Rensselaer

why not change the world?®

Generalization, Model Validation and Optimization

Ahmed Eleish

Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

February 14th 2025

Tetherless World Constellation
Rensselaer Polytechnic Institute

Model Evaluation, Generalization

Errors in Classification

- We’ve seen classification errors working with the iris classification examples.

- In classification, the model’s output is the predicted class label for the input variables and the true class label is the target.

- **If the predicted class label is different from the actual class label (true class) then there is an error with that classification**.

Misclassification Error

- The error rate is the percentage of errors made over the entire dataset
- Error rate is also known as the misclassification rate or simply called the error.
- Error = (*Number of Misclassifications*)/(Total Number of Samples)
- Accuracy = (*Number of Correct Classifications*)/(Total Number of Samples)

Misclassification Error

e.g.

	actual		
predicted	setosa	versicolor	virginica
setosa	15	0	0
versicolor	0	16	2
virginica	0	0	17

- Evaluating a kNN model trained on 2/3 of observations in the Iris dataset and tested on the remaining 1/3
- Error = 2/50 = 0.04 = 4%
- Accuracy = 48/50 = 0.96 = 96%

Evaluation of Model Training

- To robustly evaluate predictive models the training process is repeated multiple times according to commonly used sampling strategies.

- The goal is for model training to be exposed to as much of the variation in structure in the dataset as is reasonably possible.

- Each training iteration is evaluated separately, with the average performance of the model over the number of training iterations considered an indicator of training success.

Training, Validation and Test 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 Iris dataset is initially split into a training set (90% - 135 obs) and a test set (10% - 15 obs) ~ this depends on the size of the dataset.
- Over 10 iterations, the training set is split into training (100 obs) and validation (35 obs), and after training, the average training error is calculated
- The final model is tested on the test set (15 obs) and the test error is calculated

Errors

- The error on the training (validation set) 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 the new 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

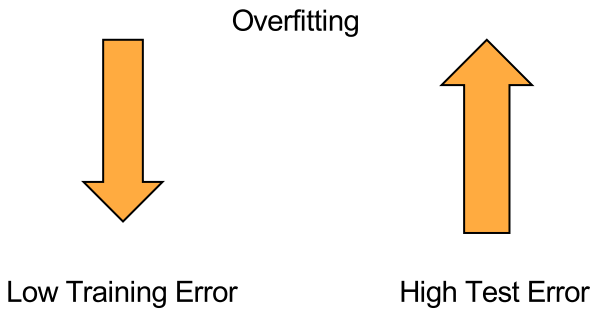
Terminology Confusion!

- ‘Test’ and ‘validation’ are used interchangeably in academia and industry!!
- That’s fine… just make sure you know which one you mean!
- It is common **NOT** to keep a separate test set for the final model, especially in non-published research. Instead, the dataset is split into training/test sets for every training iteration.
- When reporting errors, preferably specify if it’s training set error or test set error.

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

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.

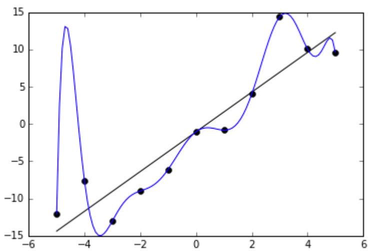


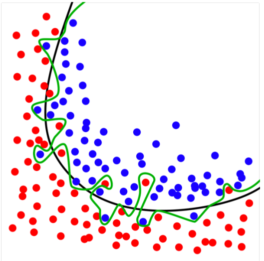
Resource/Reference: Introduction to Statistical Learning with R, 7th Edition

Overfitting

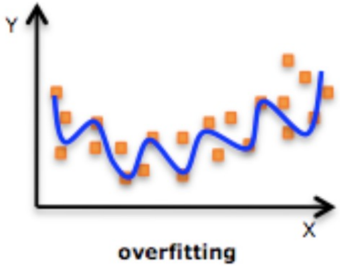
- This is a good indication that the model may have learned to ***model the noise*** in the training data, instead of the learning from the underlying structure of the data.

- Overfitting is an indication of poor generalization.

