CSC 4780/6780
Fall 2022
Homework 06

September 25, 2022

This homework is due at 11:59 pm on Sunday, Oct 2. It must be uploaded to iCollege by then. No credit will be given for late submissions. A solution will be released by noon on Monday, Oct 3.

Once again: it is a good idea to get this done and turn it in early. You can turn it in as many times as you like – iCollege will only keep the last submission.

1 What are we doing?

Often the data that we are given is not what we should feed into our model. So we might manufacture some new features from the data we have. And we might delete some irrelevant or redundant features. This is known as "Feature engineering" and "Feature selection".

2 Analyze the features

You are given `data.csv` which is, once again, real estate data that you will use to predict prices.

There are five columns:

- `property_id` Use this as the index of the dataframe.
- `sqft_hvac` This is the square foot measurement of the interior of the house.
- `lot_width` This is the width of the lot in feet.
- `lot_depth` This is the depth of the lot in feet.
- `age_of_roof` This is the age of the roof in years.
- `miles_to_school` This is the number of miles to the nearest elementary school.

- `price` This is the price that the house sold for.

Create a program called `analysis.py` to do feature selection.

Specifically, `analysis.py` should do the following:

- Read the csv into a dataframe.
- Get the input (X) and target (Y) as numpy arrays.
- Use PolynomialFeatures to expand X into 2nd degree polynomials. (This will automatically add a column of 1s)
- Create a loop that:
  - Calculates the p-value of the Pearson correlation between each input and the current residual. (The first time through the residual will just be Y.)
  - Sorts the inputs so that the p-values are in ascending order.
  - Add the input with the lowest p-value to the list of inputs you are going to actually use.
  - Do linear regression using that list of inputs.
  - Print the R^2 value.
  - Calculate a new residual.
- Finally, make two scatter plots:
  - `ResidualMiles.png` that compares the final residual and the texttmiles_to_school
  - `ResidualRoof.png` which compares the residual and the `age_of_roof`

The output should look like this:

```
> python3 analysis.py data.csv
First time through: using original price data as the residual
     "sqft_hvac" vs residual: p-value=0.0
     "sqft_hvac^2" vs residual: p-value=0.0
     "sqft_hvac lot_width" vs residual: p-value=0.0
     "sqft_hvac lot_depth" vs residual: p-value=0.0
     "sqft_hvac age_of_roof" vs residual: p-value=8.731082754175303e-55
     "lot_width lot_depth" vs residual: p-value=4.1868028438284917e-41
     "lot_depth^2" vs residual: p-value=1.1014229237235183e-34
     "lot_depth" vs residual: p-value=1.349112505398765e-32
     "lot_width^2" vs residual: p-value=5.993692365884732e-22
     "lot_width" vs residual: p-value=7.502300252189193e-22
     "sqft_hvac miles_to_school" vs residual: p-value=9.519591803622703e-21
     "lot_depth age_of_roof" vs residual: p-value=9.568489304909097e-09
     "miles_to_school" vs residual: p-value=6.314415223095078e-08
     "miles_to_school^2" vs residual: p-value=4.169604800671784e-06
```

```
     "lot_width miles_to_school" vs residual: p-value=3.3260830318919526e-05
     "age_of_roof miles_to_school" vs residual: p-value=0.002206196824401264
     "lot_width age_of_roof" vs residual: p-value=0.014189712836334
     "age_of_roof" vs residual: p-value=0.2204854707886266
     "age_of_roof^2" vs residual: p-value=0.37721705725549765
     "lot_depth miles_to_school" vs residual: p-value=0.6902534566522582
**** Fitting with ["1" "sqft_hvac" ] ****
R2 = 0.8285362066724605
Residual is updated
     "lot_width lot_depth" vs residual: p-value=4.63867435576096e-226
     "lot_depth" vs residual: p-value=1.2511219230454328e-213
     "lot_depth^2" vs residual: p-value=7.007419749277197e-210
     "sqft_hvac lot_depth" vs residual: p-value=1.3573402394546474e-79
     "miles_to_school" vs residual: p-value=1.143866412662615e-60
     "lot_width miles_to_school" vs residual: p-value=3.1979675246627807e-48
     "miles_to_school^2" vs residual: p-value=4.278477233456872e-46
     "sqft_hvac miles_to_school" vs residual: p-value=2.2362041512664205e-42
     "lot_depth age_of_roof" vs residual: p-value=5.324817730838837e-34
     "lot_width^2" vs residual: p-value=1.6087269118146826e-30
     "lot_width" vs residual: p-value=1.894797058154101e-30
     "age_of_roof miles_to_school" vs residual: p-value=3.073565041879731e-21
     "lot_width age_of_roof" vs residual: p-value=0.0024146107224859237
     "sqft_hvac lot_width" vs residual: p-value=0.007708854025643736
     "lot_depth miles_to_school" vs residual: p-value=0.012927560709559626
     "age_of_roof" vs residual: p-value=0.1250566876903974
     "sqft_hvac age_of_roof" vs residual: p-value=0.14177548490199277
     "age_of_roof^2" vs residual: p-value=0.20592884268913036
     "sqft_hvac^2" vs residual: p-value=0.955142095666689
     "sqft_hvac" vs residual: p-value=1.0000000000004001
**** Fitting with ["1" "sqft_hvac" "lot_width lot_depth" ] ****
R2 = 0.9278909375846092
Residual is updated
Making scatter plot: age_of_roof vs final residual
Making a scatter plot: miles_from_school vs final residual
```

So, you will go through the loop twice to get an X with two columns of input (and a column of 1s).

3 Make Predictions

Now stare at the two plots you made. Can you make a new variable that will let linear regression fit better? (Keep reading; there is a heavy-handed hint in the next couple of paragraphs.)

Now, make another program: `prediction.py`. Using the list of features that you found in `analysis.py`, this program should read in the csv, manufacture the meaningful variables, and

remove the less useful ones. The X matrix should have three columns. (Or four if you have a column of 1s.)

Do not use `sklearn.preprocessing.PolynomialFeatures` or `scipy.stats.pearsonr` in `prediction.py`.

Use this 3-column X and Y to make a formula for prediction. The output should look like this:

```
> python3 prediction.py data.csv
Making new features...
Using only the useful ones: ['sqft_hvac', 'lot_size', 'is_close_to_school']...
R2 = 0.96941
*** Prediction ***
Price = $23,846.11 + (sqft x $119.01) + (lot_size x $11.01)
     Less than 2 miles from a school? You get $49,300.87 added to the price!
```

4 Criteria for success

If your name is Fred Jones, you will turn in a zip file called `HW06_Jones_Fred.zip` of a directory called `HW06_Jones_Fred`. It will contain:

- `analysis.py`
- `prediction.py`
- `data.csv`
- `ResidualRoof.png`
- `ResidualMiles.png`

Be sure to format your python code with black before you submit it.

We will run your code like this:

```
cd HW06_Jones_Fred
python3 analysis.py data.csv
python3 prediction.py data.csv
```

Do this work by yourself. Stackoverflow is OK. A hint from another student is OK. Looking at another student's code is *not* OK.

The template files for the python programs have import statements. Do not use any frameworks not in those import statements.

5 Reading

You should have read through Chapter 10 : *Preparing Data for Machine Learning: Feature Selection, Feature Engineering, and Dimensionality Reduction.* (You can skip the dimensionality reduction for now – we will talk about that later.)