



Rensselaer

why not change the world?®

Evaluating Classification & Clustering Models

Ahmed Eleish

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

October 22th 2024



Evaluating Classification Models



Classification Accuracy

- *Accuracy = (Number of correct predictions) / (Overall number of predictions)*

		<i>Predicted Value</i>	
		Positive	Negative
<i>Real Value</i>	Positive	TP	FP
	Negative	FN	TN



Evaluation Metrics

- *Recall = (True Positive) / (True Positive + False Negative)*
- *Precision = (True Positive) / (True Positive + False Positive)*
- *F1 = 2 [(Recall * Precision) / (Recall + Precision)]*
 - *F1 = (True Positive) / [True Positive + 1/2*(False Positive + False Negative)]*



Evaluation Metrics

- ***Specificity = (True Negative) / (True Negative + False Positive)***
- ***Fall-out = (False Positive) / (True Negative + False Positive)***
- ***Miss Rate = (False negative) / (True positive + False negative)***

In-class exercise

<https://rpi.box.com/s/l9nc8cgzcn3xnkfd9oxsxx7cb8gu72u>

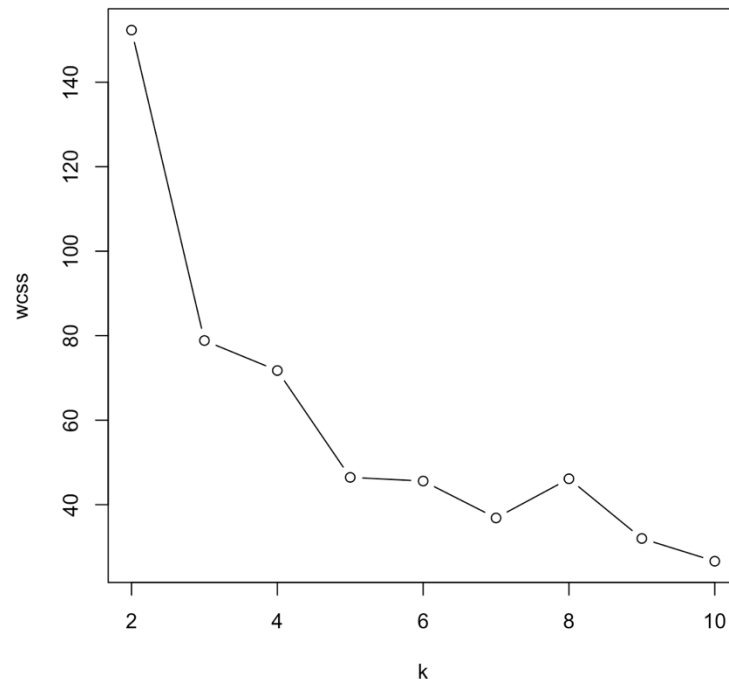


Evaluating Clustering Models



Within-Cluster Sum of Squares (Elbow Method)

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

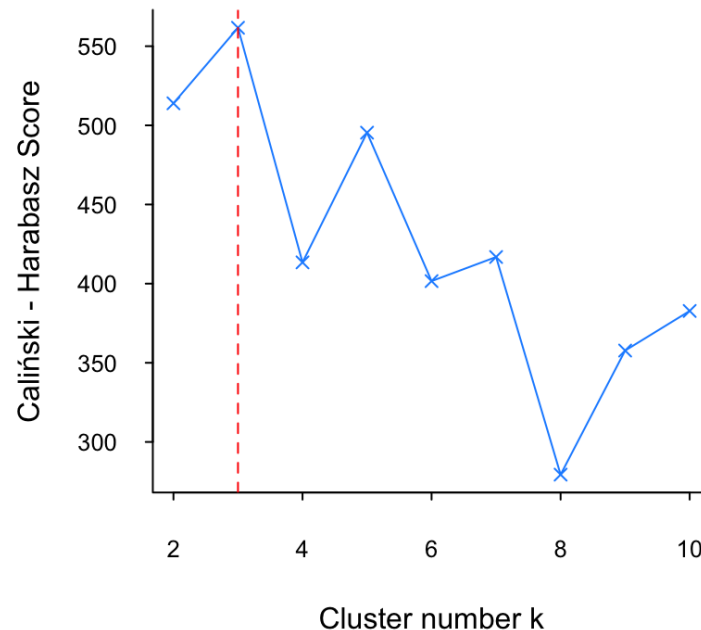


Calinski–Harabasz index (CHI)

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

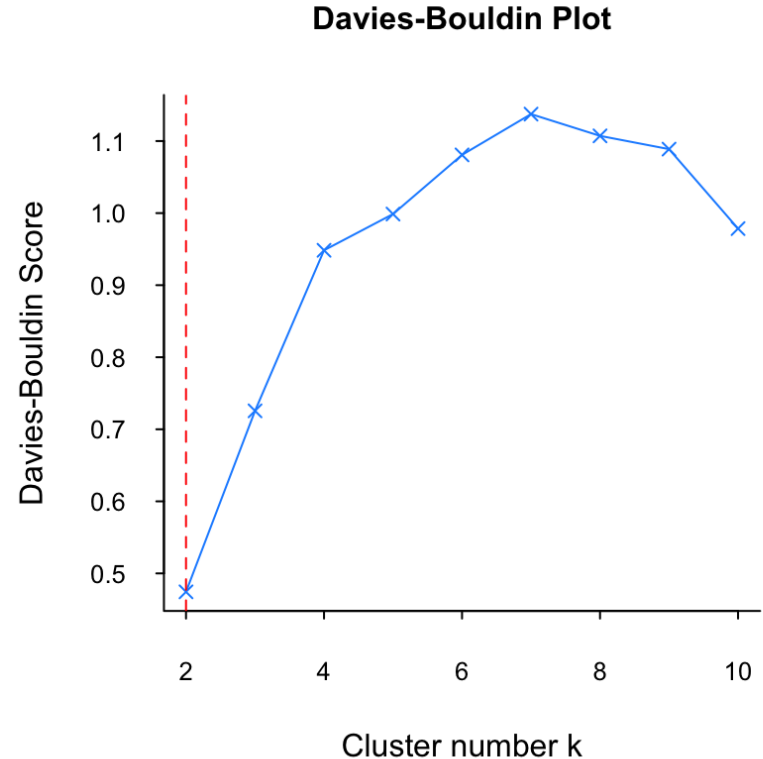
$$BCSS = \sum_{i=1}^k n_i \|\mathbf{c}_i - \mathbf{c}\|^2$$

Calinski - Harabasz Plot



Davies — Bouldin Index (DBI)

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



Thanks!