Day_3_Links

1. What is dimensionality reduction? https://youtu.be/jPmV3j1dAv4
2. PCA: https://youtu.be/HMOI_lkzW08
3. PCA and SVD explained: pca and svd with numpy
4. How does PCA really work: https://youtu.be/FgakZw6K1QQ
5. Read *Machine Learning Pocket Reference* Ch 17 (mlpr ch17 code)
6. Read *Hands on Machine Learning* Ch 20 (homl Ch 20 code)
7. Feature Selection: explanation of feature selection
8. Read *Feature Engineering and Selection* Ch 10 (feature selection)
9. Read *Machine Learning Pocket Reference* Ch 8 (mlpr Ch 8 code)
10. Read *Machine Learning Pocket Reference* Ch 9 (mlpr Ch 9 code)
11. Read *Feature Engineering and Selection* Ch 11 (forward and backward selection)
12. Variable selection in multiple regression
13. Collinearity: detecting collinearity
14. Variance inflation factors: understanding VIF

Discussion 1:

Question:

You are given a data set and asked you to perform Principal Component Analysis on it. The data set consists of 100K observations and 31 attributes, 15 of which are continuous/floating point numeric values. Some of the attributes are highly correlated and the variances of the numeric attributes are of widely disparate magnitudes.

Describe in the forum the data preparation techniques you would recommend prior to performing PCA.

Be sure to explain why your approach would be effective and why it would be preferable to other possible approaches.

Solution:

- For highly correlated and continuous attributes, first analyze their correlation types. If the data set has completely positive or negative correlation attributes,

then the performance of the model is likely to be affected by *multicollinearity*. The easiest way is to remove these attributes that are completely relevant.
- Very low variance features may also not be useful in understanding the data.
- Dimensional reduction algorithms such as kernel sparse representation are another approach. This first maps the nonlinear samples to the kernel space to alleviate the nonlinear similarity problem of the attributes. Then it takes feature data of the kernel space and restructures it to gain a sparse representation of the original dataset.
- Another option is to use feature selection to select the optimal attribute subsets for experiments.
- For numeric attributes with widely disparate variances magnitudes, we could perform feature scaling such as normalization. This method can eliminate the dimensional influence between attributes, that is, eliminate the influence of the order of magnitude.
- Split up categorical and numerical data. PCA is used for numerical data, not categorical data.

The idea of PCA is to select the axis that preserves the maximum amount of variance, as it will most likely lose less information than the other projections. That is to say, it minimizes the mean squared distance between the original dataset and its projection onto that axis. Next we choose the nth dimensional orthogonal axis that accounts for the largest amount of remaining variance. And we continue to do this for each subsequent dimension. The idea being that it is easy to identify where the amount of information cuts off using a scree chart, or some practical estimation thereof such as computing which number of the d-dimensions is required to ensure that 95% of the data is explained.

There are multiple approaches to this, one of which is very related to the SVD (singular value decomposition) - since they are both based on a very similar matrix decomposition. here is a video I created explaining what an SVD is and how to do it (using matlab):

https://www.youtube.com/watch?v=Id1TPKdrCuw

It is important to remember that in SVD and PCA you must center your data.

We know that you should not use categorical data for PCA. The question states that 15 of the 31 attributes are continuous/floating point numeric values. I am not sure if we are to assume the others are categorical or not; if so, then one should not use PCA and instead use other methods that deal with this better such as multiple correspondence analysis . If the other data points are numeric values that are finite, this would be OK to use PCA, but we must make sure to consistently scale the data. Overall, it is not suggested (in the

literature) to use PCA on binary variables, although there is some research which suggests it can be done, however, it is less meaningful than with continuous variables. http://www.macs.hw.ac.uk/bmvc2006/papers/174.pdf

Discussion 2:

Question:

Construct a regression model using correlation thresholds, variance thresholds, stepwise search methods, and VIFs. Explain why each of the feature selection methods listed are useful.

Solution:

Correlation Thresholds

If there are high correlations between variables, then there is redundancy within the data. This adds more features to the model, making it more complex, and harder to interpret and show that each variable is significant. Some threshold should be chosen to remove a feature if it is highly correlated with multiple other features. This can be done by looking at a heatmap of correlations and scatter plots.

