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

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.

For the midterm exam, our models do NOT incorporate years of education which is the most effective predictor of income. You will have a chance to include this variable in another homework assignment.

Do not be (too) concerned if you run out of time. Please try to accomplish as much as you can, in sequential order.

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

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

[a]&[b] Let your code produce the predicted income for
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = A
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = B
and report the output in a single table as shown in see `WriteupTemplate.docx`.

[c] Produce a single plot of `income` vs `jobexp` which contains
(a) observations on group A members in red
(b) model predicted `income` for group-A members in red with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$
(c) observations on group B in blue
(d) model predicted `income` for group-B members in blue with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$
(e) appropriate axis labels and legends
(f) 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 `jobexp` and `group` as the only features. Fit the tree to a maximum depth of 2 (which produces a similar R^2 as the linear regression model.)

[a] Please produce a legible visual of the fitted tree using `plot_tree()`.

[b]&[c] Let your code produce the predicted income for
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = A
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = B
and report the output in a single table as shown in see `WriteupTemplate.docx`.

[d] Produce a single plot of `income` vs `jobexp` which contains
(a) observations on group A members in red
(b) model predicted `income` for group-A members in red with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$
(c) observations on group B in blue
(d) model predicted `income` for group-B members in blue with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$

(e) appropriate axis labels and legends
(f) 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 `jobexp` and `group` as the only features. Fit the KNN model with $K = 12$ (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(12) model and take the standardization into account when making predictions.

[a]&[b] Let your code produce the predicted income for
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = A
`jobexp` = $[1, 1.5, 2, ..., 20.5, 21]$ and `group` = B
and report the output in a single table as shown in see `WriteupTemplate.docx`.

[c] Produce a single plot of `income` vs `jobexp` which contains
(a) observations on group A members in red
(b) model predicted `income` for group-A members in red with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$
(c) observations on group B in blue
(d) model predicted `income` for group-B members in blue with `jobexp` varying over 1 and 21 as $[1, 1.5, 2, ..., 20.5, 21]$
(e) appropriate axis labels and legends
(f) calculate R^2 and report it in the chart title as

"KNN: Rsquared = #.##"

Please submit your work as

- hw7.ipynb and hw7.html with your code fully executed
- WriteupTemplate.docx with your reported results

to `Canvas`.