DRIVES

Development and Research on Innovative Vocational Education Skills

COMPUTER VISION FUNDAMENTALS

U2.E5. DIMENSIONALITY REDUCTION: FEATURE SELECTION AND EXTRACTION

Computer Vision Expert

May 2021, Version 1

Co-funded by the Erasmus+ Programme of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to …

CVE.U2.E5.PC1	understands the need, definition, and purpose of dimensionality reduction.
CVE.U2.E5.PC2	list the different types of dimensionality reduction.
CVE.U2.E5.PC3	define feature selection and understand its particularities.
CVE.U2.E5.PC4	knows the different feature selection strategies.
CVE.U2.E5.PC5	define feature extraction and understand its particularities.
CVE.U2.E5.PC6	knows the different feature extraction strategies.
CVE.U2.E5.PC7	identify and understand the differences between feature selection and feature extraction.

DIMENSIONALITY REDUCTION

Dimension Reduction (DR) is the process of reducing the number of variables/features in review, and can be divided into two subcategories called **Feature Selection** and **Feature Extraction**.

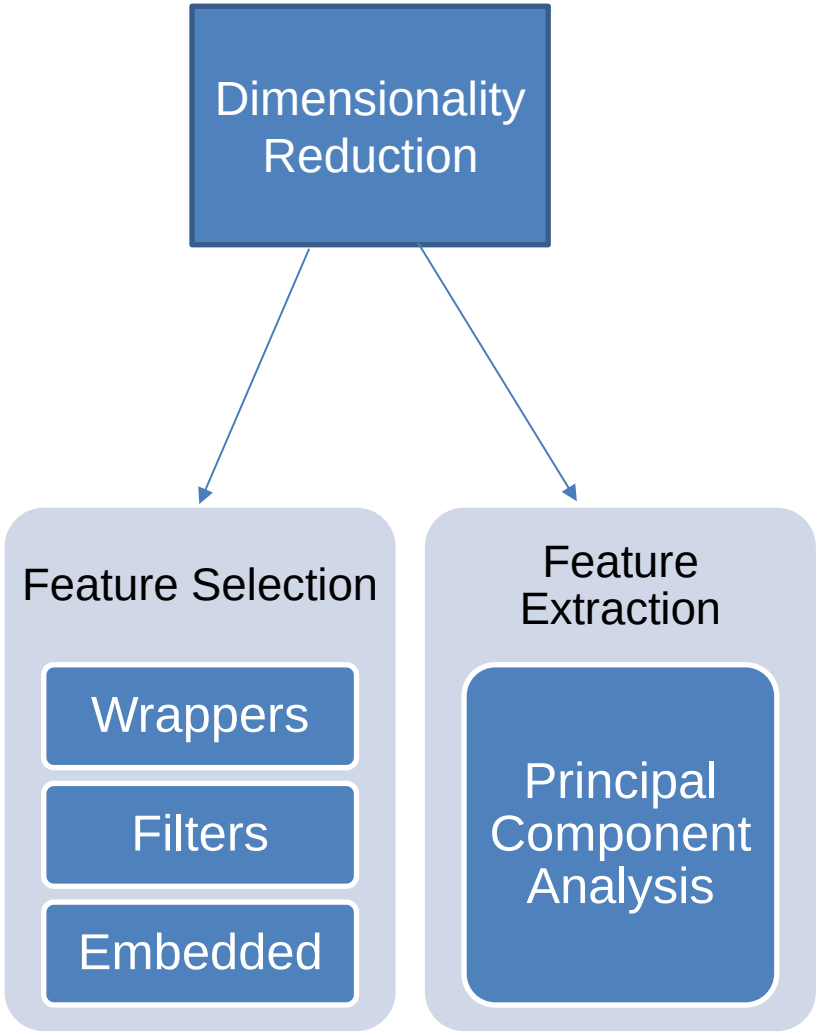
- DR improves the performance by reducing the number of features that are to be considered.
- DR techniques change the original dataset having high dimensionality and turn it into a new dataset, representing low dimensionality while maintaining the original meanings of the data.

