Assignment 1

Linear Regression, Feature Engineering, and Regularization

Instructions. This assignment focuses on predicting monthly sales for different products at the retailing giant H&M. In the dataset, each row is a "product-month" of sales at a single large H&M location. Our approach will be to trained machine learning models using sales data for a subset of products, and then use the trained model to predict sales for a set of out-of-sample test products.

Explore the Week 1 web app and answer the questions below. Stick to 3 sentences maximum per question. Upload your responses to BruinLearn as a PDF by the end of the day (March 30th, 5:00pm PT).

Question 1. Build an OLS regression model (no regularization) to predict sales for the "Hoodie" product category. Use only "Price" as a feature. What mean absolute error (MAE) do you observe on the test set? (Note: The units for predictions and errors are both "items sold per month".)

Question 2. Staying with the "Hoodie" category, add a handful of color or pattern indicator variables of your choosing to your Price-only model. Does the training MAE improve? Does the test MAE improve? What does this tell you about the informativeness of these product-attribute variables for predicting sales on this particular dataset?

Question 3. Now add “Lag 1” (last month’s sales) to your model from Question 2. What happens to the test MAE? If you observe a change in the test MAE, briefly explain why you think this occurs.

Question 4. Pick a regularization scheme (Lasso, Ridge, or ElasticNet) and re-train your model for “Hoodie” using **all** available variables. Use cross-validation (CV) to tune the regularization parameter(s). Report your CV MAE, test MAE, and the number of non-zero coefficients chosen by the model by reading them off the plots. How does test MAE performance compare to your model from Question 3?

Question 5. Suppose H&M is planning to launch a brand-new item in the “Hoodie” product category. Maximizing revenue for the new item requires accurate sales predictions, so H&M can make strategic inventory decisions accordingly. What is a potential limitation of the machine learning framework considered in this assignment for predicting sales in the first month for the new item? What other feature data might you want to collect to improve the model’s predictive accuracy?