LINEAR REGRESSION IN PYSPARK: POWER GENERATION

INTRODUCTION

In this exercise, you will use linear regression to understand the relationships between the different amounts of advertising spending and the resulting sales.

- Linear Regression
- Interaction
- Confounding

DATA USED IN THIS EXERCISE:

We will use the data from a Combined Cycle Power Plant over six years (2006-2011). We wish to predict the net hourly electrical energy output (PE) as a function of four explanatory variables: hourly average ambient temperature (T), pressure (AP), relative humidity (RH) and exhaust vacuum (V). The dataset is `power_plant.csv`.

QUESTIONS

1) Import and describe the dataset univariately (what is the mean, median, min, max, etc for each variable).
2) We will focus on predicting power output (PE). What is your best estimate of PE in the absence of any other information (just the power column)?
   a. Calculate the sample mean
   b. Write out a regression equation for the null (intercept-only) model you just fit and plug in the parameter estimate from the model
3) Now consider the pair-wise relationships between each individual predictor (T, AP, RH and V) and the corresponding outcome of PE.
   a. For each predictor, calculate the Pearson correlation coefficient between it and the target variable, PE. Does any variable appear significant?
   b. For each predictor, find the corresponding p-value.
4) We may also need to understand if our predictors are related to each other.

   a. For each possibly pairwise combination of the predictors, create a scatterplot and find the correlation between the predictors. Are any of them related to each other?

5) For the final step in this modeling process, we will fit a multivariate linear regression using (potentially) all of the predictors.
   a. For each predictor, consider if you need a non-linear fit. You may 'cheat' here, and produce a scatterplot matrix in seaborn (sns.pairplot), as the data will fit in memory. In practice this will be harder for big data sets.
   b. Using the Pipeline constructor, construct a final ‘best’ model using the linear (or non-linear) terms you decided on above, setting aside some data for validation purposes.
   c. Use this final model to summarize the relationships between the predictors and the outcome, including validation of the model. Summarize these results verbally.
6) Perform the above analysis, but using a Generalized Linear Regression model.
7) Perform a single validation split over a range of penalty values to determine the appropriate range of hyperparameters for a penalized linear regression. How does this compare to the ‘vanilla’ linear regression?