BA476 Problem set 1 (40) Due date: Friday 16 Feb at 23:55

This is an individual assignment, you may discuss it with your colleagues but do not share solutions with each other. If you get stuck on something you are always welcome to stop by office hours to talk it through with us.

Answer each question on a new page and submit a pdf on gradescope. When submitting you will be prompted to indicate on which page you answered each question. Failing to do so may result in parts of your submission not being graded.

1. Explain whether each scenario is a classification or regression problem, a prediction or inference problem, and provide n (the number of observations in the data) and p (the number of predictors in the data).

   (a) (2 points) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 12 similar products that were previously launched. For each product we have recorded whether it was a success or failure, the price charged for the product, the development budget, the marketing budget, the price of the product's main competitor, and 4 other variables.

   (b) (2 points) We want to understand which lifestyle factors cause tumorous growths. We have data on 53 patients from a clinical trial, including whether or not they had a tumor, their age, weight, income, location, activity level, diet and a feature capturing a family history of tumors.

2. Consider the following scenarios and decide whether or not the model is doing a good job. Motivate your answers.

   (a) (3 points) A classification algorithm is designed to detect fraudulent transactions in a dataset of credit card transactions. The algorithm achieves an accuracy rate of 95%. About 1 in 50 transactions in the dataset is actually fraudulent.

   (b) (3 points) An algorithm predicts home selling prices in Boston, where the average house price is \$707 000, The algorithm has out of sample MSE of 50 million.

   (c) (3 points) Roulette is a gambling game where a ball falls in a spinning wheel and comes to rest in one of 37 slots (numbers 0-36). When your correctly predict which (non-zero) slot the ball will land on, you win 36 times the amount you wagered. We build a model that correctly predicts which number a roulette ball will land 4\$ of the time, based on the handler, table, humidity, etc.

3. Consider a training set with the instances (x_i, y_i): $\{(-2,6),(-1,2),(0,1),(2,-2),(3.5,-5)\}$.

   (a) (3 points) Suppose we train a model $\hat{f}(x) = \beta_0 + \beta_1 x + \epsilon$ with $\beta_0 = 1, \beta_1 = -2$. What is the MSE of this model on the training set?

   (b) (2 points) For the same coefficients as before, what is the value of the lasso objective function for regularization parameter $\alpha = 0.1$.

   (c) (2 points) For the same coefficients as before, what is the value of the ridge objective function for regularization parameter $\alpha = 0.1$.

(d) (2 points) Compare a lasso model with regularization parameter $\alpha = 0.1$ to a lasso model with regularization parameter $\alpha = 1$ in terms of flexibility and expected training set MSE.

(e) (2 points) Which of the following models are linear, if any?

1. $\hat{f}_1(x) = \beta_0 + \beta_1 x + \beta_2 \log x + \epsilon$
2. $\hat{f}_2(x) = \beta_0 + \beta_1 x + x^{\beta_2} + \epsilon$
3. $\hat{f}_3(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \sqrt{x} + \epsilon$

(f) (2 points) Which of $\hat{f}(x)$ or $\hat{f}_1(x)$ do you expect will have lower MSE on the training set? Is this true in general (i.e. for any data)? Why or why not?

4. The table below shows, for every subset of predictors, the MSE on a specific regression problem.

Size 0		Size 1		Size 2		Size 3		Size 4	
Set	MSE	Set	MSE	Set	MSE	Set	MSE	Set	MSE
$\emptyset$	27.5	X_1	27.5	X_1, X_2	22	X_1, X_2, X_3	12	X_1, X_2, X_3, X_4	12
		X_2	26	X_1, X_3	24	X_1, X_2, X_4	13		
		X_3	25	X_1, X_4	18	X_1, X_3, X_4	14		
		X_4	26.5	X_2, X_3	25	X_2, X_3, X_4	14.5		
				X_2, X_4	14.5				
				X_3, X_4	15				

(a) (3 points) Execute forward selection for the problem in the table. For each size $1, \ldots, 4$, state which subsets of predictors are considered as well as which set is selected.

(b) (1 point) Execute backwards selection. Which sets of predictors are selected?

(c) (2 points) List the sizes where forwards or backwards selection chose the optimal subset.

5. Consider the following training set and corresponding euclidean distance matrix. We will use KNN as estimator. All numbers are reported to two decimals.

Instance	X1	X2	Y
0	1	-1	1
1	2	3	5
2	-1	4	-1
3	3	2	4
4	5	-0.5	2

dists	0	1	2	3	4
0	0.0	4.12	5.39	3.61	4.03
1	4.12	0.0	3.16	1.41	4.61
2	5.39	3.16	0.0	4.47	7.5
3	3.61	1.41	4.47	0.0	3.2
4	4.03	4.61	7.5	3.2	0.0

(a) (3 points) Let $k = 2$. For each instance, provide its k closes neighbours in the training set (potentially including the instance itself) and the resulting predicted label.

(b) (2 points) Let $k = 3$. For each instance, provide its k closes neighbours in the training set (potentially including the instance itself) and the resulting predicted label.

(c) (3 points) For a new instance with X1 = 4, X2 = 6, compute the euclidean distance to every instance in the training set and make the instance's prediction for $k = 2$. (Refer to the wiki link above for a definition of Euclidean distance if needed.)