1. Assume that you are given the following sample . Estimate the weight of people whose heights are 150, 155, 165, and 190 cm, using KNN with $k = 3$:

$$\hat{y}_{KNN} = \frac{y_1 + y_2 + \cdots + y_k}{k}$$

where $y_1, y_2, \cdots, y_k$ are the labels of the k nearest neighbors to your test instance. (10 pts)

Person	Height (cm)	Weight (kg)
1	171	80
2	168	78
3	191	100
4	182	80
5	150	65
6	178	83

2. Repeat 1, but instead of using the simple average of the labels of k nearest neighbors, which is use the following weighted average:

$$\hat{y}_{KNN} = \frac{w_1 y_1 + w_2 y_2 + \cdots + w_k y_k}{w_1 + w_2 + \cdots + w_k}$$

where the weight w_i for the label y_i of instance i is determined as $1/d_i$, where d_i the distance between the instance i and the test instance. (10 pts)

3. Assume that $J(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c$ where $\mathbf{Q} = \mathbf{Q}^T \in \mathbb{R}^{n\times n}$, $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$, and $c \in \mathbb{R}|$. Show that $\nabla_{\mathbf{x}} J(\mathbf{x}) = 2\mathbf{Q}\mathbf{x} + \mathbf{d}$ and $\mathbf{H} = \frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{Q}$. $\mathbf{H}_{ij} = \frac{\partial^2 J}{\partial x_i \partial x_j}$ and $\mathbf{H}$ is called the Hessian matrix of J. (10 pts)

4. Write down the prediction $\widehat{\mathbf{y}}$ for a test row vector $\mathbf{x}'_{1\times p}$ made by a linear regression model in terms of $\mathbf{y}$ the vector of labels of the training set and $\mathbf{X}_{n\times(p+1)}$, the (augmented) feature matrix, and explain why $\widehat{\mathbf{y}}$ can be viewed as a special case of KNN regression. (10 pts)

5. Show that the for $\mathbf{y} \in \mathbb{R}^n$, $\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a member of the column space of $\mathbf{X}$, i.e. is a linear combination of the columns of $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$. (10 pts)

6. Show that in linear regression, if $\widehat{\beta}$ minimizes RSS(β), then $\mathbf{y} - \widehat{\mathbf{y}}$ is orthogonal to the column space of $\mathbf{X}$. (10 pts)

7. **Programming Part: Vertebral Column Data Set**

This Biomedical data set was built by Dr. Henrique da Mota during a medical residence period in Lyon, France. Each patient in the data set is represented in the data set by six biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine (in this order): pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and grade of spondylolisthesis. The following convention is used for the class labels: DH (Disk Hernia), Spondylolisthesis (SL), Normal (NO) and Abnormal (AB). In this exercise, we only focus on a binary classification task NO=0 and AB=1.[1]

(a) Download the Vertebral Column Data Set from: `https://archive.ics.uci.edu/ml/datasets/Vertebral+Column`.

(b) Pre-Processing and Exploratory data analysis: (10 pts)

   i. Make scatterplots of the independent variables in the dataset. Use color to show Classes 0 and 1.

   ii. Make boxplots for each of the independent variables. Use color to show Classes 0 and 1 (see ISLR p. 129).

   iii. Select the first 70 rows of Class 0 and the first 140 rows of Class 1 as the training set and the rest of the data as the test set.

(c) Classification using KNN on Vertebral Column Data Set (20 pts)

   i. Write code for k-nearest neighbors with Euclidean metric (or use a software package).

   ii. Test all the data in the test database with k nearest neighbors. Take decisions by majority polling. Plot train and test errors in terms of k for $k \in \{208, 205, \ldots, 7, 4, 1, \}$ (in reverse order). You are welcome to use smaller increments of k. Which k^* is the most suitable k among those values? Calculate the confusion matrix, true positive rate, true negative rate, precision, and F_1-score when $k = k^*$.[2]

   iii. Since the computation time depends on the size of the training set, one may only use a subset of the training set. Plot the *best test error rate*,[3] which is obtained by some value of k, against the size of training set, when the size of training set is $N \in \{10, 20, 30, \ldots, 210\}$.[4] Note: for each N, select your training set by choosing the first $\lfloor N/3 \rfloor$ rows of Class 0 and the first $N - \lfloor N/3 \rfloor$ rows of Class 1 in the training set you created in 7(b)iii. Also, for each N, select the optimal k from a set starting from $k = 1$, increasing by 5. For example, if $N = 200$, the optimal k is selected from $\{1, 6, 11, \ldots, 196\}$. This plot is called a *Learning Curve*.

Let us further explore some variants of KNN.

