UNIVERSIDAD
NACIONAL
DE COLOMBIA

CLASIFICACIÓN Y RECONOCIMIENTO DE PATRONES

REDUCCION DE LA DIMENSIONALIDAD

JOHN W. BRANCH
Profesor Titular
Departamento de Ciencias de la Computación y de la Decisión
Director del Grupo de I+D en Inteligencia Artificial – GIDIA

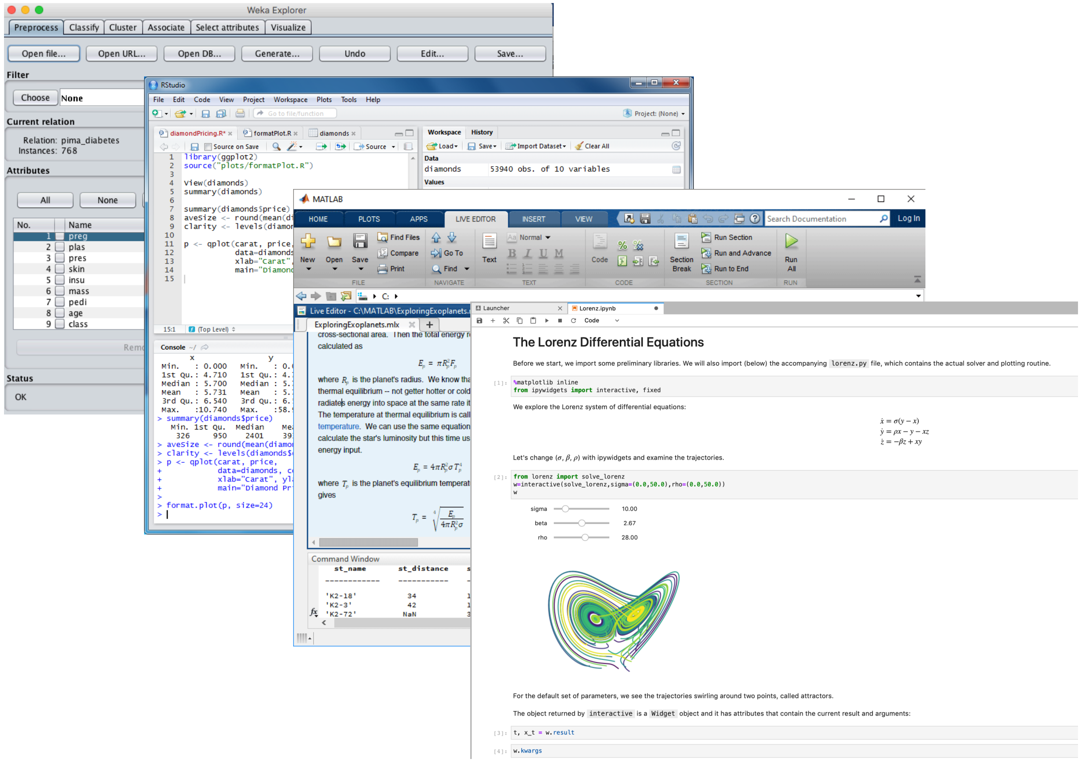
PEDRO ATENCIO ORTIZ
Profesor Invitado
Facultad de Ingenierías
Instituto Tecnológico Metropolitano

Agenda

1. Data processing Tools.
2. Dimensionality reduction using Python and Scikit:
    1. Feature Selection.
        1. The Knapsack problem.
        2. Statistical approach (Filters).
            1. Variance criteria.
            2. Statistical tests.
        3. Wrappers and Embedded models.
            1. Wrappers and search.
            2. Recursive elimination.
            3. Select from model.
    2. Feature Extraction.
        1. Feature transformation.
        2. Principal Component Analysis (PCA).
            1. SVD vs SKLearn.
            2. Non-Linear PCA (Kernel).
3. Conclusions.
4. Classwork.

