CSC 4792 - DATA MINING

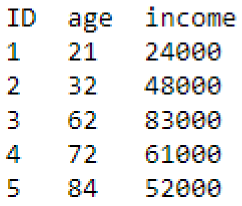
PROGRAMMING EXERCISE 1//2019

Question One

Given the data matrix $\mathbf{X} \in \mathbb{R}^{\boldsymbol{M}\mathbf{x}(\boldsymbol{d}+\mathbf{1})}$ whose rows are the inputs $\boldsymbol{x_i}$ as row vectors, and target vector $\boldsymbol{y} \in \mathbb{R}^{\boldsymbol{M}\mathbf{x1}}$ a column vector whose components are the target values $\boldsymbol{y_i}$. In class we discussed that in order to derive the linear regression algorithm, we need to minimize the in-sample error function of $\boldsymbol{w}$ and the data $\mathbf{X}, \mathbf{y}$ given by $\boldsymbol{E_{in}}(\boldsymbol{w}) = \frac{\mathbf{1}}{\mathbf{M}}\sum_{i=1}^{\mathbf{M}}(\mathbf{w}^{\mathbf{T}}\mathbf{x_i} - \mathbf{y_i})^{\mathbf{2}}$, over all possible $\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{d}+\mathbf{1}}$, the optimization problem formalized as $\boldsymbol{w} = \boldsymbol{argmin}_{w \in \mathbb{R}^{d+1}}\boldsymbol{E_{in}}(\boldsymbol{w})$. Show that besides gradient descent method, there is a closed-form solution to the problem given by $\mathbf{w} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}$, assuming $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is invertible.

Question Two

Consider the following data sampled from employees of a given organization about age against their income.



ID	age	income
1	21	24000
2	32	48000
3	62	83000
4	72	61000
5	84	52000

Tasks:

i. Plot the datapoints (age against income) to visualize the distribution in 2D.

ii. Using the closed-form solution in *question 1* or gradient descent, find the model that *suitably* predicts the income given the age of the employee in this organization.

iii. Plot your model in Question 2(ii) to visualize how it fits the datapoints.

iv. Using your model in Question (ii), what is the expected income if the employee is 50 year old?

v. If your model does not fit the dataset properly, use Polynomial regression to find the best fit model.

Prepared By: Claytone Sikasote