Rensselaer

why not change the world?®

Model Validation, Generalization, Error Estimation

Ahmed Eleish

Data Analytics ITWS/CSCI/MGMT-4600/6600

March 28th 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

Training, test and validation sets

- Training: subset of dataset used as input to the model's training algorithm
- Validation: subset used to evaluate models during training
- Test: subset used to test the final model

e.g. the training set (70%) is used to train multiple models (different features, parameters, etc.) and the validation set (20%) is used to compare and select the best performing model. The test set (10%) is then used to evaluate the selected model.

Terminology Confusion!

- ‘Test’ and ‘validation’ are used interchangeably in academia and industry!

- That’s fine… make sure to know the proper use.

https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets

Errors

- The error on the training data is called as the “**training error**”
- The error on the test data is referred to as the **”test error”**
- **The error on the test data is a good indication of how well the classifier will perform on new data and this is known as the generalization**.
- If the classifier performs well on the new data, then it is a good generalization. Generalization refers to how well the model is performing on unseen data (**data not used to train the model**)

Test error : Generalization error

- If the model generalizes well, then it will perform well on the new data sets that has the *similar structure* to the training data..

- Since the Test error is an indication of how well the model generalizes to new data, *the test error also called the generalization error.*

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition

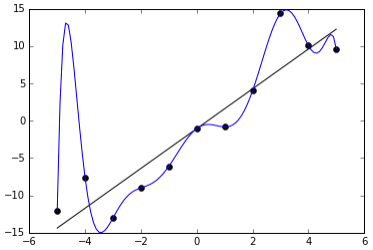
Overfitting

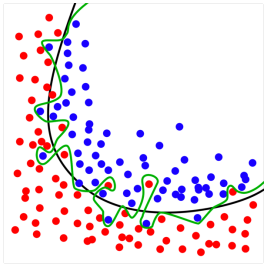
- Another related concept to Generalization is “overfitting”.
- If the model has very low training error but it has high generalization error, then it is over fitting.

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition

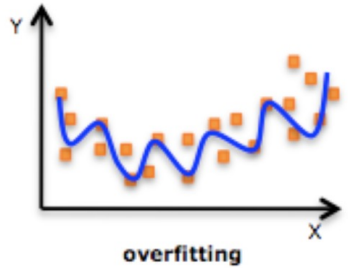
Overfitting

- This is a good indication that the model may have learned to ***model the noise*** in the training data, instead of the learning from the underlying structure of the data.
- Overfitting is an indication of poor generalization.

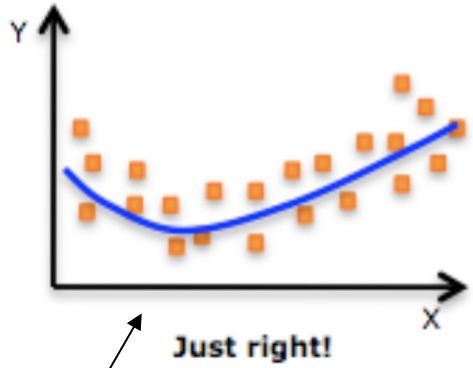




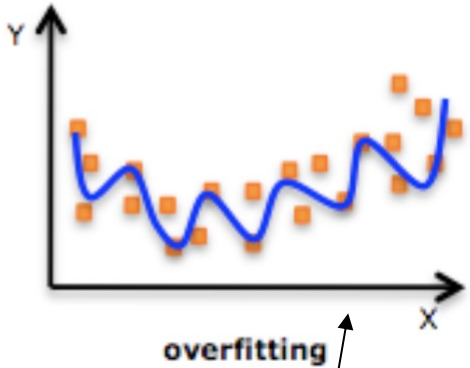
Image/Photo Credit:
https://en.wikipedia.org/wiki/Overfitting#/media/File:Overfitted_Data.png



Image/Photo Credit:
http://pingax.com/regularization-implementation-r/



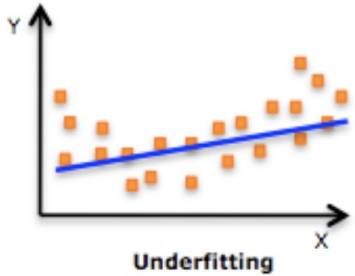
Model is fitting to
the structure of the data



Model is fitting to
the noise of the data

Underfitting

- **Underfitting** occurs when a statistical model cannot adequately capture the underlying structure of the data.
- In other words, **underfitting take place when the model has not properly learned the structure of the data**.