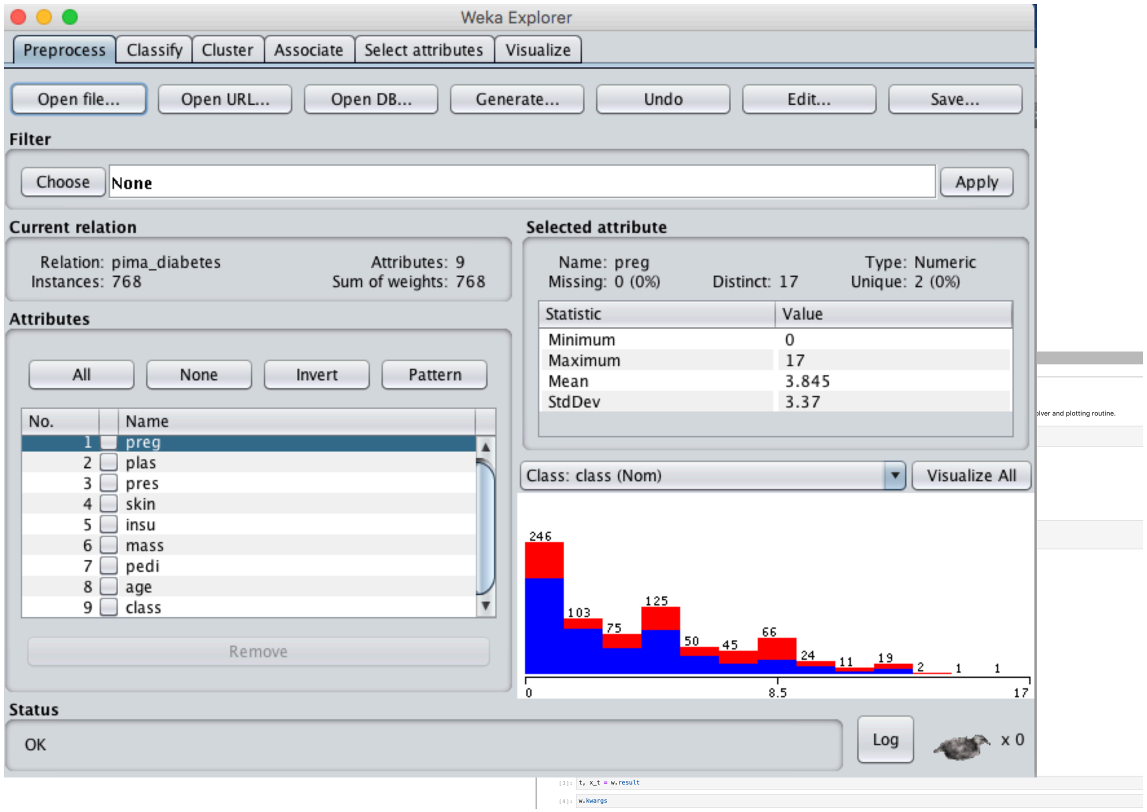
1. Data Processing Tools

- Weka
- R+RStudio
- MATLAB
- Python+Jupyter+Scikit



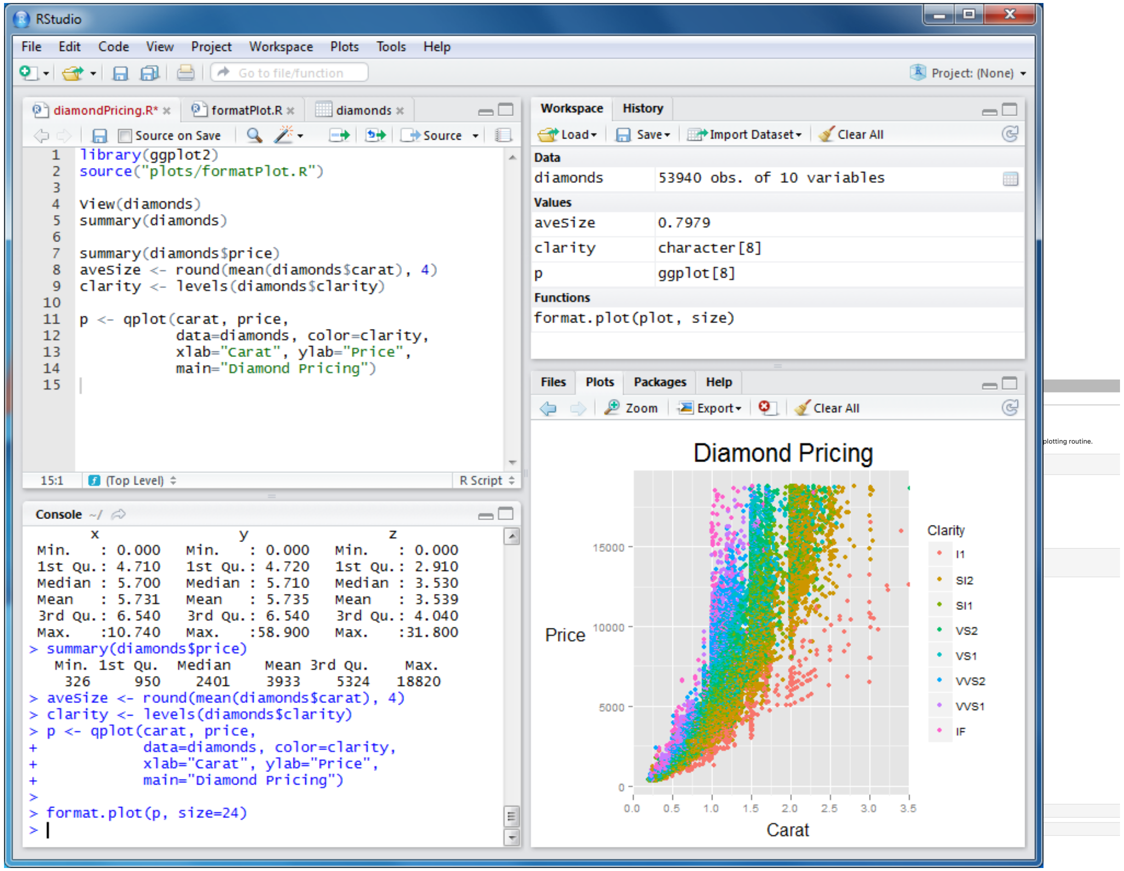
1. Data Processing Tools