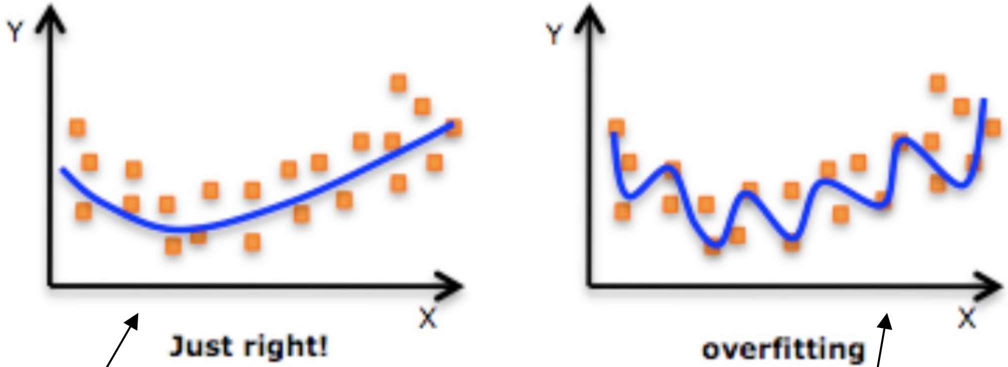




Image/Photo Credit:
https://en.wikipedia.org/wiki/Overfitting#/media/File:Overfitted_Data.png



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



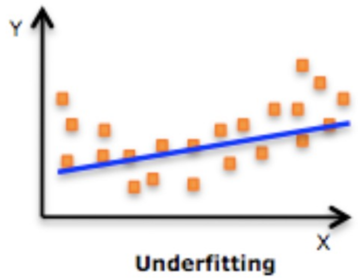
Model is fitting to
the structure of the data

Model is fitting to
the noise of the data

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

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/

Cross-Validation

Robustly Validating Models

- There are several ways to robustly evaluate/validate models
  - K-fold Cross validation
  - Monte Carlo Cross validation
  - Leave-One-Out Cross validation

https://en.wikipedia.org/wiki/Cross-validation_(statistics)

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 used all the others for training; test one the 1 observation and average at the end.
- Depending on the size of the dataset, may be computationally expensive.

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

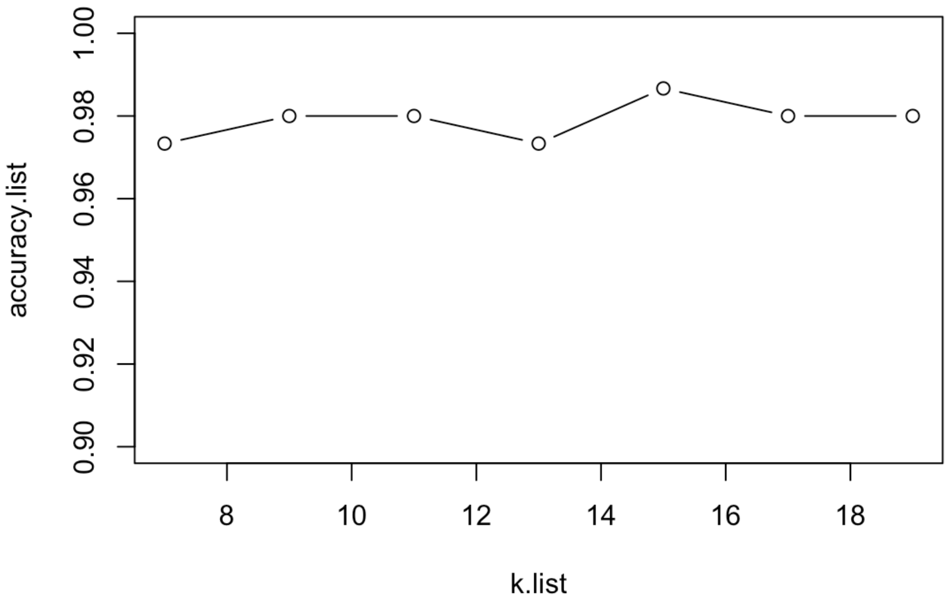
https://rpi.box.com/s/x2w6ucxzlpe72le55iwh8d4q4qioab4o

Optimizing kNN Models

- The parameter *k* represents the number of nearest neighbors used by the algorithm
- Rule of thumb: $k = n^{1/2} = \sqrt[2]{n}$
- Finding the optimal value for *k*
  - For a range of *k* values, train a kNN model and calculate classification accuracy
  - Select *k* value from best performing model

