Microsoft: DAT210x Programming with Python for Data Science

Bookmarks

- Start Here
- 1. The Big Picture
- 2. Data And Features
- 3. Exploring Data
- **4. Transforming Data**

Lecture: Transformations

Lecture: PCA
Quiz

Lab: PCA
Lab

Lecture: Isomap
Quiz

Lab: Isomap
Lab

Lecture: Data Cleansing
Quiz

4. Transforming Data > Lab: PCA > Assignment 3

Bookmark

Lab Assignment 3

You're not quite done with chronic kidney disease yet—we still need to beat it! In the previous lab assignment, you focused only on three features out of the entire dataset: **bgr**, **rc**, and **wc**. That should have seemed strange to you. How did we know to direct your attention only to those features? The answer, of course, is through PCA. By running PCA on the raw dataset data, we were able to find suitable candidate features to show the importance of feature scaling. For this lab, there will be no starter code. Copy your finished Lab 2, **assignment2.py** file over as **assignment3.py** and start working from that.

1. Head back over to the dataset page (or you can look at the kidney_disease.names file in your /Module4/Datasets/ directory). Each column has a type listed, e.g. numeric, nominal, etc. Here is a formatted list of the nominal features for your copy and pasting pleasure:

   ```
   ['id', 'classification', 'rbc', 'pc', 'pcc', 'ba', 'htn', 'dm', 'cad', 'appet', 'pe',
   'ane']
   ```

   Instead of using an indexer to select just the **bgr**, **rc**, and **wc**, alter your assignment code to **drop** all the nominal features listed above. Be sure you select the right *axis* for columns and not rows, otherwise Pandas will complain!

2. Right after you print out your dataset's dtypes, add an exit() so you can inspect the results. Does everything look like it should / properly numeric? If not, make code changes to coerce the remaining column(s).

Dive Deeper

▸ 5. Data Modeling

3. Run your assignment and then answer the questions below.
4. Alter your code so that you only drop the **id** and **classification** columns. For the remaining 10 nominal features, properly encode them by as explained in the Feature Representation section by creating new, boolean columns using Pandas `.get_dummies()`. You should be able to carry that out with a single line of code. Run your assignment again and see if your results have changed at all.

Important Notes:

Once you've completed this lab, be sure to drop by the Dive Deeper section and read the article on using PCA on boolean features! Also, it's important to keep in mind that PCA is an unsupervised learning technique. It neither knows or even cares about your data's labels and classifications. In the previous two labs, you used a pre-labeled dataset only to see how applying PCA transformations can effect other machine learning modeling process further down the analysis pipeline.

Lab Questions

(2/2 points)

After adding in all of numeric columns, do the green, non-chronic kidney disease patients group closer together than before?

Yes ▾ ✓ **Answer:** Yes

After converting the nominal features to boolean features, do the green, non-chronic kidney disease patients group even closer together than before?

Yes ▾ ✓ **Answer:** Yes

EXPLANATION

After adding in all the other features and properly encoding those that need to be, you should notice that your separation between CKD and Non-CKD patients increases significantly. Recall what was mentioned in the Determining Features > Feature Importance section about the purse strap length feature: by taking many secondary features and combining them, machine learning is usually able to come up with more informative descriptions of your data.

You have used 1 of 2 submissions