(d) Replace the Euclidean metric with the following metrics[5] and test them. Summarize the test errors (i.e., when $k = k^*$) in a table. Use all of your training data and select the best k when $\{1, 6, 11, \ldots, 196\}$. (10 pts)

[1] Make sure that you convert labels to 0 and 1, otherwise you may not obtain correct answers.

[2] We will learn in the lectures what these mean, for now research how they are computed and compute them.

[3] Obviously, use the test data you created in 7(b)iii

[4] For extra practice, you are welcome to choose smaller increments of N.

[5] You can use sklearn.neighbors.DistanceMetric. Research what each distance means.

i. Minkowski Distance:

A. which becomes Manhattan Distance with $p = 1$.

B. with $\log_{10}(p) \in \{0.1, 0.2, 0.3, \dots, 1\}$. In this case, use the k^* you found for the Manhattan distance in 7(d)iA. What is the best $\log_{10}(p)$?

C. which becomes Chebyshev Distance with $p \to \infty$

ii. Mahalanobis Distance.[6]

(e) The majority polling decision can be replaced by weighted decision, in which the weight of each point in voting is *inversely proportional* to its distance from the query/test data point. In this case, closer neighbors of a query point will have a greater influence than neighbors which are further away. Use weighted voting with Euclidean, Manhattan, and Chebyshev distances and report the best test errors when $k \in \{1, 6, 11, 16, \dots, 196\}$. (10 pts)

(f) What is the lowest training error rate you achieved in this homework? (5 pts)

[6] Mahalanobis Distance requires inverting the covariance matrix of the data. When the covariance matrix is singular or ill-conditioned, the data live in a linear subspace of the feature space. In this case, the features have to be transformed into a reduced feature set in the linear subspace, which is equivalent to using a pseudoinverse instead of an inverse.

1. Assume that you are given the following sample . Estimate the weight of people whose heights are 150, 155, 165, and 190 cm, using KNN with $k = 3$:

$$\hat{y}_{KNN} = \frac{y_1 + y_2 + \cdots + y_k}{k}$$

where $y_1, y_2, \cdots, y_k$ are the labels of the k nearest neighbors to your test instance. (10 pts)

Person	Height (cm)	Weight (kg)
1	171	80
2	168	78
3	191	100
4	182	80
5	150	65
6	178	83

191 190 71
182 78) 12
178
171
168
150

1. when height is 150 :

3 nearest neighbors → Person 5, Person 2, and Person 1

$$\hat{y}_{KNN} = \frac{65 + 78 + 80}{3} = \frac{223}{3} = 74.33 \text{ (kg)}$$

when height is 155 :

3 nearest neighbors → Person 5, Person 2, and Person 1

$$\hat{y}_{KNN} = \frac{65 + 78 + 80}{3} = \frac{223}{3} = 74.33 \text{ (kg)}$$

when height is 165 :

3 nearest neighbors → Person 2, Person 1 and Person 6

$$\hat{y}_{KNN} = \frac{78 + 80 + 83}{3} = \frac{241}{3} = 80.33 \text{ (kg)}$$

when height is 190 :

3 nearest neighbors → Person 3, Person 4 and Person 6

$$\hat{y}_{KNN} = \frac{100 + 80 + 83}{3} = \frac{263}{3} = 87.67 \text{ (kg)}$$

2. Repeat 1, but instead of using the simple average of the labels of k nearest neighbors, which is use the following weighted average:

$$\hat{y}_{KNN} = \frac{w_1 y_1 + w_2 y_2 + \cdots + w_k y_k}{w_1 + w_2 + \cdots + w_k}$$

where the weight w_i for the label y_i of instance i is determined as $1/d_i$, where d_i the distance between the instance i and the test instance. (10 pts)

when height is 150 :

3 nearest neighbors → Person 5 , Person 2 , and Person 1

d_i → 0 , 18 , and 21

w_i → $1/0$ (infinite) , so $\hat{y}_{KNN}$ = 65 (kg)

when height is 155 :

150 $\overset{168}{155}$ $\overset{177}{155}$

3 nearest neighbors → Person 5 , Person 2 , and Person 1

d_i → 5 , 13 , and 16

w_i → $1/5$, $1/13$, and $1/16$

$$\hat{y}_{KNN} = \frac{65 \times 1/5 + 78 \times 1/13 + 80 \times 1/16}{1/5 + 1/13 + 1/16} \fallingdotseq 70.11$$

when height is 165 :

168 $\overset{177}{165}$ $\overset{178}{165}$

3 nearest neighbors → Person 2 , Person 1 and Person 6

d_i → 3 , 6 , and 13

w_i → $1/3$, $1/6$, and $1/13$

$$\hat{y}_{KNN} = \frac{78/3 + 80/6 + 83/13}{1/3 + 1/6 + 1/13} \fallingdotseq 79.24$$

when height is 190 :

