Metrics for Evaluating Classification & Clustering Models

Ahmed Eleish
Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960
February 25th 2025

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

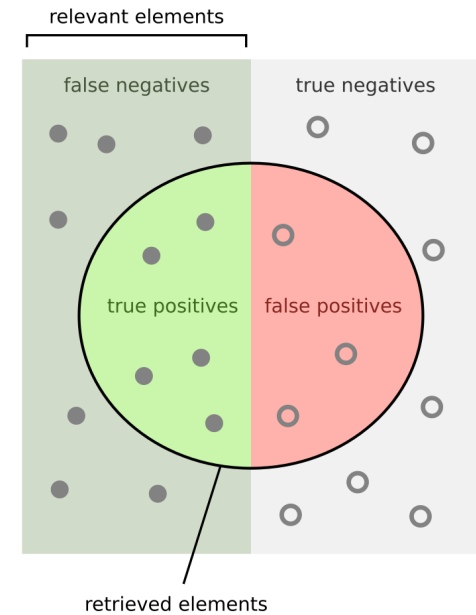- Classification error = 1 – ***Accuracy***

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

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

Rensselaer

Per Class Evaluation

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

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

Evaluation Metrics – Per Class

- *Precision = (True Positive) / (True Positive + False Positive)*

  - *Fraction of correct predictions belonging to positive class*

- *Recall = (True Positive) / (True Positive + False Negative)*

  - *Fraction of positive class correctly classified*

- *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)*

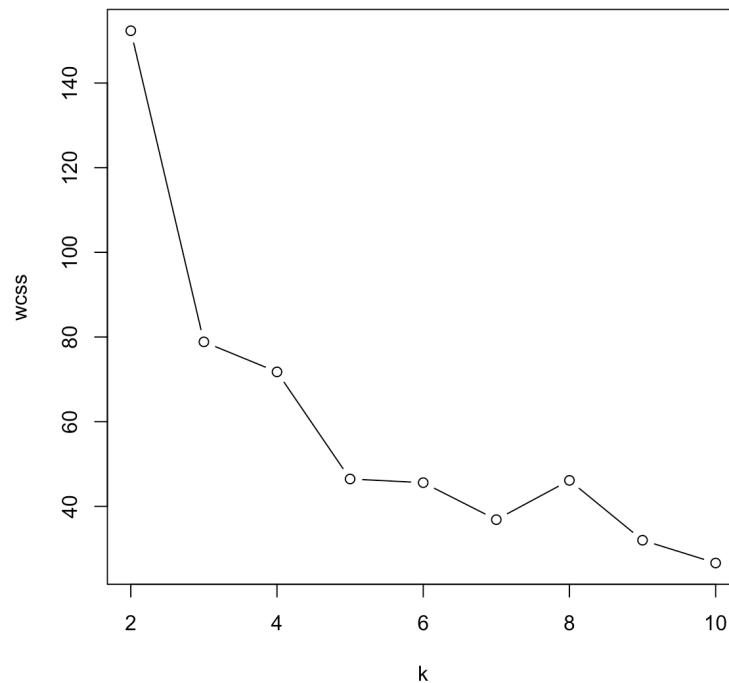 - *Fraction of positive class misclassified*

In-class exercise

https://rpi.box.com/s/8deje65k0xrhax597o8ahj1tv6bl396f

Evaluating Clustering Models

Within-Cluster Sum of Squares (Elbow Method)

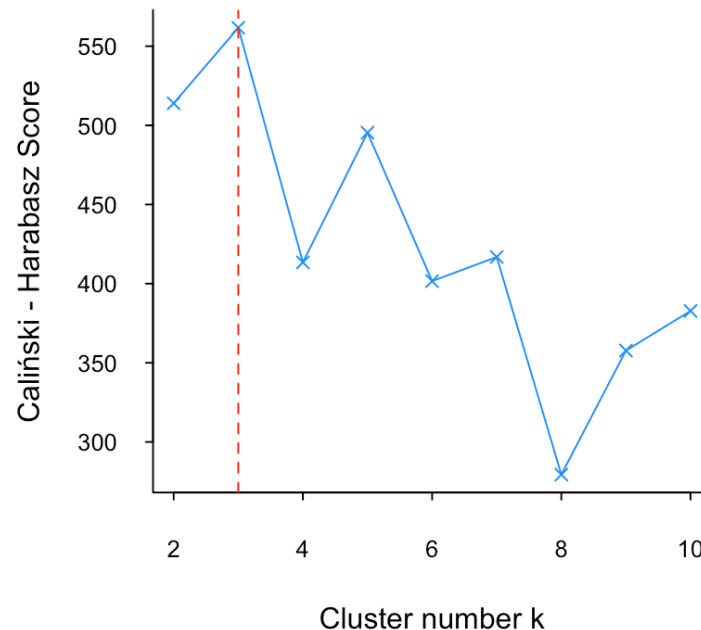$$WCSS = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \left\| \mathbf{x} - \mathbf{c}_i \right\|^2$$

Calinski–Harabasz index (CHI)
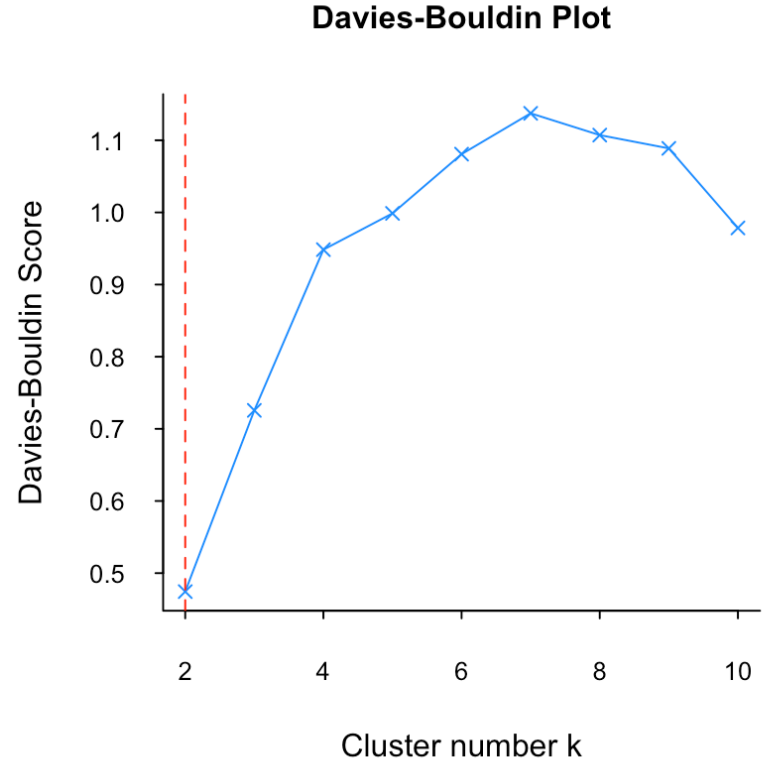
$$CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

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

Caliński - Harabasz Plot

Davies — Bouldin Index (DBI)

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

Davies-Bouldin Plot

Thanks!