DIMENSIONALITY REDUCTION

Some benefits can be acquired:

- Data storage space can be reduced when the number of dimensions comes down.
- Less computation time only.
- Redundant, irrelevant, and noisy data can be eliminated.
- Data quality can be improved.
- Reducing the dimensions helps an algorithm to work efficiently and improves accuracy.
- Reducing the dimension may allow to design and examine patterns clearly.
- Process of classification simplified and also improves efficiency.

TYPES OF DIMENSIONALITY REDUCTION



FEATURE SELECTION

Feature selection is the exhaustive process of selecting a subset of relevant features or variables from the original set of features, which are best representatives of the data. Feature selection is done in the context of an optimization problem.

In text data, features might be size of characters or some global features of the text. Feature selection provides numerous advantages: reduce the size of data, decrease needed storage, prediction accuracy improvement, overfitting evading, and reduce executing and training time from easily understanding variables.

Originally, evaluation methods in feature selection were divided into four categories: *filter*, *wrapper*, *embedded*, and *hybrid*. Recently, another type of evaluation method was developed - ensemble.

FEATURE SELECTION STRATEGIES: **WRAPPERS**

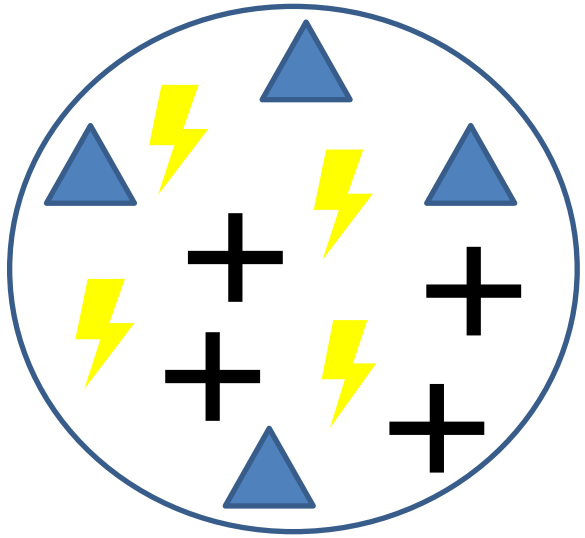
Wrappers use a predictive model that scores feature subsets based on the error-rate of the model.

Popular technique

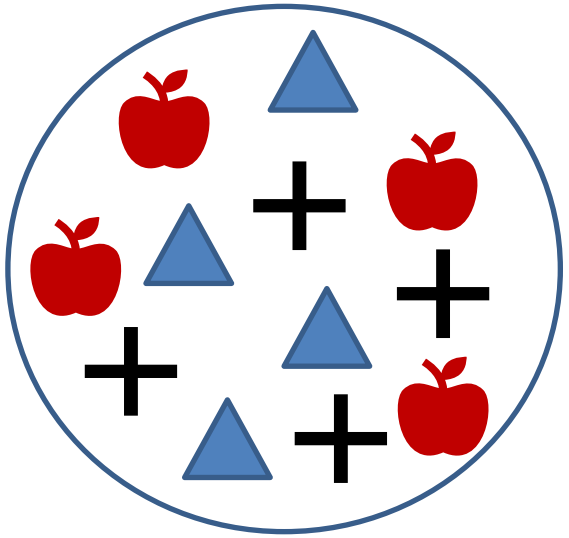
Stepwise Regression

Algorithm that adds the best feature or deletes the worst feature at each iteration.

