Machine Learning – COMP3032

Tutorial and Lab Practice 4 – Week 5

This lab pactice focuses on the concepts and techniques of model seleection.

Tutorial

1. Review the terminology and concepts introduced and algorithems taught in Lecture 4.
2. Suppose you are using Polynomial Regression. You plot the learning curves and you notice that there is a large gap between the training error and the validation error. What is happening? What are three ways to solve this?
3. Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. What is happening? Should you increase the regularization hyperparameter α or reduce it?
4. Why would you want to use Ridge Regression instead of plain Linear Regression (i.e., without any regularization)? Or Lasso instead of Ridge Regression?
5. What are the learning curves? What are they used for?

Lab Practice

1. Download, open and run the program tut04.py. Read and understand the program.
2. Revise tut04.py:

   1) Perform cross-validation for the polynomial regression model, use the *display_scores* function to print the scores, mean and the standard deviation.
   2) Perform cross-validation for the linear regression model, use the *display_scores* function to print the scores, mean and the standard deviation.
   3) Perform cross-validation for the ridge regression model, use the *display_scores* function to print the scores, mean and the standard deviation.
   4) Perform cross-validation for the lasso regression model, use the *display_scores* function to print the scores, mean and the standard deviation.
   5) Create a polynomial regression model of degree 2 From X and y. Perform cross-validation for it, use the *display_scores* function to print the scores, mean and the standard deviation.
   6) Which model would you choose based on those cross-validation scores?

3. The file bloodpressure.csv contains blood data set. Build ridge regression models of multiple linear regression to predict systolic pressure using the other features. You can use different α values (e.g. from 0 to 4 at 0.1 interval) to build the models. Use cross-validation to select the best α value. (Hint: The data has categorical features, so these need to be converted into dummy features before you can use them for regression, using function pandas.get_dummies(). You also need to drop column 0 (ID) and column 7 (systolic) when fitting the model, and use column 7 (systolic) as the lables.)