Georgia Institute of Technology

ECE 6254 - Assignment 7

Fall 2020 - v1.3

- There is 1 problem over 15 pages (including the cover page).
- Each question is graded as follows: no credit without meaningful work, half credit for partial work, full credit if essentially correct.
- Unless otherwise specified, you should concisely indicate your reasoning and show all relevant work.
- The grade on each question is based on our judgment of your level of understanding as reflected by what you have written. If we cannot read it, we cannot grade it.
- Please use a pen and not a pencil if you handwrite your solution.
- **You must submit your assignment on Gradescope**.

Problem 1: Support vector regression

The objective of this take-home exam is to investigate *support vector regression*, the counterpart of *support vector classification*. This problem involves both theoretical and programming parts. The theoretical parts rely heavily on applying concepts of convex optimization seen, which the programming part illustrates the power of the technique. You are *allowed* and *encouraged* to use built-in python functions provided by `scikit-learn` in the programming part.

In the entire problem, we consider a datset $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ that consists of N pairs of feature vectors in $\mathbb{R}^d$ and associated labels in $\mathbb{R}$. The objective of a *linear regression* is to find a model $(\mathbf{w}, b)$ that allows us to describe the relationship between feature vectors and labels as

$$y = \mathbf{w}^\intercal \mathbf{x} + b,$$

where $\mathbf{w}^\intercal$ is the transpose of $\mathbf{w}$.

The next set of questions introduce a few concepts related to hyperplanes. The questions are meant to help you appreciate the geometric meaning of the concepts.

Definition. The set of points $\mathcal{H} \triangleq \{(\mathbf{x}, y) \in \mathbb{R}^{d+1} : y = \mathbf{w}^\intercal \mathbf{x} + b\}$ is called an affine *hyperplane*.

[Q1] Sketch what a hyperplane looks like in the special case where $d = 1$. In other words, sketch the shape of the space $\mathcal{H}$ in the $\mathbb{R}^2$ plane with coordinate (x, y). Make sure you provide a short justification.

[Q2] Show that $\mathcal{H}$ as defined above is *not* a vector space in $\mathbb{R}^d$ unless $b = 0$. In other words, show that at least one property of vector spaces is not satisfied by $\mathcal{H}$.

[Q3] Show that $\mathcal{H}$ can be described as $\mathcal{H} = (\mathbf{x}_0, y_0) + \mathcal{V}$, where $(\mathbf{x}_0, y_0)$ is *any* point in $\mathcal{H}$ and $\mathcal{V}$ is a subvector space of $\mathbb{R}^{d+1}$ defined by $\mathcal{V} \triangleq \{(\mathbf{x}, y) : y = \mathbf{w}^\intercal \mathbf{x}\}$. (*Note:* you must show equality and you must prove that $\mathcal{V}$ is a subvector space)

Definition. For $\eta > 0$, the set $\{(\mathbf{x}, y) : \mathbf{w}^\intercal \mathbf{x} + b - \eta < y < \mathbf{w}^\intercal \mathbf{x} + b + \eta\}$ is called the η-band of a hyperplane $\mathcal{H}$ defined by $(\mathbf{w}, b)$.

[Q4] Sketch what an η-band of a hyperplane $\mathcal{H}$ looks like in the special case $d = 1$. No justification is required for this question.

Definition. Given a dataset $\mathcal{D}$ and $\eta > 0$, we say that a hyperplane $\mathcal{H}$ defined by $(\mathbf{w}, b)$ is the hard η-band hyperplane of $\mathcal{D}$ if all the points of $\mathcal{D}$ are in the η-band of $\mathcal{H}$, i.e.

$$\forall i \in [\![1, N]\!] \quad \mathbf{w}^\intercal \mathbf{x}_i + b - \eta < y_i < \mathbf{w}^\intercal \mathbf{x}_i + b + \eta.$$

[Q5] Show that we can always choose η large enough such that a hard η-band hyperplane of $\mathcal{D}$ exists.

[Q6] Show that there also exist some $\eta_0 \geqslant 0$ for which there always exists a hard η-band hyperplane of $\mathcal{D}$ if $\eta > \eta_0$ and such that no hard η-band hyperplane of $\mathcal{D}$ exists for $\eta \leqslant \eta_0$ (*Hint:* Think about defining η_0 as the solution of an optimization problem). Under what condition do we have $\eta_0 = 0$?

[Q7] Illustrate with a sketch that for $\eta > \eta_0$, there may exists several distinct hard η-band hyperplanes of $\mathcal{D}$. No justification is required.

The notion of hard η-band hyperplane is useful to talk about linear regression. There is little chance that the linear model $y = \mathbf{w}^\intercal \mathbf{x} + b$ is exact, and we can think of η as a way to introduce some slack and obtain an approximate model. The next set of questions are designed to guide you through the construction of a hard η-band hyperplane of $\mathcal{D}$ using ideas from linear classification.