Optimizing kNN Models

- Finding the optimal value for *k*
  - For a range of *k* values, train a kNN model and calculate classification accuracy
  - Select *k* value from best performing model

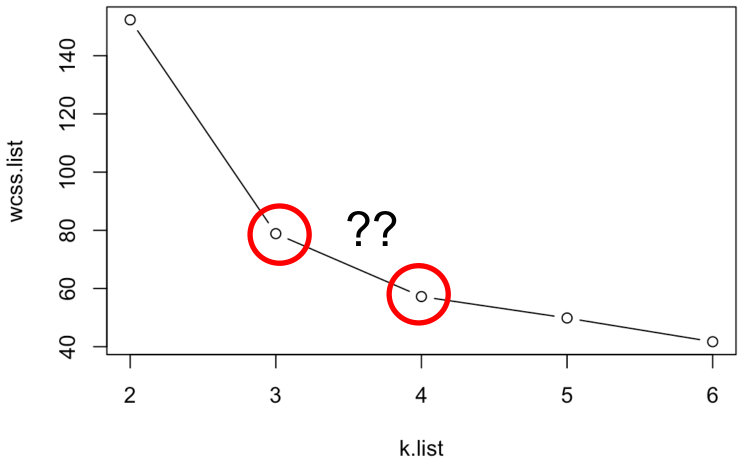


Optimizing K-Means Models

- The parameter *K* represents the number of clusters to be identified by the algorithm
- Depends on background knowledge/research question
- Finding the optimal value for *K*
  - For a range of *K* values, train a K-Means model and calculate within cluster sum of squares (WCSS)
  - Select *K* value where after which the decrease in WCSS diminishes
  - This is called the elbow method

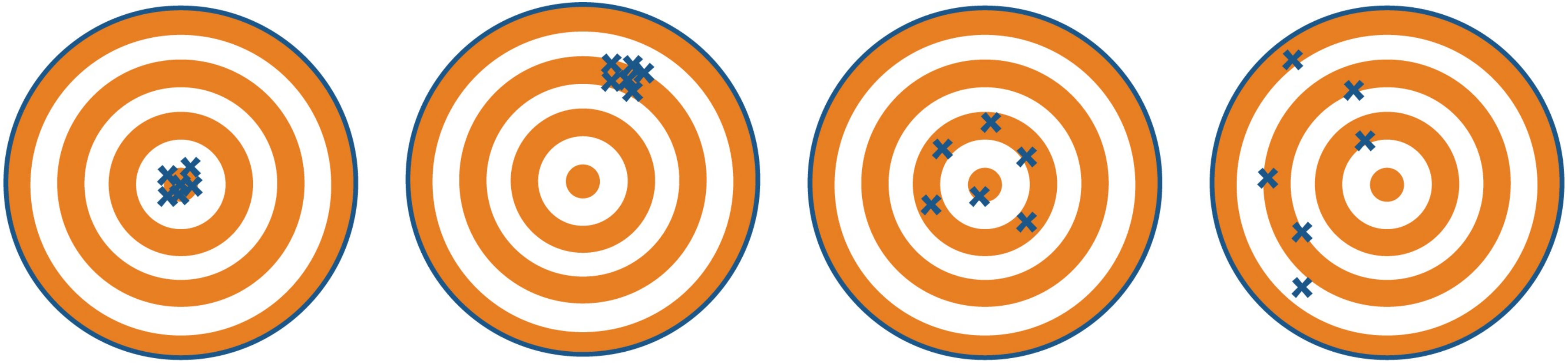
Optimizing K-Means Models

- Finding the optimal value for *K*
  - For a range of *K* values, train a K-Means model and calculate within cluster sum of squares (WCSS)
  - Select *K* value where after which the decrease in WCSS diminishes
  - This is called the elbow method



