CS249 Fall 2020
Problem Set 3: Prediction

Christopher Munoz Cortes

November 22, 2020

1 Ridge Regression

(a) Write the objective function of a ridge regression in the matrix format.

$$\text{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta$$

(b) Compute the gradient of the objective function with respect to the vector of the coefficient in the model.

Simplifying the objective function before taking the gradient yields:

$$\begin{aligned}\text{RSS}(\beta) &= (Y^T - \beta^T X^T)(Y - X\beta) + \lambda\beta^T\beta \\ &= Y^TY - \beta^TX^TY - Y^TX\beta - \beta^TX^TX\beta + \lambda\beta^T\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^T(X^TX)\beta + \lambda\beta^T\beta\end{aligned}$$

Now, taking the gradient with respect to β gives:

$$\nabla_\beta \, \text{RSS}(\beta) = -2X^TY + 2X^TX\beta + 2\lambda\beta$$

(c) Show that the solution to the ridge regression can be written in the following form:

$$\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

where X is the design matrix, Y is the vector of the outcomes, and λ is the regularization parameter.

Setting the expression from part (b) equal to zero and solving for β:

$$\begin{gathered}\nabla_\beta \, \text{RSS}(\beta) = 0 \\ -2X^TY + 2X^TX\beta + 2\lambda\beta = 0 \\ (X^TX + \lambda I)\beta = (Y^TX)^T \\ \boxed{\hat{\beta} = (X^TX + \lambda I)^{-1}\, X^TY}\end{gathered}$$

2 Regression Models

(a) In many settings, our data contains variables with missing values. Search online and find suitable ways to impute missing values in a dataset. Use one of these methods to treat such variables in the data.

Some imputation methods are:

- *Mean, median, mode imputation.* Replace each missing value with the mean, or the median, or the mode (most frequent value) of the observed values for that feature.
- *Regression imputation.* If we know there's a correlation between the missing value and other features, it is possible to get better estimates by running a linear regression for the missing feature values on other features.
- *K-nearest Neighbor Imputation.* Neighbor-based imputation replaces the missing values with the most frequent value among the k-nearest neighbor for categorical data, and mean or mode for continuous variables.
- *Multiple Imputation.* The single imputation methods above are limited in that they do not reflect the same variability from the sample data and the missing values. Multiple imputation methods create several imputed values for each missing value. Since each value is predicted from a slightly different model, it reflects sampling variability. One such method is MICE (Multivariate Imputation by Chained Equation).

For instance, if we were to take all the numerical features in our data and fill in missing values, we could use `scikit-learn`'s `IterativeImputer`. This implementation models each feature with missing values as a function of other features in a round-robin fashion.

```python
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer

# Trainset is a dataframe with the data from ‘train.csv‘
X = trainset.select_dtypes([np.number])
feat_names = X.columns
imp = IterativeImputer(max_iter=10, random_state=0, n_nearest_features=4)
X = pd.DataFrame(data=imp.fit_transform(X), columns=feat_names)
```

Listing 1: Imputing missing values using `IterativeImputer`

(b) Name two categorical features in the data, and choose ten features from the dataset that you think will be most predictive of the outcome

- Two categorical features: `OverallCond`, `BldgType`
- Ten best predictors based on correlation: `SalePrice`, `OverallQual`, `GrLivArea`, `GarageCars`, `GarageArea`, `TotalBsmtSF`, `1stFlrSF`, `FullBath`, `TotRmsAbvGrd`, `YearBuilt`, `YearRemodAdd`

(c) Transform the categorical features in your chosen set to make them suitable for modeling

See code below.

(d) Apply ridge regression to the data and find the best value of the regularization parameter using cross-validation. Report the RMSE of your best model both in training and validation sets. Do you see any overfitting in your model?

3 Feature Selection

4 Predicting The Election Outcome

Listing 2: Code for Question 1 *Hypothesis Testing*

5 Regression to the Mean