Assessment of Regression Models, Feature Selection, Model Selection

Dr. Mohamed Elshenawy
mmelshenawy@gmail.com

1

Previous session …

- What is Learning?
- Regression - OLS
- Model Capacity
- Overfitting and underfitting
- Regularization

2

This session…

- Assessment of Regression Models
  - Mean square error
  - R^2
  - Residual Plots
  - Generalization
- Bias and Variance
- Feature Selection - Regression
- Model selection

3

The session uses content from

- Book: An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0.

4

Which Metrics shall we use to assess regression models?

5

Assessing the Accuracy of the Model

- The quality of a linear regression fit is typically assessed using two related quantities:
  - Mean square error (MSE) / Residual Standard Error (RSE)
  - R^2 statistic.

6

Mean Square Error

- Assume that we have one input

$$y_i = f(x_i; \theta) = \theta_0 + \theta_1 x_i$$
$$t_i = y_i + \epsilon_i$$

- Mean square error

$$\bar{\epsilon} = \frac{\sum_{i=1}^{n}(t_i - y_i)}{n}$$

- An alternative to use the Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(t_i - y_i)^2}{n}} = \sqrt{\frac{RSS}{n}}$$

- RSS: Residual Sum of Squares (also known as Error Sum of Squares SSE)

7

Residual Standard Error (RSE)

$$RSE = \sqrt{\frac{RSS}{df}}$$

- The degrees of freedom *df* number of independent values that can vary in an analysis without breaking any constraints
- When we calculate RSE, df is equal to the sample size (n) minus the number of parameters we're trying to estimate.
- In case of 2 parameters (one input)

$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(t_i - y_i)^2}{n-2}}$$

8

Using Mean square error (MSE) / Residual Standard Error (RSE) for assessment

- RSE provides a good assessment but it is measured in the units of Y (depends on the scale of y)
- Think of a range of y [1- 10] and another scale [1- 10000]
- It is not always clear what constitutes a good RSE.

9

R^2 statistic

Define

TSS :total variance in the response t.

$$TSS = \sum(t_i - \bar{t})^2 \; where \; \bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}$$

We know that

$$RSS = \sum_{i=1}^{n}(t_i - y_i)^2$$

RSS measures the amount of variability that is left unexplained (error variability)

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

10

R^2 statistic (Cont)

- R^2 statistic: the *proportion of variability in y that can be explained using x*
  - If $R^2 = 1$ (perfect fit: all variability in y can be explained using x)
  - If $R^2 = 0$ (variability in y is not explained by

11

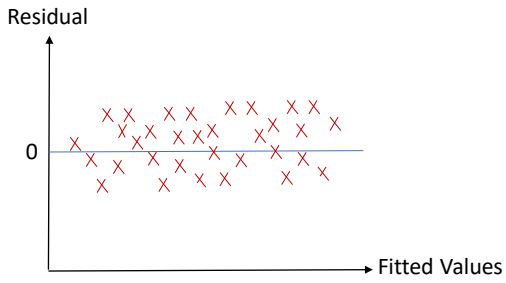
Adjusted R^2

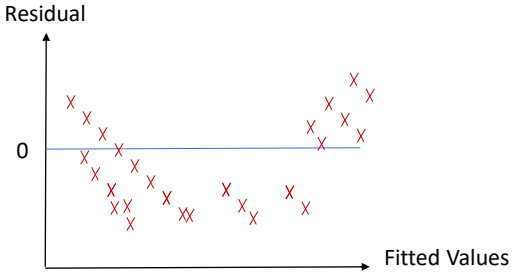
- Adjusted R-squared is a **modified version of R-squared** that has been adjusted for the number of predictors in the model.
- f you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.
- Adjusted $R^2 \leq R^2$

$$R^2\text{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - P - 1}$$

12

Residual Plots





- ϵ_i should be a normally distributed random variable with mean 0 and variance σ^2

13

Generalization

- A central challenge in machine learning is that we must perform well on new, unseen data—not just the training data. (this property separates machine learning from optimization)
- Typically, we split the dataset into two main subsets (the two sets are drawn
  from the distribution of inputs we expect the system to encounter in practice)
  - training set (used in the optimization procedure we mentioned earlier): produces a training error
  - test set: used to assess the generalization of the model (ability to perform well on unseen data): produces test error (also called generalization error)

