ACTU PS5841 Data Science in Finance & Insurance - Autumn 2022 (Y. Wang)
Assignment - 7 plus
Assigned 10/18/22, Due 10/23/22

The file `income.csv` records income of two groups of individuals, along with their years of education and job experience. Let's investigate the interaction between job experience and group membership.

Consider three models: a linear regression, a decision tree, and a K nearest neighbors regression. Fit each model specified below, use the fitted model to predict income for individuals with the same amount of education, but with varying amount of job experience.
Cohort A: `educ` = 18, `jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$, and `group` = A
Cohort B: `educ` = 18, `jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$, and `group` = B

Part A - Linear Regression
Please fit the following linear regression model with the dataset.

`income` $\sim$ `educ + jobexp + group + jobexp:group`

[*a*] Produce a single plot of `income` vs `jobexp` which contains
(a) The predicted difference in `income` between Cohort B and Cohort A as a function of `jobexp` which varies over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$. That is

$$\hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = B) - \hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = A)$$

(b) appropriate axis labels and legends
(c) calculate R^2 and report it in the chart title as "LM: Rsquared = #.##"

Part B - Decision Tree
Please fit a decision tree model with the dataset using `educ`, `jobexp` and `group` as features. Fit the tree to a maximum depth of 2 (which produces a similar R^2 as the linear regression model.)

[*b*] Produce a single plot of `income` vs `jobexp` which contains
(a) The predicted difference in `income` between Cohort B and Cohort A as a function of `jobexp` which varies over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$. That is

$$\hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = B) - \hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = A)$$

(b) appropriate axis labels and legends
(c) calculate R^2 and report it in the chart title as "Tree: Rsquared = #.##"

Part C - K Nearest Neighbors
Please fit a KNN model with the dataset using `educ`, `jobexp` and `group` as features. Fit the KNN model with $K = 43$ (which produces a similar R^2 as the linear regression model.) While `group` is a categorical feature, the sklearn KNN is OK as `group` is binary. However, you should standardize features before fitting the KNN(43) model and take the standardization into account when making predictions.

[*c*] Produce a single plot of `income` vs `jobexp` which contains
(a) The predicted difference in `income` between Cohort B and Cohort A as a function of `jobexp` which varies over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$. That is

$$\hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = B) - \hat{\text{income}}(\text{educ} = 18, \text{jobexp}, \text{group} = A)$$

(b) appropriate axis labels and legends
(c) calculate R^2 and report it in the chart title as "KNN: Rsquared = #.##"

Page 2 is the last page.

Please submit your work as

- hw7plus.ipynb and hw7plus.html with your code fully executed

to `Canvas`.