edX

Feature Selection
Feature Selection

Feature selection can be an important part of model selection. In supervised learning, including features in a model which do not provide information on the label is useless at best and may prevent generalization at worst.

Feature selection can involve application of several methods. Two important methods include:

1. Eliminating features with **low variance** and **zero variance**. Zero variance features are comprised of the same values. Low variance features arise from features with most values the same and with few unique values. One way low variance features can arise, is from dummy variables for categories with very few members. The dummy variable will be mostly 0s with very few 1s.

2. Training machine learning models with features that are **uninformative** can create a variety of problems. An uniformative feature does not significantly improve model performance. In many cases, the noise in the uniformative features will increase the variance of the model predictions. In other words, uniformative models are likely to reduce the ability of the machine learning model to generalize.

By the completion of this lab, you will:

1. Eliminate low variance features, which by their nature cannot be highly informative since they contain a high fraction of the same value.

2. Use recursive feature elimination, a cross validation technique for identifying uninformative features.

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.

2. Now, run jupyter notebook and open the "FeatureSelection.ipynb" notebook under Module 5 folder.

3. Examine the notebook and answer the questions along the way.

Question 1

1/1 point (graded)

How many features are deemed low variance?

- ◉ 12
  ✔
- ○ 17
- ○ 18
- ○ 37

Submit   You have used 2 of 2 attempts

✔ Correct (1/1 point)

Question 2

1.0/1.0 point (graded)

How many features are deemed un-important?

- ◉ 12
  ✔
- ○ 17
- ○ 18
- ○ 37

Submit   You have used 2 of 2 attempts