14

Generalization

- Training error

$$MSE_{(train)} = \frac{\sum_{i=1}^{n_{(train)}} \left(t_{i(train)} - y_{i(train)}\right)^2}{n_{(train)}}$$

- Test error

$$MSE_{(test)} = \frac{\sum_{i=1}^{n_{(test)}} \left(t_{i(test)} - y_{i(test)}\right)^2}{n_{(test)}}$$

15

Generalization – Good Model

- A good model has a low training error and a low test error as well.

16

Important Assumptions about the Data Generating Process

- We typically make a set of assumptions known collectively as the i.i.d. assumptions about the data generating process (data collection process).
- These assumptions are that the examples in each dataset are independent from each other, and that the train set and test set are identically distributed, drawn from the same probability distribution p(x,y) as each other.
- i.i.d. assumptions allow us to mathematically study the relationship between training error and test error.

17

i.i.d assumption

- The same distribution is used to generate every train example and every test example.
- For some fixed value of model parameters θ, the expected training set error is exactly the same as the expected test set error, because both expectations are formed using the same dataset sampling process. (how about if they are not?)

18

Estimators, bias and variance

- In linear regression models, we try to estimate the 'best' vector of parameters (weights) that represent a relationship between the input and output variables.
- To distinguish the estimates of parameters from their true (and unknown) value, denote the estimate of a vector of parameter θ by $\hat{\theta}$)
- We will use these notations in this part of the session only (to understand bias and variance)

19

Notations

- θ: true and unknown values of parameters that represent the actual relationship between the input and output variables.
- $\hat{\theta}$: our estimate of these parameters using the training data (recall the update rule in the gradient descent)
- Since the estimation of $\hat{\theta}$ depends on the examples, we can say that for n points

$$\hat{\theta} = g\left(X^{(1)}, X^{(2)} \ldots, X^{(n)}\right)$$

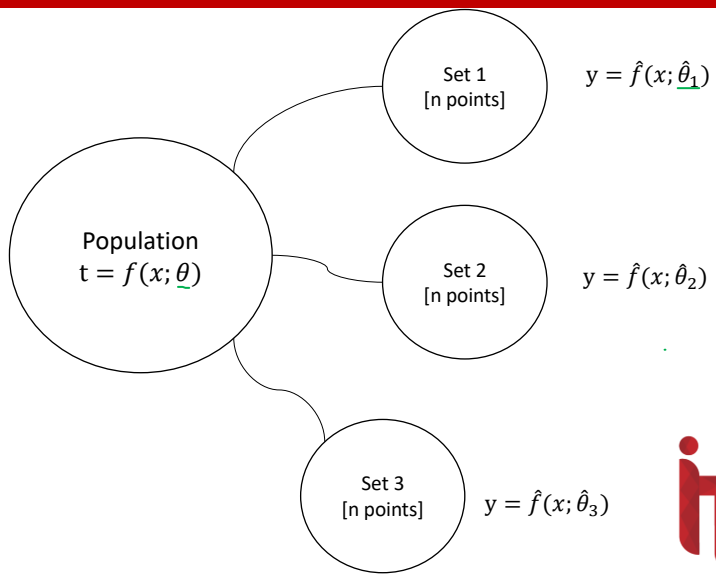
$X^{(1)}$: input at point 1, $X^{(2)}$: input at point 2,…

20

Estimation

- If we have three different sample sets (each set has the same number of examples n), we will have three different estimates of the model parameters

$\hat{\theta}$ is a random variable



21

Bias

- The bias is defined as

$$bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

- An estimator $\hat{\theta}$ is said to be unbiased if $bias(\hat{\theta}) = 0$ (that is, $\mathbb{E}(\hat{\theta}) = \theta$) .
- An estimator $\hat{\theta}$ is said to be asymptotically unbiased if

$$\lim_{n\to\infty} bias(\hat{\theta}) = 0 \text{ (that is, } \lim_{n\to\infty} \mathbb{E}(\hat{\theta}) = \theta)$$

n: number of examples used to estimate $\hat{\theta}$

22

Question

- If the true model is

$$t = 1.5 + 0.5\,x + 2x^2 + x^3 + \epsilon$$

- You are trying to fit the data generated from the true model using two parameters only

$$y = \hat{\theta}_0 + \hat{\theta}_1 x$$

- Is the fitted model unbiased?

