Final Exam

- When you finish the exam, please upload your solutions to Gradescope by **Friday 5/6 2:00pm**.
- This exam is open book, open note, and you are free to use any computational resources like Python or Matlab.
- You may **not** talk to each other about the exam and you may **not** copy solutions from the internet.
- As you work, please recall the Tufts University statement on academic integrity:

> "Academic integrity is the joint responsibility of faculty, students, and staff. Each member of the community is responsible for integrity in their own behavior and for contributing to an overall environment of integrity at the university."

1	/18
2	/10
3	/10
4	/2
Total	/40

1. For this problem, download and import the dataset `Fish.csv` found on Canvas. The dataset includes the weight in grams and various size measurements in centimeters, for 7 types of fish that are commonly sold in markets.

   (a) *(2 point)* Build a multiple regression model $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_5 X_5$ to predict the `Weight` ($\hat{Y}$) in grams using the five size measurements as the predictors $(X_1, X_2, ..., X_5)$.

   (b) *(2 point)* Assuming you used an in-built solver to build your model, what is the optimization problem the linear-least squares estimator you used tries to solve? Write the problem in terms of a matrix A, the coefficient vector $\beta = [\beta_0, ..., \beta_5]$, and Y.

   (c) *(2 point)* Interpet the coefficient, β_4, corresponding to the predictor, `Height` (i.e. X_4), that is, describe the expected mean change in `Weight` in grams for 1 centimeter increase in `Height`.

   (d) *(2 point)* What is the R^2 for your model? Is this a good model based on this metric?

   (e) *(2 point)* Create a histogram of the residuals, $r_i = Y_i - \hat{Y}_i$, for this model. Does the model satisfy the assumption that the residuals need to be normally distributed?

   (f) *(3 point)* Based on your model, what is the expected `Weight`, $\hat{Y}$, in grams for an average-sized fish in the market?

   Rebuild your model with one difference: standardize the dependent variable `Weight`, name this model $\hat{Y}_{st} = \beta_{0,st} + \beta_{1,st} X_1 + \beta_{2,st} X_2 + ... + \beta_{5,st} X_5$.

   (g) *(3 point)* Are your coefficients the same as before? Derive an expression to find the original coefficient, β_3, from the coefficient, $\beta_{st,3}$, of the model with the standardized `Weight`. Derive an expression to find the original coefficient, β_0, from the coefficient, $\beta_{st,0}$, of the model with the standardized Weight.

   (h) *(2 point)* A friend gives you an idea: instead of building `Weight` models using the fish size measurements as predictors, you could build a multiple regression model $\hat{Y} = \beta_0 + \beta_6 X_6 + \beta_7 X_7$ where X_6 = a binary variable which takes on the value 1 if the fish species is a perch and is 0 for non-perch fish and X_7 = a binary variable which takes on the value 1 for non-perch fish and is 0 for perch fish. Is this a good idea? Why or why not?

2. Given n points $\{x_1, ..., x_n\}$ and their labels $\{y_1, ..., y_n\}$ for $1 \leq i \leq n$ with each $x_i \in \mathbb{R}^d$, consider the dual optimization problem corresponding to the hard margin SVM classification problem

$$\max_{\substack{\alpha_i > 0 \\ i=1,...,n}} -\frac{1}{2} \sum_{j,k} \alpha_j \alpha_k \cdot y_j y_k \cdot \langle x_j, x_k \rangle + \sum_{i=1}^{n} \alpha_i$$

subject to

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

Let $\{\alpha_i^*\}_{i=1}^n$ denote the optimal solution to the problem above. Assume that there exists a point x_s such that $y_s(w^* x_s + b^*) = 1$ where y_s is the label of x_s and w^* and b^* denote the optimal solutions to the primal (i.e. non-dual) optimization problem.

(a) *(4 point)* What is the form of the classifier function in terms of $\{\alpha_i^*\}_{i=1}^n$ and $\{(x_i, y_i) : 1 \leq i \leq n\}$? *(Hint: the classifier function should only depend on the dual variables and the data points and their labels. It should not be given in terms of b^* and w^*.)*

(b) *(4 point)* Suppose that we use a Gaussian kernel,

$$K(x_i, x_j) = \exp\left(\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

for the SVM classification problem. What is the form of the classifier in terms of $\{\alpha_i^*\}_{i=1}^n$, $\{(x_i, y_i) : 1 \leq i \leq n\}$ and the Gaussian kernel. *(Hint: the classifier function should only depend on the dual variables and the data points and their labels. It should not be given in terms of b^* and w^*.)*

(c) *(2 point)* Let $n = 1000$. Assume that $\alpha_3^* = 0$. Suppose we remove the third training data point (x_3, y_3). How does the classifier function change? Justify your answer.

3. (a) *(5 point)* Suppose we have a network of perceptrons with a fixed set of weights and biases for each perceptron; that is, for the i^{th} perceptron of the j^{th} layer, the output is given by

$$y_{ij} = \begin{cases} 0 & w_{ij}^{\intercal}x + b_{ij} \leq 0 \\ 1 & w_{ij}^{\intercal}x + b_{ij} > 0 \end{cases}$$

where w_{ij} is a vector of weights and b_{ij} is a scalar bias. Show that multiplying all of the weights and biases by a constant $c > 0$ doesn't change the behavior of the network.

(b) *(5 point)* Now replace the perceptrons from part (a) by neurons with sigmoid activation functions; that is, for the i^{th} neuron of the j^{th} layer, the output is given by

$$y_{ij} = \frac{1}{1 + e^{-(w_{ij}^{\intercal}x + b_{ij})}}.$$

Now multiplying all of the weights and biases by a constant $c > 0$ will change the behavior of the network. To keep things simple, we will assume that for our data, $w^{\intercal}x + b$ is never equal to 0. Show that as $c \to \infty$ this network of neurons is the same as the network of perceptrons from part (a).

4. (a) *(1 point)* What made you interested in taking this class?

   (b) *(1 point)* Which programming language did you prefer to use this semester?