SI 670 Problem Set 2

Q1 (15 points). Suppose that you are working as a data scientist at the sales department for an activewear clothing company, say like Lululemon or Nike, and you would like to train a machine learning model using sales records from last year to forecast sales for this year.

(1) Is your model a classifier or a regressor? Why?

It is a regressor. Because we want to predict the number of sales of this year but not classify something into different categories.

(2) How would your training data be similar to testing data? List three ways you expect your training data to be similar to testing data, and discuss the reasons.

Similar features: The training and test dataset should contain same predictors or features so that models can make prediction or test accuracy.

Similar distribution in features: We would expect the distributions of the features to be similar to ensure our prediction model to be consistent. Otherwise we need to do some feature scaling or manipulation

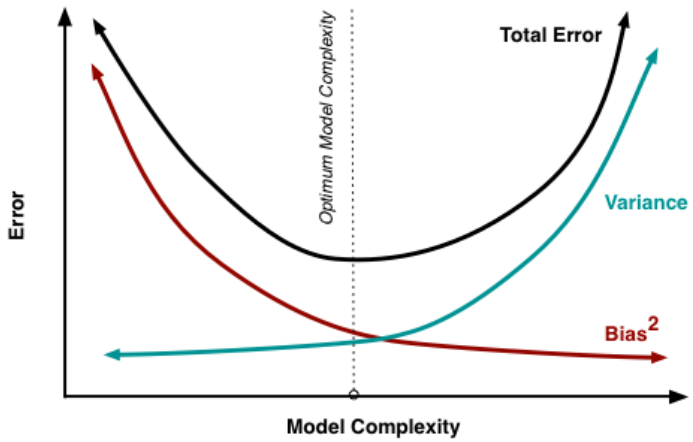
Similar DGP: We would expect the underlying data generating process to remain similar so that we can use the same set of features from last year to predict this year's sales. Otherwise, we would expect larger prediction error.

(3) How would your training data be different from your testing data? List three ways you expect your training data to be different from testing data, and discuss the reasons.

Different representativeness: training data and testing data may be different samples of the population.

Different economic situation from last years.
Different marketing strategies used by the team from last year's.

Q2 (16 points). We learned about Ridge vs Lasso regularizations in class. Recall that "alpha" (α) is a hyperparameter that determines the impact of the regularization term. The following figure illustrates the bias-variance tradeoff. Use it to answer the first two questions.



(1) As model complexity increases, what happens to the bias and variance of the model?
As complexity increases, bias decreases and variance increases. Because complex model can capture more patterns in the data, the prediction error will be smaller. But the model is more sensitive to the sample data, so the variance is larger.

(2) In ridge regression, what happens if we set α = 0? What happens as α approaches ∞?
If α = 0, there will be no regularization. The model will be equivalent to the linear regression model. If alpha term approaches infinity, the coefficients will approach to zero.

(3) If we have a large number of features (10,000 +) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in identifying useful features?
Lasso regression would be more helpful. Because we only want to keep the important features and minimize the coefficients of other variables to zero.

(4) What are the benefits of using Ridge regression compared to standard linear regression (minimizing RSS)?
Ridge regression includes regularization term can help control the overfitting problem which gives us smaller RSS and better prediction in test data.

Q3 (12 points). We learned about k-NN regression and linear regression in class.

(1) Can you think about a real-life situation where k-NN regression would work better than linear regression? Describe the situation and explain why k-NN regression is better.

Predicting housing prices. The pattern of house prices is more likely to be non-linear. Also, there are more local patterns in the house prices.

(2) Can you think about a real-life situation where linear regression would work better than k-NN regression? Describe the situation and explain why linear regression is better.

Predicting overall GPA using weekly assignment grades. The relationship between assignment grades, course grades, and GPA is mostly linear.

(3) Summarize what are the advantages and disadvantages of k-NN/linear regression, based on your examples above.

k-NN: advantages: better at capturing local and non-linear patterns, better with data including multiple similar data points (similar houses or houses in the neighboring region), flexible (without linearity assumption)

disadvantages: worse in generalization (if the testing house is very different from the houses in training data, larger error), computationally expensive in high dimensional data, sensitive to the choice of number of neighbors (k)

Linear regression: better when relationship is linear (course final grades and weekly assignment grades), easier to interpret the relationships between features and outcomes, better to deal with high-dimensional large scale datasets.

Disadvantages: too simplifying the model leads to underfitting, missing local patterns (hard to capture the local pattern in house prices if no relevant predictors are included in the model)

Q4-Q7 in si670f25_hw2.ipynb