STAT GR 5241 Statistical Machine Learning
Homework 1
Yonathan Amare
February 21, 2025

Question 1: Data Analysis (60 Points)
Part (a): Important Features Analysis (30 points)

(i) Feature Importance Comparison

Different methods identified varying sets of important features:

1. Least Squares Statistical Significance:
   - PctHousLess3BR (coefficient = 0.892)
   - PctVacMore6Mos (coefficient = 0.726)
   - blackPerCap (coefficient = 0.368)
   - pctUrban (coefficient = 0.359)
   - PersPerOwnOccHous (coefficient = 0.318)

2. Recursive Feature Elimination:
   Selected the following top features:
   - population
   - racepctblack
   - pctUrban
   - whitePerCap
   - PctBSorMore
   - PctUnemployed
   - PctWorkMomYoungKids
   - MedNumBR
   - PctHousOwnOcc
   - NumStreet

3. Lasso:
   Most significant features:
   - PctImmigRecent (0.203)
   - racepctblack (0.171)
   - FemalePctDiv (0.154)
   - NumStreet (0.137)
   - PctWorkMomYoungKids (0.113)

4. Elastic Net:
   Selected features similar to Lasso:
   - PctImmigRecent (0.190)
   - racepctblack (0.159)
   - FemalePctDiv (0.143)
   - NumStreet (0.127)
   - PctWorkMomYoungKids (0.104)

(ii) Regularization Path Analysis

I visualized regularization paths for:
- Lasso: Shows sharp transitions to zero as α increases
- Elastic Net with α = 0.3 and α = 0.7: Shows smoother transitions than Lasso
- Ridge: Shows gradual coefficient shrinkage with no exact zeros

(iii) Reflection on Results

1. Different methods selected somewhat different features because:
   - Least squares considers individual significance without accounting for correlations
   - Lasso tends to select one from correlated groups
   - Ridge keeps all features but shrinks coefficients
   - RFE looks at recursive importance

2. Tuning parameters were selected through:
   - Cross-validation for Lasso and Elastic Net
   - Grid search for optimal α values
   - Different α values did affect feature selection, particularly in Elastic Net

3. Consistently important features across methods:
   - racepctblack
   - PctWorkMomYoungKids
   - Housing-related features (PctHousLess3BR, PctHousOwnOcc)

   These features likely have strong, stable relationships with crime rates.

Part (b): Prediction Performance (30 points)

(i) MSE Comparison

After running 10 trials with 60/20/20 splits, the average MSE results were:

1. Ridge Regression: 0.0184 (±0.0019)
2. Least Squares: 0.0186 (±0.0018)
3. Lasso: 0.0186 (±0.0019)
4. Elastic Net: 0.0186 (±0.0019)
5. RFE: 0.0203 (±0.0022)
6. Best Subsets: 0.0212 (±0.0019)

(ii) Results Visualization
Created bar plot showing mean MSE with error bars for each method.

(iii) Reflection

1. Ridge regression performed best, suggesting:
   - Many features contribute useful information
   - Multicollinearity exists in the dataset
   - Complete feature elimination may be too aggressive

2. The similar performance of Lasso and Elastic Net indicates:
   - Feature selection alone doesn't improve predictions
   - Most features contain some predictive value

3. RFE and Best Subsets performed worse, suggesting:
   - Too aggressive feature selection loses information
   - Complex relationships between features matter

Question 2: Regression Properties (50 points)

Part (a): Intercept Term Equivalence (10 points)

Demonstrated that the following approaches are equivalent:

1. With intercept term: $y = \beta_0 + X\beta + \varepsilon$
2. Centered data: $(y - \bar{y}) = (X - \bar{X})\beta + \varepsilon$
3. Column of ones: $y = [1|X]\gamma + \varepsilon$

Using simulated data with $n=100$, $p=3$, the maximum difference between predictions was $< 10^{-12}$, confirming numerical equivalence.

Part (b): Zero Training Error (10 points)

For p > n case, demonstrated using:
- n = 50 observations
- p = 100 features
- Random X and y

Achieved training MSE ≈ 10^{-15}, confirming zero training error due to perfect interpolation in overparameterized regime.

Yes, let's revise Part (c) to match our new empirical results. Here's how we should write it:

Part (c): Correlated Features Properties (30 points)

(i) Least Squares Variance
Using highly correlated features (ρ = 0.99), bootstrap analysis with 1000 resamples demonstrated the instability of least squares estimates through their coefficients of variation:
- CV($\hat{\beta}_1$) ≈ 0.28
- CV($\hat{\beta}_2$) ≈ 0.65

These results show substantial instability in coefficient estimates, particularly for the second feature, demonstrating how ordinary least squares become unreliable with highly correlated features.

(ii) Ridge Regression Grouping
For correlated features with true coefficients β = [1, 2], our ridge regression analysis showed how coefficients converge as regularization strength increases. We examined this using multiple α values and demonstrated the grouping effect through regularization paths.

(iii) Lasso Selection
Using increasingly strong regularization (α from 0.5 to 2.5), we observed Lasso's variable selection property:
- At α = 0.5: Both features retained [1.74, 0.67]
- At α = 1.0: Second coefficient shrinking [1.48, 0.42]
- At α = 1.5: Strong selection pressure [1.23, 0.17]
- At α = 2.5: Complete selection of one feature [0.40, 0.00]

This progression demonstrates how Lasso tends to select one feature from a correlated pair as the regularization strength increases, while completely eliminating the other feature's effect.