EE 8591 **Homework 2 (10 pts)** **Fall 2021**

Due date: Oct 14, 2021

Topics: *VC-theory + model selection.*

Problem 1 (3 pts)

Use linear least squares regression to estimate dependency between life expectancy (response variable) and 4 input variables using the data shown below. Inputs include: murder rate, percentage of High School graduates, annual income, and illiteracy rate.

(a) Analyze the regression coefficients and explain/discuss how the value of regression coefficients can be used to interpret the model.

(b) Estimate life expectancy as a linear function of murder rate (single input). Show this dependency in a graphical form.

(c) can selecting a subset of input variables improve prediction accuracy of the model estimated in part (a)? Explain your answer. For example, you can include

- additional modeling results showing mean squared error estimated via resampling;
- arguments based on interpretation of regression coefficients in part (a).

STATE NAME	LIFE EXPECT.	MURDER	HSGRAD	INCOME	ILLITERACY
Alabama	69.1	15.1	41.3	3624	2.1
Alaska	69.3	11.3	66.7	6315	1.5
Arizona	70.6	7.8	58.1	4530	1.8
Arkansas	70.7	10.1	39.9	3378	1.9
California	71.7	10.3	62.6	5114	1.1
Colorado	72.1	6.8	63.9	4884	0.7
Connecticut	72.5	3.1	56.0	5348	1.1
Delaware	70.1	6.2	54.6	4809	0.9
Florida	70.7	10.7	52.6	4815	1.3
Georgia	68.5	13.9	40.6	4091	2.0
Hawaii	73.6	6.2	61.9	4963	1.9
Idaho	71.9	5.3	59.5	4119	0.6
Illinois	70.1	10.3	52.6	5107	0.9
Indiana	70.9	7.1	52.9	4458	0.7
Iowa	72.6	2.3	59.0	4628	0.5
Kansas	72.6	4.5	59.9	4669	0.6
Kentucky	70.1	10.6	38.5	3712	1.6
Louisiana	68.8	13.2	42.2	3545	2.8
Maine	70.4	2.7	54.7	3694	0.7
Maryland	70.2	8.5	52.3	5299	0.9
Massachusetts	71.8	3.3	58.5	4755	1.1
Michigan	70.6	11.1	52.8	4751	0.9
Minnesota	73.0	2.3	57.6	4675	0.6
Mississippi	68.1	12.5	41.0	3098	2.4
Missouri	70.7	9.3	48.8	4254	0.8
Montana	70.6	5.0	59.2	4347	0.6

Nebraska	72.6	2.9	59.3	4508	0.6
Nevada	69.0	11.5	65.2	5149	0.5
NewHampshire	71.2	3.3	57.6	4281	0.7
NewJersey	70.9	5.2	52.5	5237	1.1
NewMexico	70.3	9.7	55.2	3601	2.2
NewYork	70.6	10.9	52.7	4903	1.4
NorthCarolina	69.2	11.1	38.5	3875	1.8
NorthDakota	72.8	1.4	50.3	5087	0.8
Ohio	70.8	7.4	53.2	4561	0.8
Oklahoma	71.4	6.4	51.6	3983	1.1
Oregon	72.1	4.2	60.0	4660	0.6
Pennsylvania	70.4	6.1	50.2	4449	1.0
RhodeIsland	71.9	2.4	46.4	4558	1.3
SouthCarolina	68.0	11.6	37.8	3635	2.3
SouthDakota	72.1	1.7	53.3	4167	0.5
Tennessee	70.1	11.0	41.8	3821	1.7
Texas	70.9	12.2	47.4	4188	2.2
Utah	72.9	4.5	67.3	4022	0.6
Vermont	71.6	5.5	57.1	3907	0.6
Virginia	70.1	9.5	47.8	4701	1.4
Washington	71.7	4.3	63.5	4864	0.6
WestVirginia	69.5	6.7	41.6	3617	1.4
Wisconsin	72.5	3.0	54.5	4468	0.7
Wyoming	70.3	6.9	62.9	4566	0.6

Problem 2. (2 pts)
Find analytic estimate of VC-dimension for a set of functions formed by the union of two spheres in two-dimensional space. Each sphere is defined by its center and radius parameters.

Problem 3 (3 pts)
Consider model selection for k-nearest neighbors' regression.
Consider two model selection criteria to select an optimal *k* value:

- VC bound: Eq (4.14) in the textbook, where VC-dimension is $h=n/k/n^{1/5}$.
- Using the L_2 measure, where L_2 measure is the MSE between the prediction *y* and true target value *t(x)*. Note that using this L_2 measure always yields the 'best possible' value of *k*.

For empirical comparison, apply these two methods to:
(a) sine-squared data set generated according to: $y_i = \sin^2 (2\pi x_i) + \xi$, where inputs x_i are sampled from the uniform random distribution in [0,1], and noise ξ is Gaussian N(0, 0.1).
(b) pure noise data (x, y) where x-values are from uniform random distribution in [0,1], and y-values are Gaussian noise with $\sigma = 1$.

Present your comparisons in a graphical form, showing boxplots for optimal k-values selected by each method, for 100 random realizations of training data. Show results for three different sample sizes n= 25, 50 and 100 samples. That is, each Figure displays 3 box plots of selected k-values (one box plot for each sample size). There will be total of 4 figures (~ two methods applied to two different data sets).

4. Problem 4 (2 pts)

Consider signal denoising using Fast Fourier Transform (FFT) – where a signal (given as a function of time) is represented in the frequency domain, as a superposition of harmonic functions. Then harmonic functions with 'small' coefficients are discarded, and the denoised signal contains only harmonic functions with large coefficients. If you are not familiar with FFT denoising, see this short video https://www.youtube.com/watch?v=s2K1JfNR7Sc

Provide VC-theoretical interpretation of FFT signal denoising, i.e. how it can be related to structural risk minimization. That is,

- what is the type of structure implemented in FFT signal denoising);
- can you suggest a different structure (complexity ordering) on a set of harmonic functions.

Also, comment on the differences between signal denoising (in classical signal processing) and VC-theoretical framework.