23

Bias

- More flexible methods and higher capacity models result in less bias.

24

Variance

- The variance of an estimator is simply the variance:

$$Var(\hat{\theta})$$

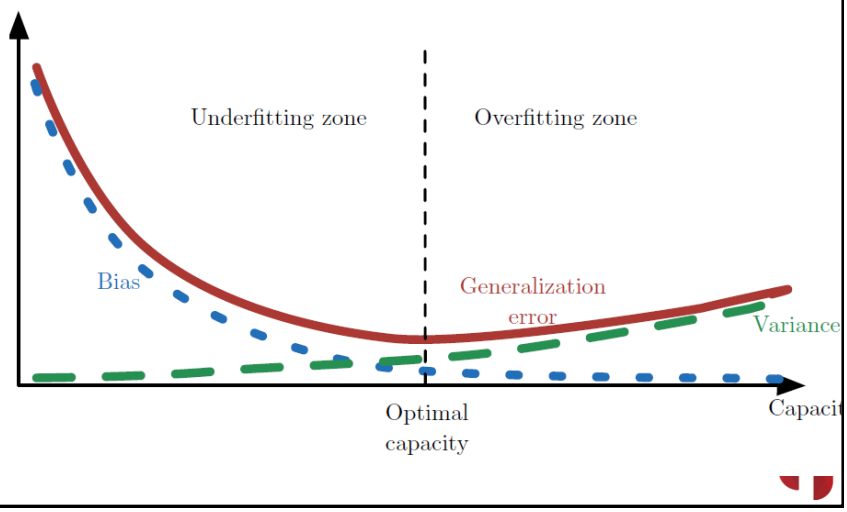
- The variance or the standard error of an estimator provides a measure of how we would expect the estimate we compute from data to vary as we independently resample the dataset from the underlying data generating process.
- If a machine learning method has high variance then small changes in the training data can result in large changes in our estimation.
- In general, more flexible statistical methods have higher variance.
- A good model has low bias and low variance.

25

Bias-Variance tradeoff

Increasing capacity tends to increase variance and decrease bias

Bias and variance are tightly linked to the concepts of capacity, underfitting and overfitting



26

Feature Selection

How to select input variables in linear models?

27

Best Subset Selection

- We fit a separate least squares regression for each possible combination of the p predictors.
- That is, we fit all d models selection that contain exactly one predictor, all models that contain exactly two predictors, and so forth.
- We then look at all of the resulting models, with the goal of identifying the one that is best.
- Think of p variables. Each variables has a value of true (included in the model) or false (not included). What is the possible number of models
- We have 2^p models

28

Sequential Feature Selection¶

- For computational reasons, best subset selection cannot be applied with very large p.
- The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on unseen data.
- For these reasons, stepwise methods, , which explore a far more restricted set of models, can be used
- Two methods
  - Forward Stepwise Selection
  - Backward Stepwise Selection

29

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

30

Backward Stepwise Selection

- Begin with the full least squares model containing all d predictors
- Iteratively removes the least useful predictor, one-at-a-time.

31

Backward Stepwise Selection

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

32

Other techniques (we will review them in scikit learn)

33

Removing features with low variance

- A simple approach for feature selection
- Zero-variance (constant) features do not explain the variability in the output.

34

Univariate feature selection

- Univariate feature selection works by selecting the best features based on univariate statistical tests.

35

Model-based selection

- Use models such as Lasso regression to select features

36

Model Selection

- In machine learning tasks, we need to:
  - Choose, from a range of different types, the best algorithm for the particular problem at hand
  - Finding appropriate values for hyperparameters within a given model.
- We cannot use the test data for tuning hyperparameters. Why?
- How can we perform model selection?