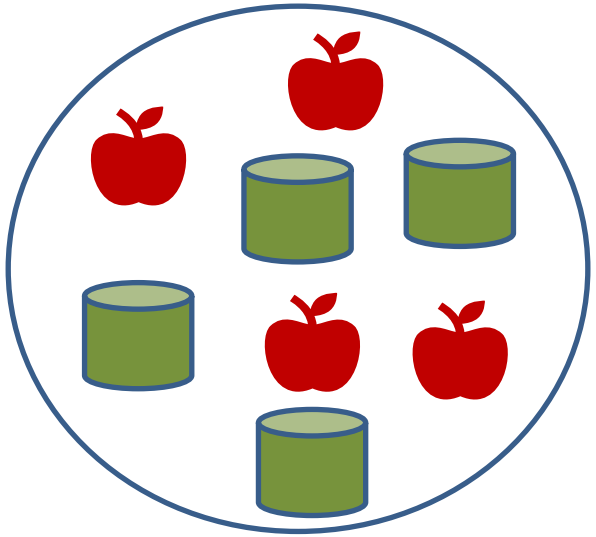
Lowest Error:



Mid-range Error:



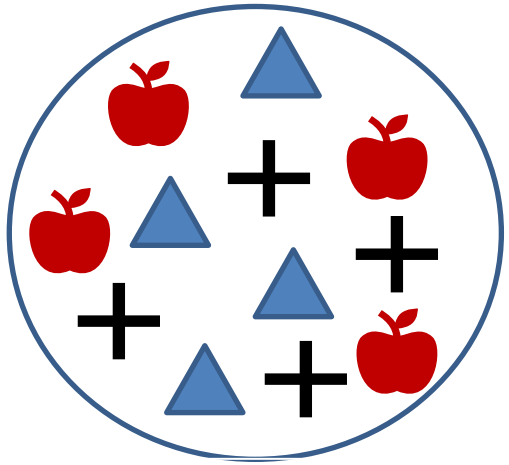
Highest Error:



FEATURE SELECTION STRATEGIES: **FILTERS**

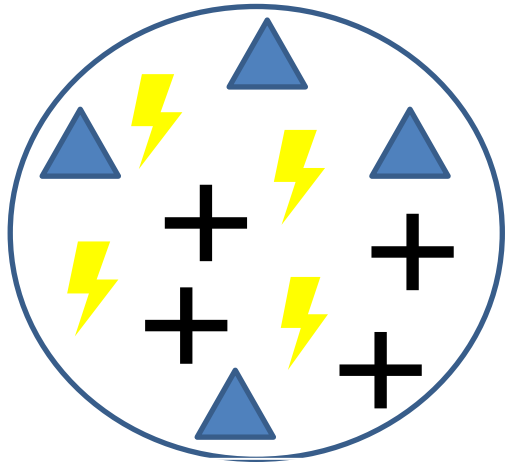
Filters use a proxy measure that is less computationally intensive but slightly less accurate.

Best Prediction:



Instead of

Best Answer:



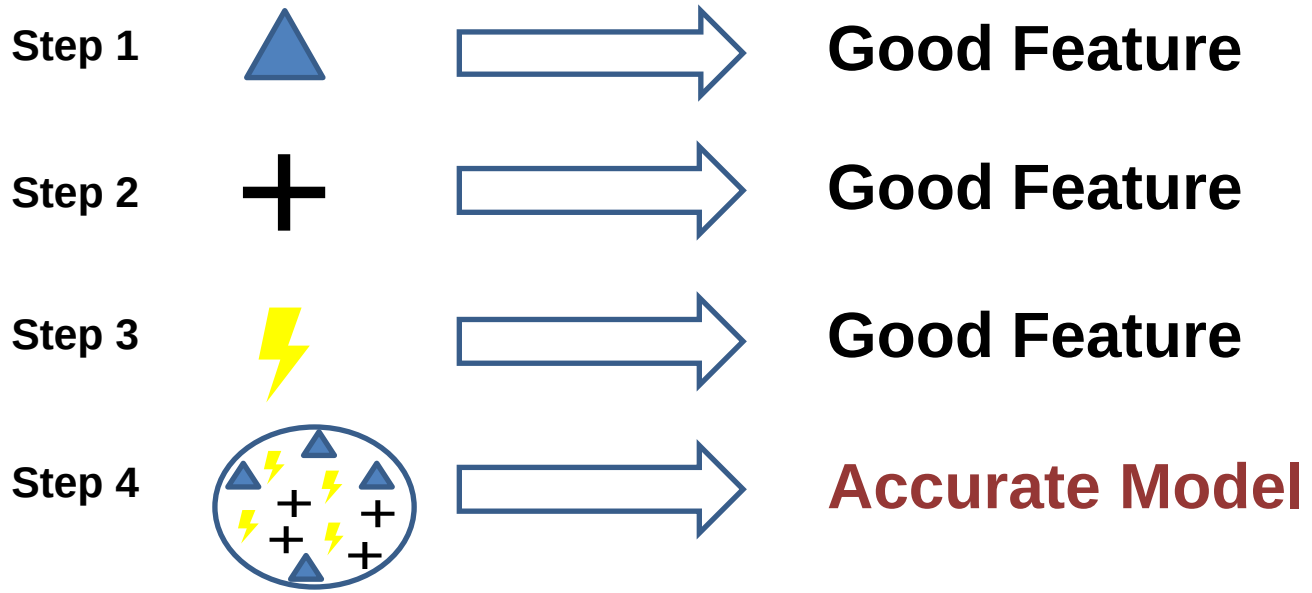
Filters produce a feature set that do not contain assumptions based on the predictive model, making it a useful tool for exposing relationships between features, such as which variables are "bad" together and, as a result, drop the accuracy.

BAD

GOOD

Model Building Process - Regularization Models

Embedded algorithms learn about which feature best contribute to an accurate model during the model building process. The most common type is **Regularization Model**.



FEATURE SELECTION STRATEGIES: **HYBRID AND ENSEMBLE**

DRIVES
Development and Research on Innovative Vocational Education Skills

Hybrid and **ensemble** methods can be represented in the hybrid method. It can be developed either by integrating two various methods (e.g. wrapper and filter), two methods with the same criteria, or two feature selection approaches.

Ensemble method is a method that has the goal of building a group of feature subsets and then producing an aggregated result out of the group. This method depends on various subsampling techniques where a particular feature selection method is implemented on a variety of subsamples and the obtained features are merged to create a more stable subset.