This method will be needed in order to ensure that the model is not overfit. One should remove variables that are highly correlated. For instance we can say that we consider that the correlation is high if it is above 0.5.

Variance Thresholds

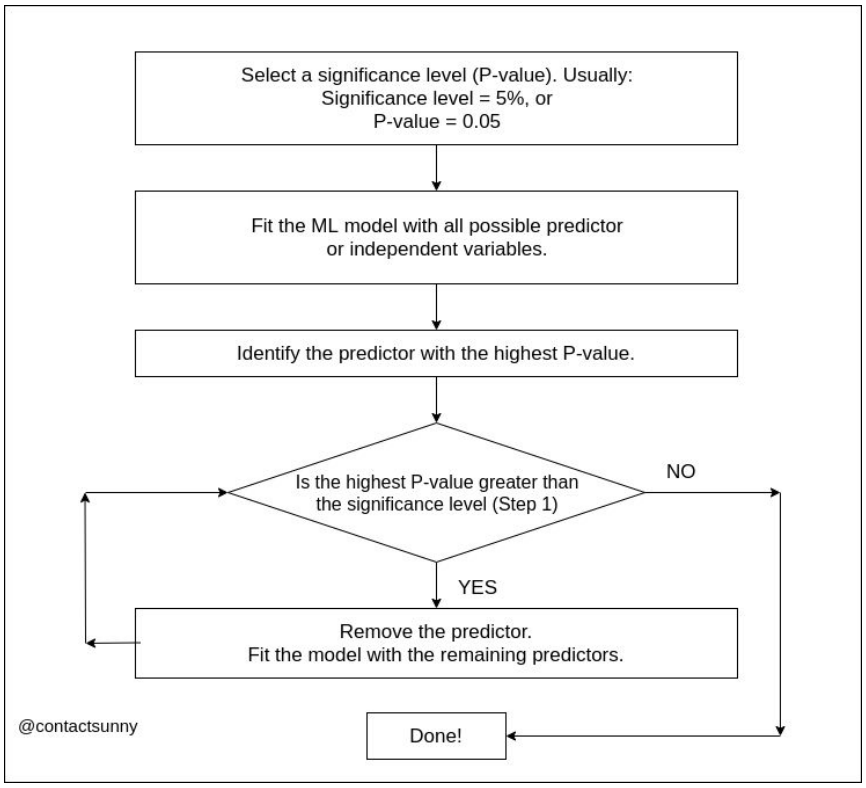
If a number has a very low variance, the feature is nearly constant and won't be contributing much to the model. In this case, this variable should be eliminated as before to make the model less complex.

Stepwise Search Methods

- Backward method: Start the test with all available predictor variables, then starting with the highest *p-value* variable. Delete as you go to produce the

regression model. Continue this process until only terms with *p*-values below a chosen threshold remain(*p*-value thresholds usually could be 0.05 or 0.10).

- Forward method: Start the test with a model containing only the intercept. Then slowly add variables to the model. The first predictor should be the lowest *p*-value. This continues until all the remaining terms that are not included in the model are above a specified *p*-value threshold.



VIF

The Variance Inflation Factor (VIF) helps to find multicollinearity/correlation between the explanatory variables, by taking one and regressing it every other predictor in the model. The result will show what percentage the variance/behavior of a predictor is inflated by the interaction with others. In this context the VIF can help in reducing the number of unnecessary variables, hence the model complexity making it easier to understand.

VIF estimates how much each explanatory variable is explained by the others calculating regression models and calculating the R-squared. The VIF is 1/(1-R-squared). If the VIF is 1, it means that the variable is independent of the others entirely. There are different

rules to evaluate which is the right threshold for the VIF coefficient. Below 4 the variable is not in a multicollinearity relationship with the others.

This can be used as a more objective measure of how collinear features are. A rule of thumb is if a vif is greater than 5, there is high collinearity. As before, a high vif may indicate that the feature should be dropped depending on things like if you are including polynomial variables.