37

Key techniques

1. Use a validation set
2. Cross Validation

38

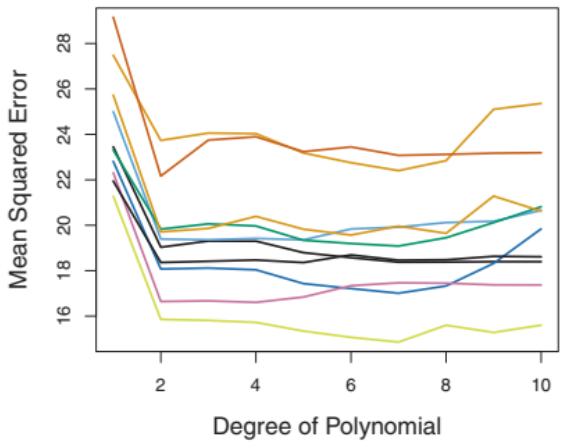
Validation set

- Validation set: a set of examples that the training algorithm does not observe.
- Using a validation set involves randomly dividing the available training set of into two parts, a *training set* and a *validation set* or *hold-out set*.

39

Using a validation set - Drawbacks

- For small datasets, the validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In figure: The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set.



Chapter 5 from the book : An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0.

40

Using a validation set - Drawbacks

- Statistical methods tend to perform worse when trained on fewer observations
- Since in the validation approach, only a subset of the observations is used for training, the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire data set.

41

Alternative approach: Cross-Validation

- Cross-Validation: an alternative procedure that enables us to use all of the examples in the estimation of the mean test error, at the price of increased computational cost.
- Methods
  - *Leave-One-Out Cross-Validation*
  - *k-Fold Cross-Validation*

42

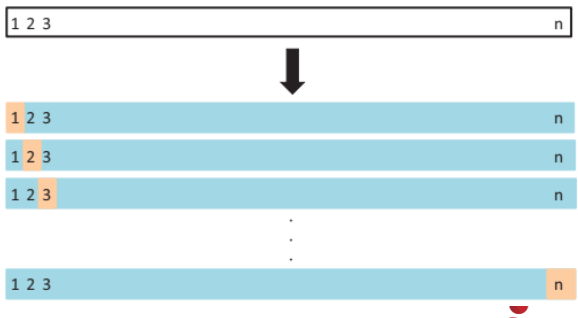
Leave-P-Out Cross-Validation

- Select p out of n observations as the validation set and the remaining as the training set.
- Repeat for all possible combinations.
- Problem with this approach: how many ways we have?
  - C_p^n
- Calculate the combinations for n =150 and p =50
  - 2.01286609 E+40
- Choose p =1, what is the number of combinations?
  - 150

43

Leave-One-Out Cross-Validation (LOOCV) - procedure

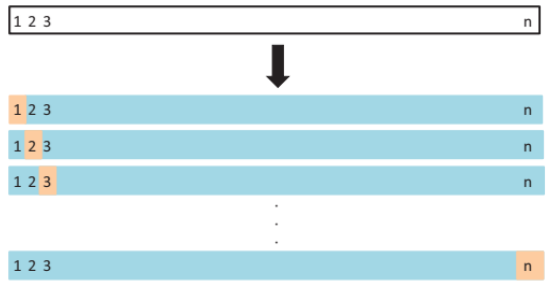
1. A single observation (x_1, t_1) is used for the validation set, and the remaining observations $\{(x_2, t_2), \ldots, (x_n, t_n)\}$ make up the training set.
2. The statistical learning method is fit on the $n - 1$ training observations, and a prediction is made for the excluded observation



44

LOOCV – procedure (Cont.)

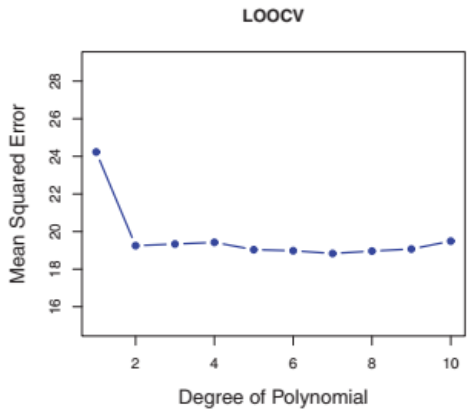
3. Rpeat the procedure by selecting (x_2, t_2) for the validation data, training the statistical learning procedure on the n - 1 observations $\{(x_1, t_1), (x_3, t_3) \ldots, (x_n, t_n)\}$ make up the training set.
4. The statistical learning method is fit on the $n - 1$ training observations, and a prediction is made for the excluded observation
5. The test error of the cross validation is the average error of these n test error estimates



45

Why does *LOOCV* the problems of the validation set approach?