For some $\eta > \eta_0$ and a dataset $\mathcal{D}$, define two sets

$$\mathcal{D}^+ \triangleq \{(\mathbf{x}_i, y_i + \eta) : i \in [\![1, N]\!]\} \qquad \mathcal{D}^- \triangleq \{(\mathbf{x}_i, y_i - \eta) : i \in [\![1, N]\!]\}$$

[Q8] Sketch an example of what the sets $\mathcal{D}$, $\mathcal{D}^+$ and $\mathcal{D}^-$ look like in the special case $d = 1$. No justification is required.

[Q9] Show that for a training set $\mathcal{D}$ and $\eta > 0$, $\mathcal{H} \triangleq \{(\mathbf{x}, y) : y = \mathbf{w}^\intercal \mathbf{x} + b\}$ is a hard η-band hyperplane of $\mathcal{D}$ *if and only if* $\mathcal{D}^+$ and $\mathcal{D}^-$ are located on both sides of $\mathcal{H}$ and no point is in $\mathcal{H}$.

The result above is useful because it provides us with a way to transform a regression problem into an equivalent classification problem. Specifically, we can interpret a hard η-band hyperplane of $\mathcal{D}$ as a separating hyperplane for the classes $\mathcal{D}^+$ and $\mathcal{D}^-$, to which we associate the label $+1$ and -1, respectively. In turn, we can now formulate the problem of finding a good hard η-band hyperplane of $\mathcal{D}$ as a margin maximization for a classification problem.

Note that the equivalent classification problem is in $\mathbb{R}^{d+1}$. We therefore set $\mathbf{z} \triangleq [\mathbf{x}, y]^\intercal \in \mathbb{R}^{d+1}$. The equation of a separating hyperplane is therefore $\tilde{\mathbf{w}}^\intercal \mathbf{z} + b = 0$, which upon writing $\tilde{\mathbf{w}} \triangleq [\mathbf{w}, \omega]^\intercal$ takes the form

$$\mathbf{w}^\intercal \mathbf{x} + \omega y + b = 0.$$

[Q10] Let $\eta > \eta_0$. Show (using what we have discussed in class) that the maximum margin optimization problem associated to the equivalent classification problem takes the form

$$\min_{\mathbf{w},\omega,b} \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{2}\omega^2 \text{ such that for } i \in [\![1, N]\!] \begin{cases} \mathbf{w}^\intercal \mathbf{x}_i + \omega(y_i + \eta) + b \geqslant 1, \\ \mathbf{w}^\intercal \mathbf{x}_i + \omega(y_i - \eta) + b \leqslant -1, \end{cases} \tag{1}$$

In what follows, we let $\mathbf{w}^\sharp, \omega^\sharp, b^\sharp$ be the solution to the optimization problem (1).

[Q11] Why does the solution $\mathbf{w}^\sharp, \omega^\sharp, b^\sharp$ exist when $\eta > \eta_0$?

[Q12] Show that the optimal solution must satisfy $\omega^\sharp \neq 0$ (*Hint:* look at the constraints in the original optimization problem).

[Q13] Define $\epsilon \triangleq \eta - \dfrac{1}{\omega^\sharp}$. Using again the constraints of the optimization problem (1), show that $\omega^\sharp \geqslant \frac{1}{\eta}$. Conclude that $0 \leqslant \epsilon < \eta$.

[Q14] Define $\mathbf{w}^* \triangleq -\dfrac{\mathbf{w}^\sharp}{\omega^\sharp}$ and $b^* \triangleq -\dfrac{b^\sharp}{\omega^\sharp}$. Show that $(\mathbf{w}^*, b^*)$ is the solution of the optimization problem

$$\min_{\mathrm{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 \text{ such that } \forall i \in [\![1, N]\!] \begin{cases} \mathbf{w}^\intercal \mathbf{x}_i + b - y_i \leqslant \epsilon \\ \mathbf{w}^\intercal \mathbf{x}_i + b - y_i \geqslant -\epsilon. \end{cases} \tag{2}$$

[Q15] Reasoning by contradiction and inspecting the constraints of the problem (2), show that $\epsilon \geqslant \eta_0$.

[Q16] The solution $(\mathbf{w}^*, b^*)$ of the problem (2) may not be unique. In fact, show graphically with a sketch that if ϵ is large enough, there are many solutions of the form $(\mathbf{0}, b^*)$.

This result suggests that we can solve the regression problem by picking some $\epsilon > \eta_0$ and focusing on solving the optimization problem (2) above, which is in $\mathbb{R}^d$. The result gives us a linear regression function $y = (\mathbf{w}^*)^\intercal \mathbf{x} + b^*$. The next set of questions are designed to walk you through the kernelization of the problem. We will follow the procedure already seen in class that consists in using Lagrangian duality.

[Q17] Form the Lagrangian associated to the optimization problem (2). Make sure to denote by α_i the Lagrange multipliers associated to inequalities $\mathbf{w}^\intercal \mathbf{x}_i + b - y_i \leqslant \epsilon$ and to denote by β_i those associated to the inequalities $\mathbf{w}^\intercal \mathbf{x}_i + b - y_i \geqslant -\epsilon$. The resulting Lagrangian should therefore be $\mathcal{L}(\mathbf{w}, b, \{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N)$.

