in X_+} K(x'_+, x_+)} \end{aligned}$$

similarly :

$$\begin{aligned} d_-(x) &= ||\Phi(x) - \mu_-||_{\mathcal{F}} \\ &= \sqrt{\langle \Phi(x) - \frac{1}{n_-}\sum_{x_-\in X_-}\Phi(x_-), \Phi(x) - \frac{1}{n_-}\sum_{x_-\in X_-}\Phi(x_-)\rangle_{\mathcal{F}}} \\ &= \sqrt{\langle \Phi(x), \Phi(x)\rangle_{\mathcal{F}} - \frac{1}{n_-}\sum_{x_-\in X_-} 2\langle \Phi(x), \Phi(x_-)\rangle_{\mathcal{F}} - \frac{1}{n_-^2}\sum_{x_-\in X_-}\sum_{x'_-\in X_-}\langle \Phi(x'_-), \Phi(x_-)\rangle_{\mathcal{F}}} \\ &= \sqrt{K(x,x) - \frac{1}{n_-}\sum_{x_-\in X_-} 2K(x,x_-) - \frac{1}{n_-^2}\sum_{x_-\in X_-}\sum_{x'_-\in X_-} K(x'_-, x_-)} \end{aligned}$$

Therefore $h(x)$ a classification rule, involving only kernel products and the sign function, that assigns to a new test point the label of the class whose mean is closest in the feature space.