Image/Photo Credit: http://pingax.com/regularization-implementation-r/

Robustly Validating Models

- There are several ways to create the evaluate/validate models
  - Holdout method
  - K-fold Cross validation
  - Monte Carlo Cross validation
  - Leave-One-Out Cross validation

Holdout Method

- Split the dataset into 2 subsets, one for training and another for testing.
- The training set is usually larger than the test set.
- Not recommended for robust validation.

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

K-fold Cross Validation

- In k-fold cross validation, the data are segmented in to *k* number of **disjoint partitions**.
- During each iteration, one partition is used as the test set and the remaining k-1 (combined) for training; The process is repeated *k* times.
- Each time using a different partition for testing, so that each partition is used exactly one time for the validation.

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

Monte Carlo Cross Validation (Repeated random sub-sampling)

- In Monte Carlo cross validation, the dataset is split into training/test sets over *n* iterations with the samples in each selected at random.
- The size of each partitions may be constant or vary over the iterations.
- Commonly used in research, considered robust because of the averaging effect over multiple iterations.
- Downside: since selection is random, some observations may not end up in test sets and some may be oversampled

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

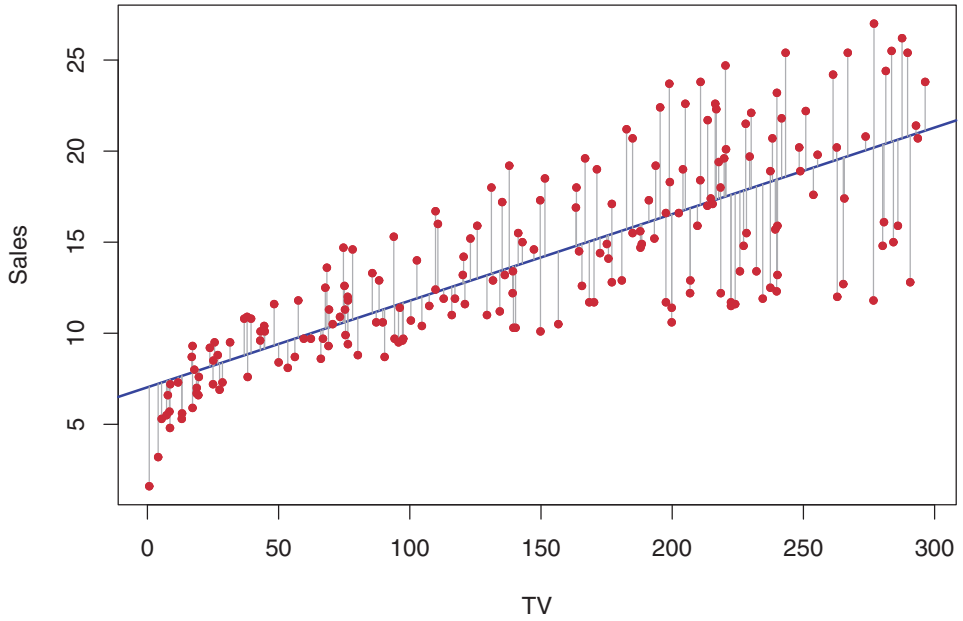
Leave One Out Cross Validation (LOOCV)

- For as many iterations as there are observations, drop one observation and use all the others for training; test on the 1 observation and average at the end.
- Every observation is tested once.
- Depending on the size of the dataset, may be computationally expensive.

Resource/Reference: Introduction to Statistical Learning with R, 7th Edition - Chapter 5

Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $



Evaluating Linear Models

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)},$$

Residual Standard Error

- Mean sales ≈ 14,000 units

RSE = 3.26 = 3,260 units
good/bad?

R^2
- measures the proportion of the variability in *Y* that can be explained using *X*
- has a value between 0,1

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

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

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

Residual Sum of Squares (RSS)

For given data *(x1,y1), ..., (x$_n$,y$_n$)* $\in \mathbb{R}\times\mathbb{R}$,

- Residual Sum of Squares (RSS), the *i*th residual $e_i = y_i - \hat{y}_i$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Mean Absolute Error

- Mean(||Predicted value - Real value||)

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

Mean Squared Error

- Mean((Predicted value - Real value)^2)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error

- SquareRoot(Mean((Predicted value - Real value)^2))

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}}$$

in-class exercise

https://rpi.box.com/s/f0ipdgmdul7dig11kxyov3l2v0fk8027

Thanks!
Have a great weekend!

Work on your assignment/project!!!