Week 7: Feature Selection and Dimensionality Reduction

Introduction

Contents of the Week

- Introduction
- Readings
- Supplemental Materials
- 7.1 Discussion/Participation
- 7.2 Exercise: Dimensionality Reduction and Feature Selection

Topics

- A Feature extraction
- B Feature selection
- C Dimensionality reduction
- D Principal component analysis
- E Data snooping/leakage

Readings

Read the following:

- Chapters 9-10 of *Machine Learning with Python Cookbook*

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings | **Videos**

- How to Choose a Feature Selection Method For Machine Learning
- Feature Selection Techniques in Machine Learning
- A One-Stop Shop for Principal Component Analysis
- Chi-Square Test for Feature Selection in Machine learning
- ANOVA for Feature Selection in Machine Learning
- Data Leakage in Machine Learning
- Data Snooping

7.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is the difference between feature extraction and feature selection?
2. What is the difference between supervised and unsupervised feature selection?
3. What is the purpose of dimensionality reduction?
4. How should you expect dimensionality reduction to affect your model results?
5. What is principal component analysis (PCA)?

6. What is meant by a kernel PCA? What happens if you use a linear kernel?
7. What are some methods for feature selection?
8. How does the type of data affect which feature selection method you choose?
9. What is recursive feature elimination?
10. What is data snooping/leakage? What are some subtle ways that these can occur?

7.2 Exercise: Dimensionality Reduction and Feature Selection

Download the labeled training dataset from this link: House Prices - Advanced Regression Techniques.

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

Submit your code, output, and answers at the link above. Comment all your code and answer any questions that are asked in the instructions. It is perfectly fine to answer a question by displaying output from your code, but make sure you are displaying the appropriate output to answer the question. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Exercise Rubric

Part 1: PCA and Variance Threshold in a Linear Regression

1. Import the housing data as a data frame and ensure that the data is loaded properly.
2. Drop the "Id" column and any features that are missing more than 40% of their values.
3. For numerical columns, fill in any missing data with the median value.
4. For categorical columns, fill in any missing data with the most common value (mode).
5. Convert the categorical columns to dummy variables.
6. Split the data into a training and test set,

where the **SalePrice** column is the target.

7. Run a linear regression and report the R^2-value and RMSE on the test set.
8. Fit and transform the training features with a PCA so that 90% of the variance is retained (see section 9.1 in the *Machine Learning with Python Cookbook*).
9. How many features are in the PCA-transformed matrix?
10. Transform but **DO NOT** fit the test features with the same PCA.
11. Repeat step 7 with your PCA transformed data.
12. Take your original training features (from step 6) and apply a min-max scaler to them.
13. Find the min-max scaled features in your training set that have a variance above 0.1 (see Section 10.1 in the *Machine Learning with Python Cookbook*).
14. Transform but **DO NOT** fit the test features with the same steps applied in steps 11 and 12.
15. Repeat step 7 with the high variance data.
16. Summarize your findings.

Part 2: Categorical Feature Selection

Download the data from this link Mushroom Classification. Based on several categorical features, you will predict whether or not a mushroom is edible or poisonous.

1. Import the data as a data frame and ensure it is loaded correctly.
2. Convert the categorical features (all of them) to dummy variables.
3. Split the data into a training and test set.
4. Fit a decision tree classifier on the training set.
5. Report the accuracy and create a confusion matrix for the model prediction on the test set.
6. Create a visualization of the decision tree.
7. Use a χ^2-statistic selector to pick the five best features for this data (see section 10.4 of the *Machine Learning with Python Cookbook*).
8. Which five features were selected in step 7? Hint: Use the **get_support** function.
9. Repeat steps 4 and 5 with the five best features selected in step 7.
10. Summarize your findings.