Metrics for Evaluating Classification & Clustering Models

Accurate vs. Precise



High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

http://climatica.org.uk/climate-science-information/uncertainty

Classification Metrics

Classification Accuracy

- ***Accuracy = (Number of correct predictions) / (Total number of data points)***

$$= \frac{TP+TN}{N}$$

- Simplistic evaluation of model

- Classification error = 1 – ***Accuracy***

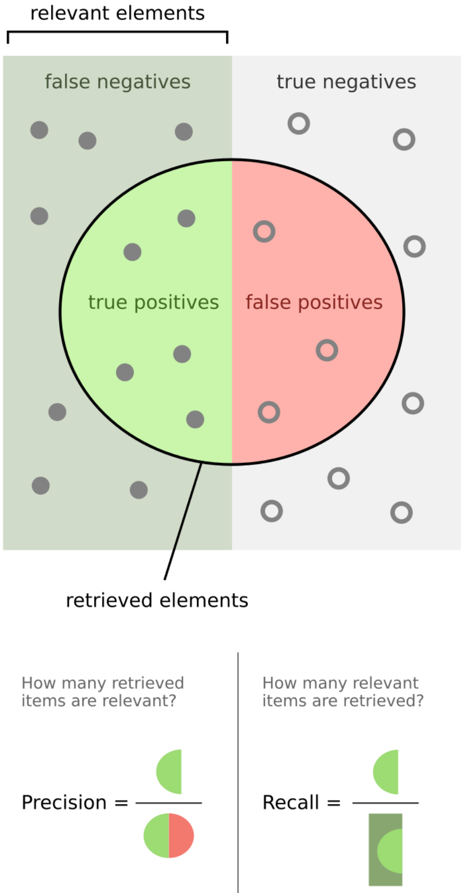
$$= \frac{FP+FN}{N}$$

		Predicted Value	
		Positive	**Negative**
Real Value	**Positive**	TP	FP
	Negative	FN	TN

Per Class Evaluation

$$\text{Precision} = \frac{\text{Relevant retrieved instances}}{\text{All } \textbf{retrieved} \text{ instances}}$$

$$\text{Recall} = \frac{\text{Relevant retrieved instances}}{\text{All } \textbf{relevant} \text{ instances}}$$



https://en.wikipedia.org/wiki/Precision_and_recall

Evaluation Metrics – Per Class

- ***Precision = (True Positive) / (True Positive + False Positive)***
  - ***Proportion of positive predictions that are correct***
- ***Recall = (True Positive) / (True Positive + False Negative)***
  - ***Proportion of positive class correctly identified***
- ***F1 = 2 [(Recall * Precision) / (Recall + Precision)]***
  - ***F1 = (True Positive) / [True Positive + 1/2*(False Positive + False Negative)]***
  - ***Harmonic mean (weighted average) of precision and recall***

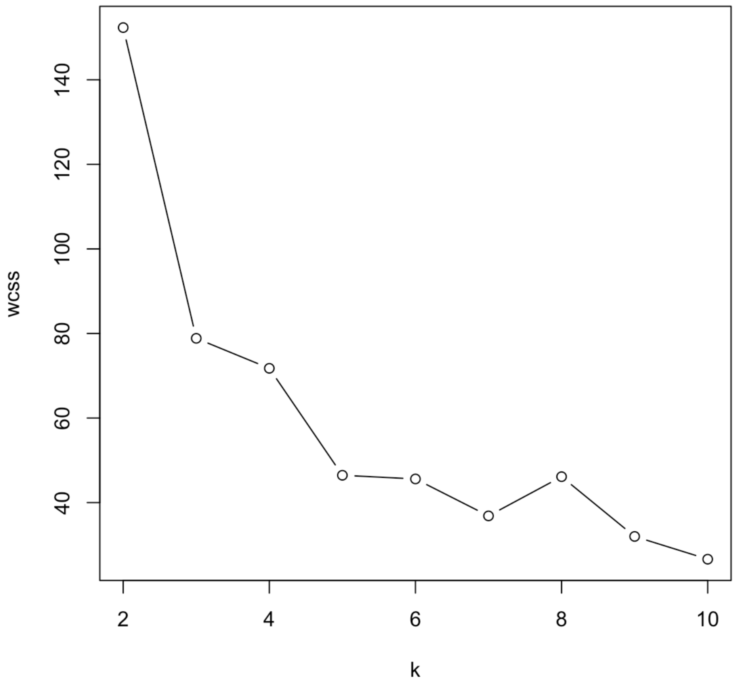
Evaluation Metrics – Per Class

- ***Specificity = (True Negative) / (True Negative + False Positive)***
  - *Fraction of correct predictions belonging to negative class*
- ***Fall-out = (False Positive) / (True Negative + False Positive)***
  - *Fraction of negative class correctly classified*
- ***Miss Rate = (False negative) / (True positive + False negative)***
  - *Fraction of positive class misclassified*

