INFO 6105
Data Science Engineering Methods and Tools

Lecture 4
Cross-validation & Model Selection

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

Accuracy Metrics

Question How accurate our predictions are? It depends on

- what is being predicted,
- what accuracy measure is used, and
- what data set is used for computing the accuracy measure

Accuracy Metrics

Question How accurate our predictions are? It depends on

- what is being predicted,
- what accuracy measure is used, and
- what data set is used for computing the accuracy measure

Regression

- Mean Absolute Error (MAE)
- Root Mean Squared Error
- R^2, Adjusted R^2

Classification

- Accuracy, Balanced Acuracy
- FP and TP rates
- AUC

Accuracy Metrics

Accuracy metrics are used to evaluate model performance for

- Feature Selection
- Model Selection
- Tuning parameters

Accuracy Metrics

Accuracy metrics are used to evaluate model performance for

- Feature Selection
- Model Selection
- Tuning parameters

How to estimate accuracy metrics?

Validation Set Approach

It is a typical approach to randomly divide the dataset into two parts:

- Training Set: A model is built using this data
- Test Set (also called out-of-sample data or hold-out data)

Validation Set Approach

It is a typical approach to randomly divide the dataset into two parts:

- Training Set: A model is built using this data
- Test Set (also called out-of-sample data or hold-out data)

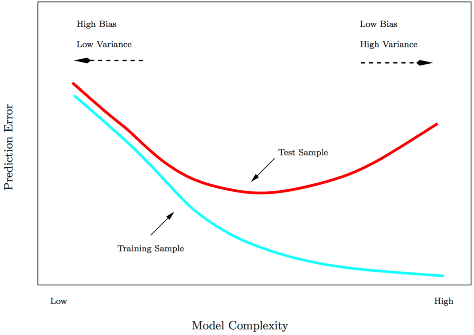
We measure the accuracy metrics on the test data to evaluate how well a model performs on new data and estimate the prediction error or accuracy.

Validation Set Approach

It is a typical approach to randomly divide the dataset into two parts:

- Training Set: A model is built using this data
- Test Set (also called out-of-sample data or hold-out data)

We measure the accuracy metrics on the test data to evaluate how well a model performs on new data and estimate the prediction error or accuracy.

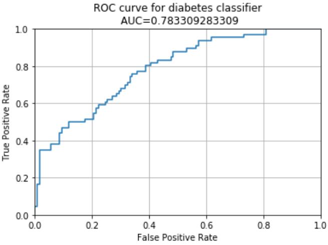


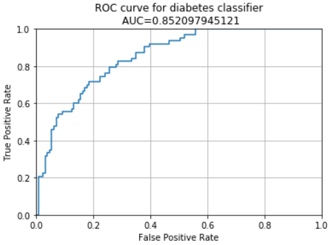
Validation Set Approach

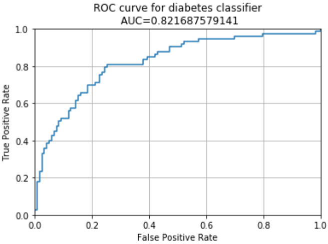
Two potential drawbacks

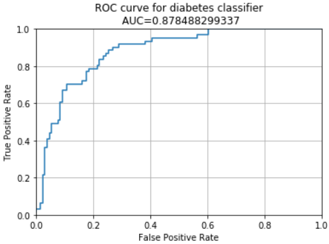
- The estimated error can be highly variable depending on which sample are included in the training and test sets
- Only a subset of the samples are used to fit the model.

Validation Set Approach









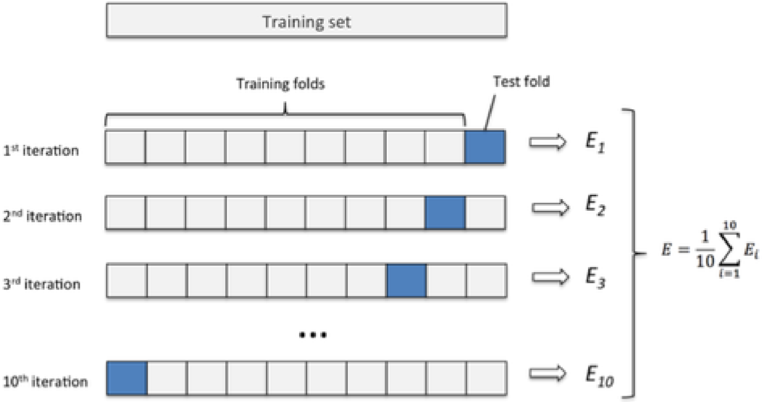
K-fold cross validation

Widely used for model selection and estimating the test error

- Randomly divide the data into K equal-sized parts
- For $k = 1, \ldots, K$, do
  - leave out part k
  - fit the model to the other $k - 1$ parts (combined)
  - calculate the test error E_k on the left-out k^{th} part
- Calculate the cross-validation error:

