STAT 154: Homework 6

Release date: **Sunday, April 7**

Due by: **11 PM, Sunday, April 21**

The honor code

(a) Please state the names of people who you worked with for this homework. You can also provide your comments about the homework here.

(b) Please type/write the following sentences yourself and sign at the end. We want to make it *extra* clear that nobody cheats even unintentionally.

I hereby state that all of my solutions were entirely in my words and were written by me. I have not looked at another students solutions and I have fairly credited all external sources in this write up.

Submission instructions

- It is a good idea to revisit your notes, slides and reading; and synthesize their main points BEFORE doing the homework.
- No .Rnw file is provided. You may use templates from previous homeworks if you want.
- **For the parts that ask you to implement/run some R code, your answer should look something like this (code followed by result):**

```
myfun<- function(){
show('this is a dummy function')
}
myfun()

## [1] "this is a dummy function"
```

Note that this is automatically generated if you use the R sweave environment.

- You need to submit the following:

  1. A pdf of your write-up to “HW6 write-up” that includes code for Problem 4.
  2. No *separate* code submission is required for this HW. You have to include the code in your submission for Problem 4 in the write-up itself.

- Ensure a proper submission to gradescope, otherwise it will not be graded. **This time we will not entertain any regrade requests for improper submission.**

Homework Overview

This homework covers kernel ridge regression and classification. The first problems attempts to make you comfortable with computational complexity related questions.

1 Computational complexity (10 pts)

Read about the big-O notation for the wiki page https://en.wikipedia.org/wiki/Big_O_notation and then answer the following using the big-O notation.

In the following questions, by computational complexity we refer to the number of addition/multiplication operations between two real numbers. To elaborate, we assume that two real numbers or multiplying them takes unit operations. Adding k real numbers has k computational complexity. Computing the multiplication of k pairs of numbers would have k computational complexity as well.

Now let $\mathbf{a}, \mathbf{v} \in \mathbb{R}^d$ and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times d}$ and answer the following questions:

1. (2 pts) What is the computational complexity of computing $\mathbf{a}+\mathbf{v}$? What is the computational complexity of computing $\mathbf{a}^\top \mathbf{v}$?

2. (2 pts) What is the computational complexity of computing the matrix $\mathbf{A}+\mathbf{B}$? How much space does storing the matrix $\mathbf{A}$ require?

3. (4 pts) What is the computational complexity of computing the vector $\mathbf{A}\mathbf{v}$? How about computing the matrix $\mathbf{A}^\top\mathbf{B}$?

4. (2 pts) What is the complexity of computing the vector $\mathbf{A}^\top\mathbf{B}\mathbf{v}$?
Hint: There are two ways to do it, one is naive and one is smart. You are encouraged to think and report both of them in your answer.

2 Kernel Methods (23 pts)

For the following problems, you can assume that inverting a $p\times p$ matrix takes order p^3 operations, i.e., the computational complexity of matrix inversion of size p is $O(p^3)$. Also for this problem, we use slightly different notation for dimensions in order to remain consistent with the note and the lecture.

1. (2 pts) Let $\mathbf{X} \in \mathbb{R}^{n\times \ell}$ and $\mathbf{y} \in \mathbb{R}^n$. Recall that the ridge estimate for the problem
$$\min_{\theta\in\mathbb{R}^\ell} \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \lambda \|\theta\|_2^2 \tag{1}$$
is given by $\widehat{\theta}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda \mathbf{I}_\ell)^{-1}\mathbf{X}^\top\mathbf{y}$. What is the computational complexity of computing this estimate?
Hint: You may use answers from previous question.

2. (2 pts) Show that $\widehat{\theta}_\lambda = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_n)^{-1}\mathbf{y}$ is also a valid estimate for the ridge problem 1. What is the complexity of computing this estimate?

3. (2 pts) Compare and contrast the computational complexities from the previous two parts.

4. (2 pts) Suppose we modify the problem (1) using a feature map $\phi : \mathbb{R}^\ell \to \mathbb{R}^d$ as follows:

$$\min_{\widetilde{\theta} \in \mathbb{R}^d} \left\| \mathbf{\Phi} \widetilde{\theta} - \mathbf{y} \right\|_2^2 + \lambda \|\widetilde{\theta}\|_2^2 \tag{2}$$

$$\text{where} \quad \mathbf{\Phi} = \begin{bmatrix} \text{—}\phi(\mathbf{x}_1)^\top\text{—} \\ \vdots \\ \text{—}\phi(\mathbf{x}_n)^\top\text{—} \end{bmatrix} \in \mathbb{R}^{n \times d}. \tag{3}$$

Can you provide a few reasons why we may want to do this? (Usually $d \geq \ell$ when we extend the $\mathbf{x}$ vectors in this fashion.)

5. (8 pts) As discussed in class, often the choice of ϕ is (implicitly or explicitly) made such that

$$\phi(\mathbf{x})^\top \phi(\mathbf{z}) = k(x, z) \tag{4}$$

where $k : \mathbb{R}^\ell \times \mathbb{R}^\ell \to \mathbb{R}$ denotes the kernel function (that satisfies some nice properties). Can you compute the kernel functions for the following feature maps:

   (a) $\phi(x) = [x_1 x_2, \frac{x_1^2}{\sqrt{2}}, \frac{x_2^2}{\sqrt{2}}]^\top$ where $\mathbf{x} \in \mathbb{R}^2$ with $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$.

   (b) $\phi(x) = [1, x, \frac{x^2}{\sqrt{2!}}, \frac{x^3}{\sqrt{3!}}, \ldots]$ where $x \in \mathbb{R}$.

On the reverse side can you compute the feature map for the following kernel functions (it may not be possible in some cases):

   (c) $k(x, z) = (1 + \mathbf{x}^\top \mathbf{z})^2 + \mathbf{x}^\top \mathbf{z}$ for $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$.

   (d) $k(x, z) = e^{-(x-z)^2}$ for $x, z \in \mathbb{R}$.

6. (2 pts) What are the two different ways of computing the solution to the problem (2)?

7. (5 pts) We now compare the complexity of the two estimates for a polynomial kernel. Compare and contrast the computational complexity of the two estimates when $\mathbf{x} \in \mathbb{R}^\ell$ and the feature map ϕ is chosen such that the corresponding kernel function (4) is given by

$$k(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^\top \mathbf{z})^p.$$

Discuss when is an estimate better than the other (on computational grounds).

3 LDA and linear regression (10 pts)

ESLII book https://web.stanford.edu/~hastie/Papers/ESLII.pdf: Question 4.2 (all 5 parts)

4 Applied problem for classification (7 pts)

ISL Book: 4.11 (all 7 parts) **Show code in the write-up pdf for all parts.**