HW1

Yue Xu

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

① for 150 cm

- Person 1: $|171-150| = 21$ cm
- Person 2: $|168-150| = 18$ cm
- Person 3: $|191-150| = 41$ cm
- Person 4: $|182-150| = 32$ cm
- Person 5: $|150-150| = 0$ cm
- Person 6: $|178-150| = 28$ cm

∴ the 3 nearest neighbors are Person 5, 2, 1

∴ $\hat{y}^{KNN} = \frac{65+78+80}{3} = 74.33$ kg

② for 155 cm

- Person 1: $|171-155| = 16$ cm
- Person 2: $|168-155| = 13$ cm
- Person 3: $|191-155| = 36$ cm
- Person 4: $|182-155| = 27$ cm
- Person 5: $|150-155| = 5$ cm
- Person 6: $|178-155| = 23$ cm

∴ the 3 nearest neighbors are Person 5, 2, 1

∴ $\hat{y}^{KNN} = \frac{65+78+80}{3} = 74.33$ kg

③ for 165 cm

- Person 1: $|171-165| = 6$ cm
- Person 2: $|168-165| = 3$ cm
- Person 3: $|191-165| = 26$ cm
- Person 4: $|182-165| = 17$ cm
- Person 5: $|150-165| = 15$ cm
- Person 6: $|178-165| = 13$ cm

∴ the 3 nearest neighbors are Person 2, 1, 6

∴ $\hat{y}^{KNN} = \frac{78+80+83}{3} = 80.33$ kg

④ for 180 cm

- Person 1: $|171-180| = 19$ cm
- Person 2: $|168-180| = 22$ cm
- Person 3: $|191-180| = 1$ cm
- Person 4: $|182-180| = 8$ cm
- Person 5: $|150-180| = 40$ cm
- Person 6: $|178-180| = 12$ cm

∴ the 3 nearest neighbors are Person 3, 4, 6

∴ $\hat{y}^{KNN} = \frac{100+80+83}{3} = 87.67$ 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)

① for 150 cm

∵ distance between Person 5 to this instance is 0 , cannot be dividend

∴ we suppose $d = 0.1$ (enough small)

$$\hat{y}^{KNN} = \frac{\frac{1}{0.1}\times 65 + \frac{1}{18}\times 78 + \frac{1}{21}\times 80}{\frac{1}{0.1} + \frac{1}{18} + \frac{1}{21}} = 65.14 \text{ kg}$$

② for 155 cm

$$\hat{y}^{KNN} = \frac{\frac{1}{5}\times 65 + \frac{1}{13}\times 78 + \frac{1}{16}\times 80}{\frac{1}{5} + \frac{1}{13} + \frac{1}{16}} = 70.71 \text{ kg}$$

③ for 165 cm

$$\hat{y}^{KNN} = \frac{\frac{1}{3}\times 78 + \frac{1}{6}\times 80 + \frac{1}{13}\times 83}{\frac{1}{3} + \frac{1}{6} + \frac{1}{13}} = 79.24 \text{ kg}$$

④ for 180 cm

$$\hat{y}^{KNN} = \frac{\frac{1}{1}\times 100 + \frac{1}{8}\times 80 + \frac{1}{12}\times 83}{\frac{1}{1} + \frac{1}{8} + \frac{1}{12}} = 86.76 \text{ kg}$$

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)

① ∵ i) $\nabla_x Ax = A^T$

ii) $\nabla_x A^T x = A$

iii) $\nabla_x x^T Ax = Ax + A^T x$ (if $A = A^T$ then $\nabla_x x^T Ax = 2Ax$)

$\therefore \nabla_x J(x) = \frac{\partial J}{\partial x} = \frac{\partial}{\partial x}(x^T Qx + d^T x + c)$

$= \frac{\partial}{\partial x}(x^T Qx) + \frac{\partial}{\partial x}(d^T x) + \frac{\partial}{\partial x}(c)$

$= Qx + Q^T x + d + 0$

$= 2Qx + d$ ($\because Q = Q^T$)

② $H = \frac{\partial^2 J}{\partial x \partial x^T} = \frac{\partial}{\partial x^T}(2Qx + d)$

$= 2Q + 0$

$= 2Q$

4. Write down the prediction $\hat{\mathbf{y}}$ 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 $\hat{\mathbf{y}}$ can be viewed as a special case of KNN regression. (10 pts)

① ∵ in linear regression model $\hat{y} = x'w$

$w = (X^TX)^{-1}X^Ty$

∴ $\hat{y} = x'(X^TX)^{-1}X^Ty$

② ∵ $\hat{y} = x'(X^TX)^{-1}X^Ty$

∴ we can view $x'(X^TX)^{-1}X^T$ as the weight of each training points

$\hat{y} = \sum_{i=1}^{i=n} w_i y_i$

∴ it is equivalent to KNN, where
neighborhood is every training point
weights are determined by the linear regrassion geometry

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)

Suppose $\beta = (X^TX)^{-1}X^Ty$

$\therefore \hat{y} = X\beta$

$\therefore \beta \in \mathbb{R}^{p+1}$ is a weight vector, it provides the coefficients of the linear combination

$\therefore$ each column of X is multiplied by the corresponding element of β

$\therefore \hat{y}$ is a member of the column space of X

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)

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

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

$= \frac{\partial}{\partial \beta}(y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta)$

$= \frac{\partial}{\partial \beta}(y^Ty - 2y^TX\beta + \beta^TX^TX\beta)$

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

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

$\because \hat{\beta}$ minimizes $RSS(\beta)$

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

$\therefore X^T(y-\hat{y}) = 0 \quad (\because \hat{y} = X\hat{\beta})$

$\therefore y-\hat{y}$ is orthogonal to each column of X

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