- LOOCV, we repeatedly fit the statistical learning method using training sets that contain $n - 1$ observations, almost as many as are in the entire data set (in contrast to the validation set approach, which uses a considerable portion of the dataset for validation).
- Performing LOOCV multiple times will and calculating the average test error will always yield the same results (variable results for selected validation dataset is not an issue).

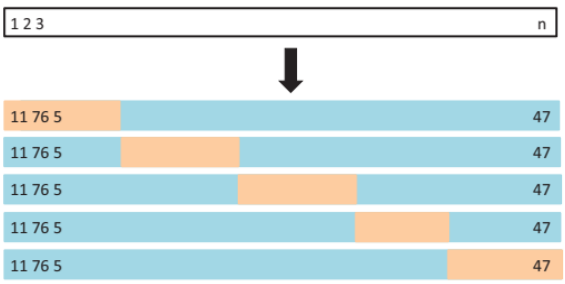


46

k-Fold Cross-Validation

- An alternative to LOOCV is *k-fold CV*

1. *Randomly divide the set of n observations into k groups, or folds, of approximately equal size.*
2. *The first fold is treated as a validation set, and the method is fit on the remaining k − 1 folds.*
3. *The error is calculated using the* observations in the held-out fold

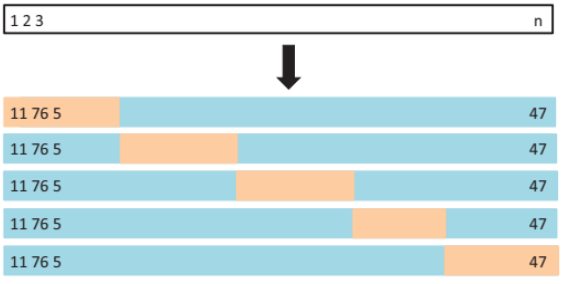


5-fold CV

47

k-Fold Cross-Validation (Cont.)

4. *Repeat the process k times,* each time, a different group of observations is treated as a validation set.
5. The test error of the cross validation is the average error of these k test error estimates

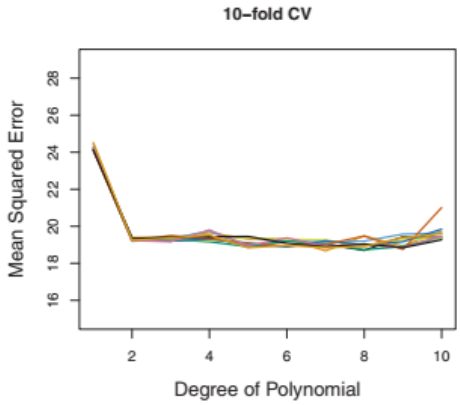


5-fold CV

48

Advantages of k-fold compared to LOOCV

- LOOCV is a special case of k-fold CV in which k is set to equal n
- What is the advantage of using a small k (typically between 5 and 10)
- The most obvious advantage is computational. LOOCV requires fitting the statistical learning method n times. This has the potential to be computationally expensive (especially if n is extremely large).
- In figure: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts.



49

Cross-Validation - Considerations

- The number of training runs that must be performed is increased by a factor of S (problematic for models in which the training is itself computationally expensive)
- There might have multiple hyperparameters for a single model (for instance, several regularization parameters). Exploring combinations of settings for such parameters could, in the worst case, require a number of training runs that is exponential in the number of parameters.

50

What to do if the learning fails

- Get more data (a larger sampl)e
- Change the model by:
  - Increase the capacity
  - Reduce the capacity
  - Change the technique/method
  - Changing the hyperparameters you consider
- Change the feature representation of the data
- Change the optimization algorithm used to apply your learning rule (e.g. loss function modifications)

51