**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?
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.
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.
4. Regressor. The output is a continuous numerical value
5. 1. A similar group of target customers compared to last year’s;

2. Similar categories of products compared to last year’s;

3. Similar database structure and features used in the company compared to last year’s;

…

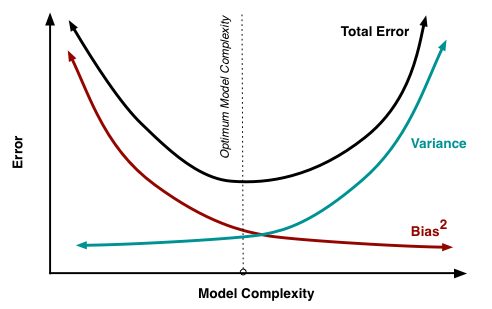
1. 1. The competitive environment in the market compared to last year’s;

2. Economic situation different from last year’s;

3. Different marketing strategies compared to 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?
2. In ridge regression, what happens if we set 𝛼 = 0? What happens as 𝛼 approaches ∞?
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?
4. What are the benefits of using Ridge regression compared to standard linear regression (minimizing RSS)?
5. Bias decreases, variance increases
6. When alpha = 0, we have OLS; alpha → inf, the penalty dominates and the coef becomes 0. So we have the simplest model where we predict a constant.
7. Lasso; bad features will be set to 0 so we have implicit feature selection
8. Ridge regularization prevents overfitting to our data, so it has smaller test error an also smaller variance compared to OLS.

**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.
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.
3. Summarize what are the advantages and disadvantages of k-NN/linear regression, based on your examples above.
4. Predict HOA fees of a community: could be more related to the surrounding areas (likely to be in the same homeowners’ association) instead of housing features (square footage, number of bedrooms and bathrooms, etc)
5. Predict how many miles a car can drive with one gallon of gas given the car’s weight. Heavier cars use more gas.
6. Advantages of k-NN Regression:

* Flexibility in capturing complex, non-linear relationships
* Adaptability to local variations in data

Disadvantages of k-NN Regression:

* Sensitivity to the choice of k
* May not perform well in high-dimensional spaces due to the curse of dimensionality

Advantages of Linear Regression:

* Easy to interpret the relationship between features
* Can handle higher dimension data
* Captures linear relationships well

Disadvantages of Linear Regression:

* Limited ability to capture complex, non-linear relationships
* Inflexibility when dealing with local variations in data

**Q4-Q7 in si670f25\_hw2.ipynb**