ML - OVERFITTING, SHRINKAGE, AND LINEAR MODELS

MGMT 675

AI-Assisted Financial Analysis

Kerry Back

RICE | BUSINESS
Jones Graduate School of Business

MACHINE LEARNING IN FINANCE

- Fraud detection
- Credit risk analysis
- Return prediction
- Valuation
- Text analysis
- Time series forecasting

REGRESSION VS CLASSIFICATION

- Regression means to predict a continuous variable (not necessarily linear regression).
- Classification is to predict a categorical variable. Binary or multiclass.

OVERFITTING AND SHRINKAGE

UNDERFITTING AND OVERFITTING

- Fitting or training a model means estimating its parameters (like linear regression).
- A model underfits the data on which it was trained if it makes poor predictions on that data.
- A model overfits the data on which it was trained if it makes good predictions on that data but performs poorly on new data.
- Overfitting means that the chosen parameters reflect chance relationships in the training data.

MODEL COMPLEXITY AND SHRINKAGE

- We can create more complex models by increasing the number of parameters (like adding variables in a linear regression).
- We can reduce complexity by reducing the number of parameters. Or by choosing less influential parameter values (like regression coefficients closer to zero).
- A more complex model is less likely to underfit but more likely to overfit.

TRAIN AND TEST

- To check that we have not overfit our model, we train and test.
- Split data randomly into train and test samples. Train (fit) on the training data. Test on test data.
  - Test data is also called holdout data.
  - Also called out-of-sample testing.
- Performance on the test data is the performance that we can expect on new data.

TEST CRITERIA

- How do we decide if performance is good or bad?
- For continuous variables,
  - usually want to achieve a low sum of squared errors
  - equivalently, achieve a high R^2.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- For categorical, can use % accurately classified.

VALIDATION

- We could consider different models - for example, models of different complexity - and train them all, test them all on the test data, and then choose the model that performs best on the test data.
- But then we run a risk of overfitting our test data. The performance may not generalize to new data.
- So, we can hold out some data within the training data (called validation data) for doing model comparison.

LASSO AND RIDGE EXAMPLES

DATA

- Download ml1.xlsx from the course website
- Upload it to Julius and ask Julius to read it and describe it.
- The data was created by generating 51 sets of 100 standard normals.
  - The first 50 sets are labeled x1, …, x50.
  - The last set was used as the noise to generate y1 as x1 + noise.
  - So, x2, …, x50 are irrelevant for y1, but they may be correlated with y1 by chance.

LINEAR REGRESSION

- Ask Julius to do a train-test split of the data with 20% of the data in the test set.
- Ask Julius to train a linear regressor on the training data with x1, …, x50 as the features and y1 as the target.
- Ask Julius to report the coefficient estimates.
- Ask Julius to compute the R-squared on the test data.

SHRINKAGE

- To induce selection of parameter values that are less influential, we can penalize large values.
- Usually in linear regression, we minimize SSE (sum of squared errors).
- LASSO: minimize SSE + penalty $\times$ sum of $|\beta_i|$.
- Ridge: minimize SSE + penalty $\times$ sum of β_i^2.

TRAIN AND TEST

- Ask Julius to train a ridge regressor on the training data, report the parameter estimates, and compute the R-squared on the test data.
- Ask Julius to traing a lasso regressor on the training data, report the parameter estimates, and compute the R-squared on the test data.
- The penalty in lasso and ridge is called alpha in scikit-learn. You can specify it.

OPTIMIZING THE PENALTY

- The penalty is called a hyperparameter, because it is not fit by training but instead is specified in advance.
- The penalty controls the degree of complexity. Larger penalty = less complex.
- To optimize the penalty, we can cross validate.

CROSS VALIDATION

- Split the training data into random subsets, say, A, B, C, D, and E.
- Use $A \cup B \cup C \cup D$ as training data and test on E.
- Then use $B \cup C \cup D \cup E$ as training data and test on A.
- Then, ..., until we have trained and tested 5 times.

- Average the 5 test scores.
- Repeat for each model configuration (each hyperparameter value).
- Choose the hyperparameter value with the highest average score.
- Then proceed to testing on the test data.

CROSS VALIDATE LASSO AND RIDGE

- Ask Julius to run GridSearchCV on lasso regression.
- You can specify a set of alpha values to try.
- Ask Julius what the optimal alpha (penalty) is.
- Ask Julius what the coefficient estimates are.
- Ask Julius what the score is on the test data.
- Repeat for ridge regression.