$$\mathrm{CV}_K = \frac{\sum_{k=1}^{K} E_k}{K}$$

K-fold Cross Validation

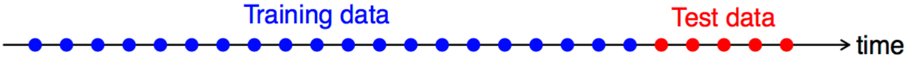


Time-series Forecasting

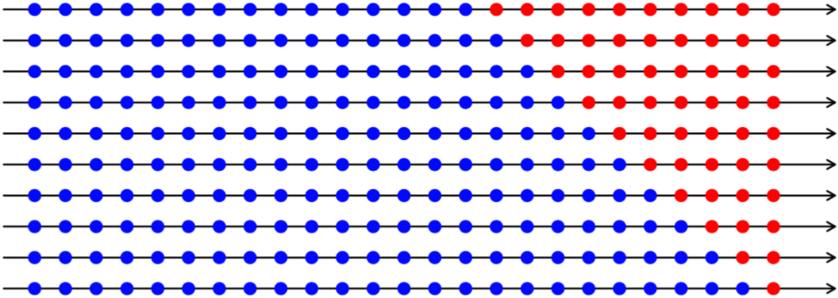
Training and Test Sets

Time-series Forecasting

Training and Test Sets



Cross-validation



Validation Set vs Cross Validation

- Computational time
- Better estimate for the test error

Feature Selection

There are three main approaches for excluding irrelevant features from a regression/classification model:

- Subset Selection: We identify a subset of features that we believe to be related to the response
- Regularization: We fit a model involving all features, but the estimated parameters are shrunken toward zero relative to the cost function.
- Dimension Reduction: We project the m features into a ℓ-dimensional space where $\ell < m$.

Subset Selection

- Best Subset Selection: Require to fit 2^m models. Not practical.
- Forward Stepwise Selection:
  - We begin with the *null* model with no features and then add features to the model one-at-a-time until all of the features are in the model.
- Backward Stepwise Selection
  - We begin with the full model containing all m features, and then iteratively remove the least useful feature, one-at-a-time.

Forward Stepwise Selection

1. Let M_0 denote the *null model*, which contains no features.
2. For $k = 0, 1, \ldots, m - 1$:
   - Consider all $m - k$ models that augment the features in M_k with one additional feature
   - Choose the best among these $p - k$ models, and call it M_{k+1}. Here best is defined as having highest R^2 for regression and highest AUC for classification.
3. Select a single best model from among $M_0, \ldots, M_m$ using cross-validation.

Backward Stepwise Selection (When $n > m$)

1. Let M_m denote the *full model*, which contains all features.
2. For $k = m, m-1 \ldots, 1$:
   - Consider all k models that contain all but one of the features in M_k , for a total of $k-1$ features.
   - Choose the best among these k models, and call it M_{k-1}. Here best is defined as having highest R^2 for regression and highest AUC for classification.
3. Select a single best model from among $M_0, \ldots, M_m$ using cross-validated.

Regularization or Shrinkage Methods

- The subset selection methods use accuracy metrics to fit a linear model that contains a subset of the features.
- As an alternative, we can fit a model containing all m features using a technique that constrains the model parameters and shrinks the them towards zero.

Ridge and Lasso regression

- Recall that the least squares fitting procedure estimates $\beta_0, \beta_1, \ldots, \beta_m$ using the values that minimize

$$\text{RSS} := \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} \beta_j x_{ij} \right)^2$$

- Ridge Regression: We estimate the model parameters to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{m} \beta_j^2$$

- Lasso Regression: We estimate the model parameters to minimize

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{m} |\beta_j|$$

Here $\lambda \geq 0$ is a tuning parameter that can be determined using cross-validation.