Modeling Wine Quality Using Classification and Regression

Mario Wijaya
MGT 8803
November 28, 2017
Motivation

Quality
- How to assess it?
- What makes a good quality wine?

Good or Bad Wine?
- Subjective?
- Wine taster

Who cares?
- Consumer
- Wine industry

Data Science
- Classification
- Regression

Goal
- Predict quality of a given wine
- Classify whether a wine is good or bad
Dataset

Consists of

- White wine: 4898 samples
- Red wine: 1599 samples
- Variables:
  - Fixed acidity
  - Volatile acidity
  - Quality
  - GoodBad
    - Quality > 5: Class 1
    - Quality <=5: Class 0
  - etc
- Potential problem?
  - Class imbalance
  - Bias
  - High variance

Solution

- Oversampling underrepresented class
- Downsampling overrepresented class
- Overweight underrepresented classes in loss function
- Normalization for classification and regression (SGD)

General Strategy

Train

Validation

Optimal Parameters

Test

Tune parameters

10-fold cross validation

Find Prediction Accuracy or $R^2$

Tools: Python3 with Scikit-learn package, Matplotlib & Seaborn (Plot & Visualization)
Models & Challenges

Regression
- Multi linear regression
- Stochastic Gradient Descent
- Ridge Regression
- Lasso Regression
- Decision Tree Regression

Classification
- SVM
- K-Nearest Neighbor
- Decision Tree Classification
- Used PCA to do dimension reduction
  - 11 variables mapped to 2 dimension

Challenges
- Find optimal parameters
  - SVM: C, gamma
  - Etc
- Find model that can be generalized
- Prevent overfitting
  - K-fold cross validation
Quick Lecture

**Stochastic Gradient Descent**
The standard gradient descent algorithm updates the parameters $\theta$ of the objective $J(\theta)$ as,

$$
\theta = \theta - \alpha \nabla_{\theta} E[J(\theta)]
$$

where the expectation in the above equation is approximated by evaluating the cost and gradient over the full training set. Stochastic Gradient Descent (SGD) simply does away with the expectation in the update and computes the gradient of the parameters using only a single or a few training examples. The new update is given by,

$$
\theta = \theta - \alpha \nabla_{\theta} J(\theta; x^{(i)}, y^{(i)})
$$

---

**Regression**
- Ridge Regression
  - L-2 penalty
- Lasso
  - L-1 Penalty
- Decision Tree

---

**Classification - KNN**
Regression

- Correlation Matrix
  - Look at possible high correlation feature
Regression

- Correlation Matrix
  - Look at possible high correlation feature
- Multiple Linear Regression
  - $Y = X_1\beta_1 + X_2\beta_2 + \ldots + X_n\beta_N + E$
  - $R^2 = 0.325$
  - Pretty bad!
- SGD - $R^2$: 0.323
- Lasso and Ridge equally bad
- Used interaction terms and remove high p-value -> bad
- Forward selection -> not good either

---

Regression Results

| Variable                  | Coef  | Std. Error | t-Stat | P>|t| | [0.025| 0.975] |
|---------------------------|-------|------------|--------|-----|--------|--------|
| Intercept                 | 49.9334 | 10.625 | 4.700 | 0.000 | 29.105 | 70.762 |
| fixed_acidity             | 0.0566 | 0.014 | 4.014 | 0.000 | 0.029 | 0.084 |
| volatile_acidity         | -1.2555 | 0.068 | -18.378 | 0.000 | -1.389 | -1.122 |
| citric_acid               | -0.0784 | 0.077 | -1.021 | 0.307 | -0.229 | 0.072 |
| residual_sugar           | 0.0376 | 0.005 | 7.623 | 0.000 | 0.028 | 0.047 |
| chlorides                | -1.1647 | 0.283 | -4.118 | 0.000 | -1.719 | -0.610 |
| free_sulfur_dioxide       | 0.0053 | 0.001 | 6.510 | 0.000 | 0.004 | 0.007 |
| total_sulfur_dioxide      | -0.0027 | 0.000 | -0.472 | 0.000 | -0.002 | -0.002 |
| density                   | -48.4174 | 10.867 | -4.456 | 0.000 | -69.720 | -27.115 |
| pH                        | 0.2880 | 0.091 | 3.151 | 0.002 | 0.109 | 0.467 |
| sulphates                 | 0.8255 | 0.068 | 12.181 | 0.000 | 0.693 | 0.958 |
| alcohol                   | 0.2607 | 0.015 | 17.448 | 0.000 | 0.231 | 0.290 |

---

Omnibus: 155.500
Prob(Omnibus): 0.000
Jarque-Bera (3B): 356.307
Skew: 4.159
Prob Skew: 0.25e-78
Kurtosis: -0.817
Cond. No.: 1.96e+05
Regression

- Correlation Matrix
  - Look at possible high correlation feature
- Multiple Linear Regression
  - $Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + E$
  - $R^2 = 0.325$
  - Pretty bad!
- SGD - $R^2: 0.323$
- Lasso and Ridge equally bad
- Used interaction terms and remove high p-value -> bad
- Forward selection -> not good either
Classification

- Normalize data (0,1)
- Varies parameter of C and gamma
  - 10-fold cross validation
  - Find best model that gives lowest error rate or highest accuracy rate
- ~83% prediction accuracy but clearly linear kernel is better in this case from support vector drawn
- How do you draw 11 dimensions into 2 dimensions?
  - PCA

Prediction Accuracy - RBF Kernel
Varying C and gamma

Prediction Accuracy - Linear Kernel
Varying C
**Classification - KNN**

- **Classification**
  - Ad-hoc knowledge:
    - $K = \frac{1}{\sqrt{\text{# of samples}}} \approx 99$
  - Use 10-fold CV
    - Determine error rate
    - Use it to find best $K$
    - $K = 40 \rightarrow K = 100$
      - Not much different
  - Higher $K$ -> smoother curves
  - Relatively good for classification
    - Easily overfitting
    - Careful!

---

**Error Prediction Rate with Variety of Ks**

- $\text{KNN Classification with $K=5$}
- $\text{KNN Classification with $K=99$}
Classification - Decision Tree

**Classification**

- Recursively find label
- Used Gini Index for splitting
  - Other methods: Information Gain (Entropy)
- 88% prediction accuracy
- Also tried with testing data
- Need to set depth, otherwise we will have overfitting
Conclusion & Discussion

**Conclusion**
- Several clustering algorithm works well with the dataset
- Bad performance with regression
  - Possibly need more work in determining which features to keep
- Combat subjective result from wine taster when we can use Data Science to answer the question

**Discussion**
- If good regression model can be found then a Python based application can be build for interactivity
- Need to understand dataset well and find optimal parameters
Modeling Wine Quality

★ Ran several algorithm on multiple linear regression
  ○ Ordinary Least Square (Linear Regression)
  ○ Ridge Regression
  ○ Lasso Regression
  ○ Stochastic Gradient Descent
  ○ Forward Selection
  ○ Decision Tree Regression

★ Created several classification models to predict whether the quality of a given wine is good or bad
  ○ K-Nearest Neighbors
  ○ SVM
  ○ Decision Tree Classification
  ○ Used PCA for dimensionality reduction