191 $\overset{190}{182}$ $\overset{190}{178}$

3 nearest neighbors → Person 3 , Person 4 and Person 6

d_i → 1 , 8 , and 12

$$w_i \rightarrow 1/1 ,\ 1/8 ,\ \text{and}\ 1/12$$

$$\hat{y}_{KNN} = \frac{100/1 + 80/8 + 83/12}{1/1 + 1/8 + 1/12} \doteqdot 96.76$$

3. Assume that $J(\mathbf{x}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c$ where $\mathbf{Q} = \mathbf{Q}^T \in \mathbb{R}^{n\times n}$, $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$, and $c \in \mathbb{R}|$. Show that $\nabla_{\mathbf{x}} J(\mathbf{x}) = 2\mathbf{Q}\mathbf{x} + \mathbf{d}$ and $\mathbf{H} = \frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{Q}$. $\mathbf{H}_{ij} = \frac{\partial^2 J}{\partial x_i \partial x_j}$ and $\mathbf{H}$ is called the Hessian matrix of J. (10 pts)

3.

$$\begin{aligned} \nabla_x J(x) &= \nabla_x (x^T Q x + d^T x + c) \\ &= \nabla_x (x^T Q x + d^T x) + 0 \\ &= \nabla_x (x^T Q x) + d + 0 \\ &= 2Qx + d + 0 \\ &= 2Qx + d \end{aligned}$$

$$\begin{aligned} H = \frac{\partial^2 J}{\partial x \partial x^T} &= \nabla_x^2 J(x) \\ &= \nabla_x \nabla_x J(x) \\ &= \nabla_x (2Qx + d) \\ &= \nabla_x (2Qx) \\ &= 2Q \end{aligned}$$

4. Write down the prediction $\widehat{\mathbf{y}}$ for a test row vector $\mathbf{x}'_{1\times p}$ made by a linear regression model in terms of $\mathbf{y}$ the vector of labels of the training set and $\mathbf{X}_{n\times(p+1)}$, the (augmented) feature matrix, and explain why $\widehat{\mathbf{y}}$ can be viewed as a special case of KNN regression. (10 pts)

4. $x'_{1\times p} = (x'_1 \cdot x'_2 \cdots x'_p)$

$$\hat{y} = x' \times a = x' (X^T X)^{-1} X^T y$$

and then, $\hat{y}$ can be :

$$\hat{y} = \sum_{i=0}^{p} (x'_i \times a_i) = \sum_{i=0}^{p} (x'_i (X^T X)^{-1} X^T y_i),$$ i means the i-th sample

so $\hat{y}$ can be represented as the weighted average of samples

if we use distance to measure the similarity between samples. we can use a_i, which is weight, to represent it.

therefore. $\hat{y}$ is a special case of KNN regression.

5. Show that the for $\mathbf{y} \in \mathbb{R}^n$, $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a member of the column space of $\mathbf{X}$, i.e. is a linear combination of the columns of $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$. (10 pts)

5. $\hat{y} = X(X^TX)^{-1}X^Ty$. if $a = (X^TX)^{-1}X^Ty$.

$\rightarrow \hat{y} = Xa$, a is a vector and $X \in \mathbb{R}^{n\times(p+1)}$

when $X * a$, the result will be the column space of X.

this result is $\hat{y}$. Therefore , $\hat{y}$ is a column space of X.

6. Show that in linear regression, if $\widehat{\beta}$ minimizes RSS(β), then $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the column space of $\mathbf{X}$. (10 pts)

6. $RSS(\beta) = \|y - \hat{y}\|^2$

$= \|y - X\beta\|^2$

minimize $\hat{\beta}$, $\frac{\partial}{\partial\beta} RSS(\beta) = 0$

$\frac{\partial}{\partial\beta} RSS(\beta) = \frac{\partial}{\partial\beta}(\|y - X\beta\|^2) = \frac{\partial}{\partial\beta}(y - X\beta)^T(y - X\beta)$

$= 0 - 2X^Ty + 2X^TX\beta$

$= -2X^Ty + 2X^TX\beta$

$= -2X^T(y - X\beta)$

$\frac{\partial}{\partial\beta} RSS(\beta) = -2X^T(y - X\hat{\beta}) = 0$

$\cancel{2}X^TX\hat{\beta} = \cancel{2}X^Ty$

$\hat{\beta} = (X^TX)^{-1}X^Ty$ and $-2X^T(y - X\hat{\beta}) = 0$

$\Rightarrow X^T(y - \hat{y}) = 0$

$\Rightarrow$ $y - \hat{y}$ is orthogonal to the column space of X.