[Q18] Using the stationary part of the KKT conditions, show that the optimal value $\mathbf{w}^*$ can be expressed as a function of the Lagrange multipliers $\{\alpha_i\}$ and $\{\beta_i\}$ as

$$\mathbf{w}^* = \sum_{i=1}^{N} (\beta_i - \alpha_i) \mathbf{x}_i.$$

[Q19] Use the result above to show that $\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N)$ takes the form

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N) = -\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N(\beta_i - \alpha_i)(\beta_j - \alpha_j)\mathbf{x}_i^\intercal \mathbf{x}_j - \epsilon\sum_{i=1}^N(\beta_i + \alpha_i) + \sum_{i=1}^N y_i(\beta_i - \alpha_i) - b\sum_{i=1}^N(\beta_i - \alpha_i).$$

[Q20] Use the above to show that the dual function is

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N) = -\infty \text{ if } \sum_{i=1}^N(\beta_i - \alpha_i) \neq 0$$

and

$$\min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N) = -\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N(\beta_i - \alpha_i)(\beta_j - \alpha_j)\mathbf{x}_i^\intercal \mathbf{x}_j - \epsilon\sum_{i=1}^N(\beta_i + \alpha_i) + \sum_{i=1}^N y_i(\beta_i - \alpha_i) \text{ if } \sum_{i=1}^N(\beta_i - \alpha_i) = 0.$$

[Q21] Use the result above to show that the dual optimization problem is

$$\min_{\{\alpha_i\}_{i=1}^N, \{\beta_i\}_{i=1}^N} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\beta_i - \alpha_i)(\beta_j - \alpha_j) \mathbf{x}_i^\intercal \mathbf{x}_j + \epsilon \sum_{i=1}^N (\beta_i + \alpha_i) - \sum_{i=1}^N y_i (\beta_i - \alpha_i) \tag{3}$$

such that $\sum_{i=1}^N (\beta_i - \alpha_i) = 0$ and $\forall i \in [\![1, N]\!]$ $\alpha_i, \beta_i \geqslant 0$.

Note that the dual optimization problem (3) is a convex quadratic program that can be solved efficiently! Most importantly, we can actually relate the solutions of the dual problem in (3) to the primal problem in (2). In the following, we denote the optimal solution of the dual problem (3) by $\left\{\alpha_i^*\right\}_{i=1}^N, \left\{\beta_i^*\right\}_{i=1}^N$.

[Q22] Using the KKT conditions, we have already established that $\mathbf{w}^* = \sum_{i=1}^N (\beta_i^* - \alpha_i^*)\mathbf{x}_i$. Use the complementary slackness condition to argue that for $i \in [\![1, N]\!]$ such that $\alpha_i^* \neq 0$ we must have

$$b = y_i - (\mathbf{w}^*)^\intercal \mathbf{x}_i + \epsilon.$$

and that for $j \in [\![1, N]\!]$ such that $\beta_j^* \neq 0$ we must have

$$b = y_j - (\mathbf{w}^*)^\intercal \mathbf{x}_j - \epsilon.$$

[Q23] Conclude from the above that we cannot simultaneously have $\beta_i \neq 0$ and $\alpha_i \neq 0$.

The points $(\mathbf{x}_i, y_i)$ for which either $\alpha_i^* \neq 0$ or $\beta_i^* \neq 0$ are called *support vectors*. To appreciate this terminology, it is useful to develop a geometric understanding of what they are.

[Q24] Assume $\epsilon > \eta_0$. Using the previous result, show that the support vectors are on one of the boundaries of the ϵ-band of the hyperplane $y = (\mathbf{w}^*)^\intercal \mathbf{x} + b^*$. Also conclude that the points that are not support vectors are *inside* the ϵ-band.

The next questions are programming questions. You do not really have to program anything, the goal is to make sure that you understand how to use the function `SVR` in `sklearn.svm`.

[Q25] Initialize your python notebook as follows

```
from sklearn.svm import SVR
np.random.seed(2020) ;
x = np.arange(100)
y = 2*x + 50*np.random.randn(x.size)
eps = 100
```

```
6    Model = SVR(C=1e7, epsilon=eps, kernel='linear')
7    Model.fit(x.reshape(-1,1),y)
```

Explain why setting `C=1e7` is required to reproduce the results analyzed theoretically so far. This will require you to read the information associated to the model carefully.

[Q26] Provide the plot and the associate code showing in a 2D plot: 1) the data using red dots (●); 2) the model learned as a blue line (–); 3) the ϵ-band around the model as two additional blue lines; and 4) the support vectors as blue squares (■). All the information you need is available in `scikit`, but you *must explicitly* indicate in your code (I suggest adding comments) how you retrieve and use the information required to create the plot.

[Q27] Repeat the experiment with `eps=50`. What is happening? If possible, use a plot to support your answer.

[Q28] Repeat the experiment with `eps=190`. What is happening? If possible, use a plot to support your answer.

[Q29] Repeat the experiment with `eps=250`. What is happening? If possible, use a plot to support your answer.

[Q30] Explain how you would kernelize this support vector regression.