Evaluating Clustering Models

Within-Cluster Sum of Squares (Elbow Method)

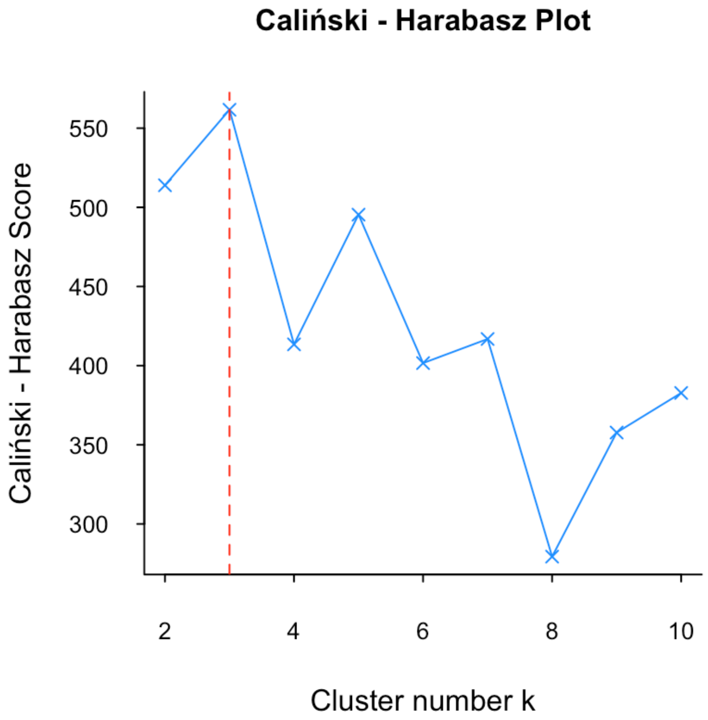
$$WCSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mathbf{c}_i||^2$$



Calinski–Harabasz index (CHI)

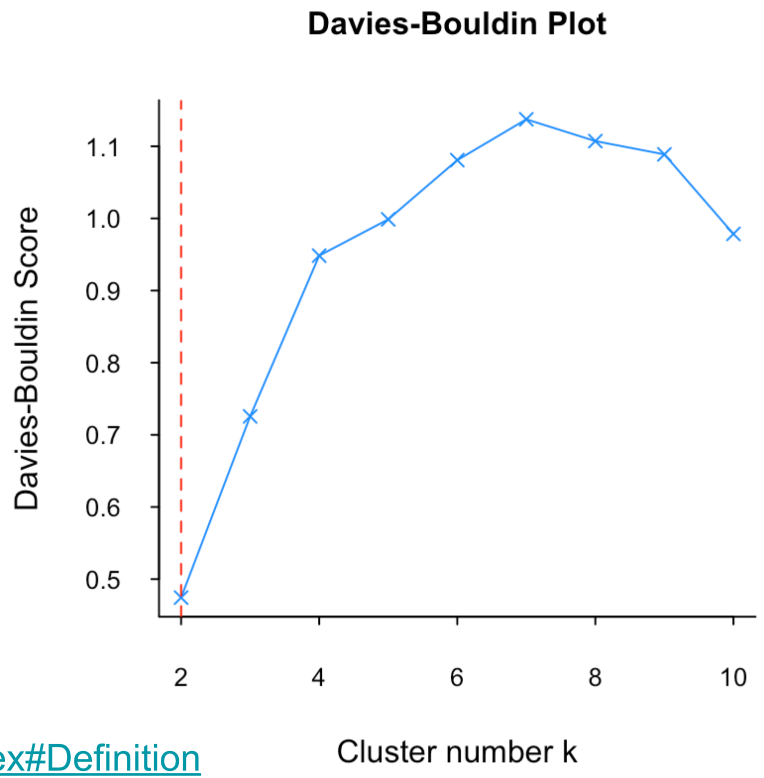
$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

$$BCSS = \sum_{i=1}^{k} n_i ||\mathbf{c}_i - \mathbf{c}||^2$$



Davies — Bouldin Index (DBI)

- Lower index value -> better clustering
- Indicates increased separation between clusters and decreased variation within clusters



https://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index#Definition

Thanks!