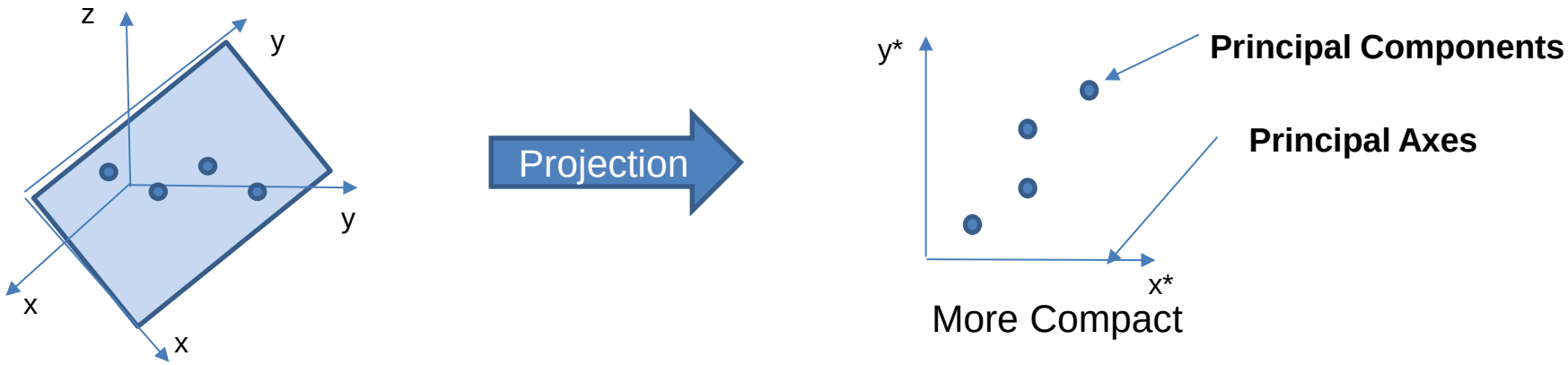
FEATURE EXTRACTION

Feature extraction is the process of transforming or projecting a space composing of many dimensions into a space of fewer dimensions. It is a transformation of the features that sometimes is not reversible because some information is lost in the process of dimensionality reduction. A new set of features is built from the original feature set.

Examples: extraction of contours in images, extraction of digrams from a text, extraction of phonemes from recording of spoken text, etc.

FEATURE EXTRACTION STRATEGIES: **PRINCIPLE COMPONENTS ANALYSIS**

The main linear technique is called **Principle Components Analysis**. This technique refers to the reduction of higher vector spaces to lower orders through projection. It can be used to visualize the dataset through compact representation and compression of dimensions.



DIFFERENCES BETWEEN FEATURE SELECTION AND FEATURE EXTRACTION

$$a + b + c + d = e$$

$$ab = a + b \quad \textit{Feature Extraction}$$

$$ab + c + d = e$$

$$c = 0 \quad \textit{Feature Selection}$$

$$ab + d = 0$$

If you can equate *ab* = *a* + *b*, making a representation of two variables into one, you are using **Feature Extraction** to reduce the number of variables.

If *c* was equal to 0, it would be irrelevant, therefore it could be removed from the equation. By doing so, you would be using **Feature Selection** because you would be selecting only the relevant variables and leaving out the irrelevant ones.

DIFFERENCES BETWEEN FEATURE SELECTION AND FEATURE EXTRACTION

Method	Main concept	Pros	Cons
Feature Extraction	Summarize the dataset by creating linear combinations of the features	Preserves the original, relative distance between covers latent structure, objects	Not sufficient enough in the existing of a huge number of irrelevant features
Feature Selection	A sublist of relevant features can be selected depending on defined criteria	Strong against irrelevant features	Latent structure does not cover

SUMMARY

- **Dimension Reduction** is the process of reducing the number of variables/features in review and can be divided into two subcategories called **Feature Selection** and **Feature Extraction**.

- **Feature selection** is the process of selecting a subset of relevant features or variables.

- **Feature extraction** is the process of transforming or projecting a space composing of many dimensions into a space of fewer dimensions.

REFERENCES

Machine Learning - Dimensionality Reduction - Feature Extraction & Selection - YouTube. (n.d.). Retrieved July 5, 2021, from https://www.youtube.com/watch?v=AU_hBML2H1c&list=RDCMUCoSiBUY5b9ixwuKfS99_rSw&start_radio=1&t=283s&ab_channel=CognitiveClass

What is the exact difference between feature selection and dimensional reduction? | Data Science and Machine Learning | Kaggle. (n.d.). Retrieved July 4, 2021, from https://www.kaggle.com/general/59245

Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*, *1*(2), 56–70. https://doi.org/10.38094/jastt1224

REFERENCE TO AUTHORS

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

0000-0002-2988-196X

Ana Luísa Sousa

- PhD student in Information System and Tecnologies
- Research Collaborator of the Algoritmi Research Center

0000-0001-5731-3583

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

0000-0003-2326-2153

REFERENCE TO AUTHORS

António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0001-6457-0756

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-4121-6169

Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-1819-7051

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Computer Vision Expert”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

DRIVES

Development and Research on Innovative Vocational Education Skills

Thank you for your attention

DRIVES project is project under **The Blueprint for Sectoral Cooperation on Skills in Automotive Sector**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

Co-funded by the Erasmus+ Programme of the European Union

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.