- **Weka**
- R+RStudio
- MATLAB
- Python+Jupyter+Scikit



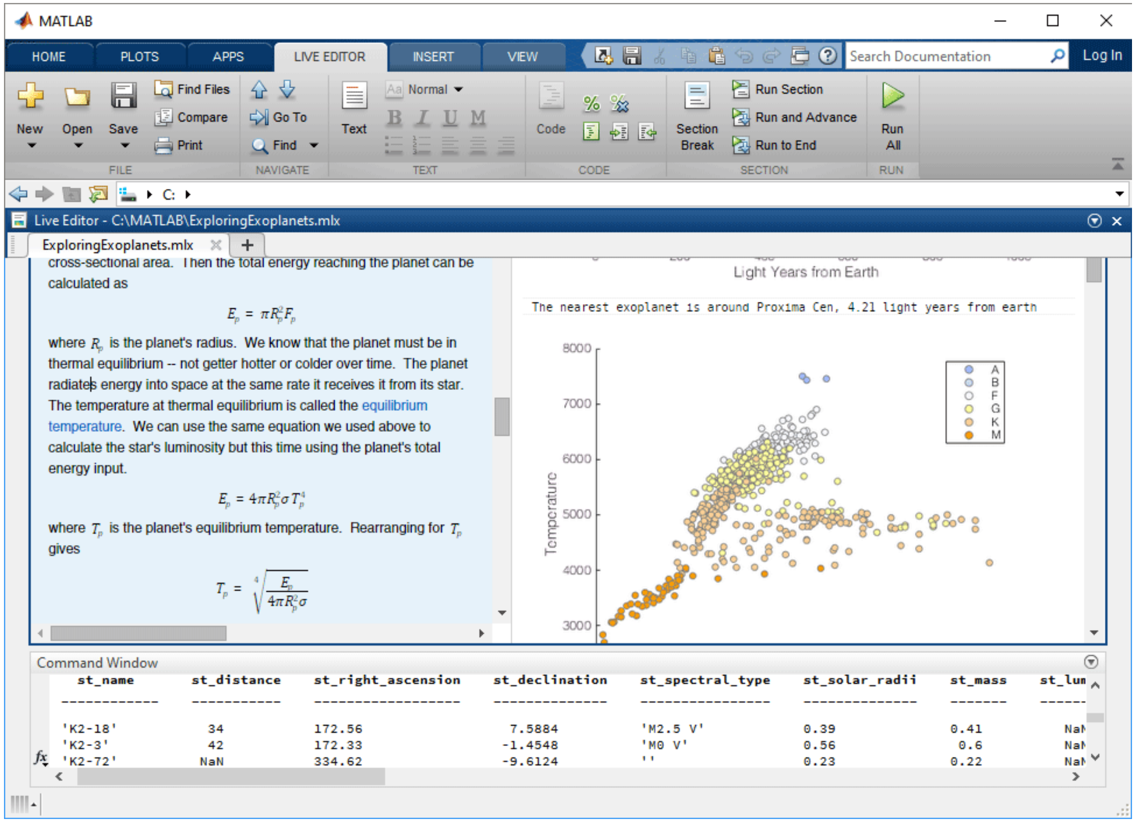
1. Data Processing Tools

- Weka
- **R+RStudio**
- MATLAB
- Python+Jupyter+Scikit



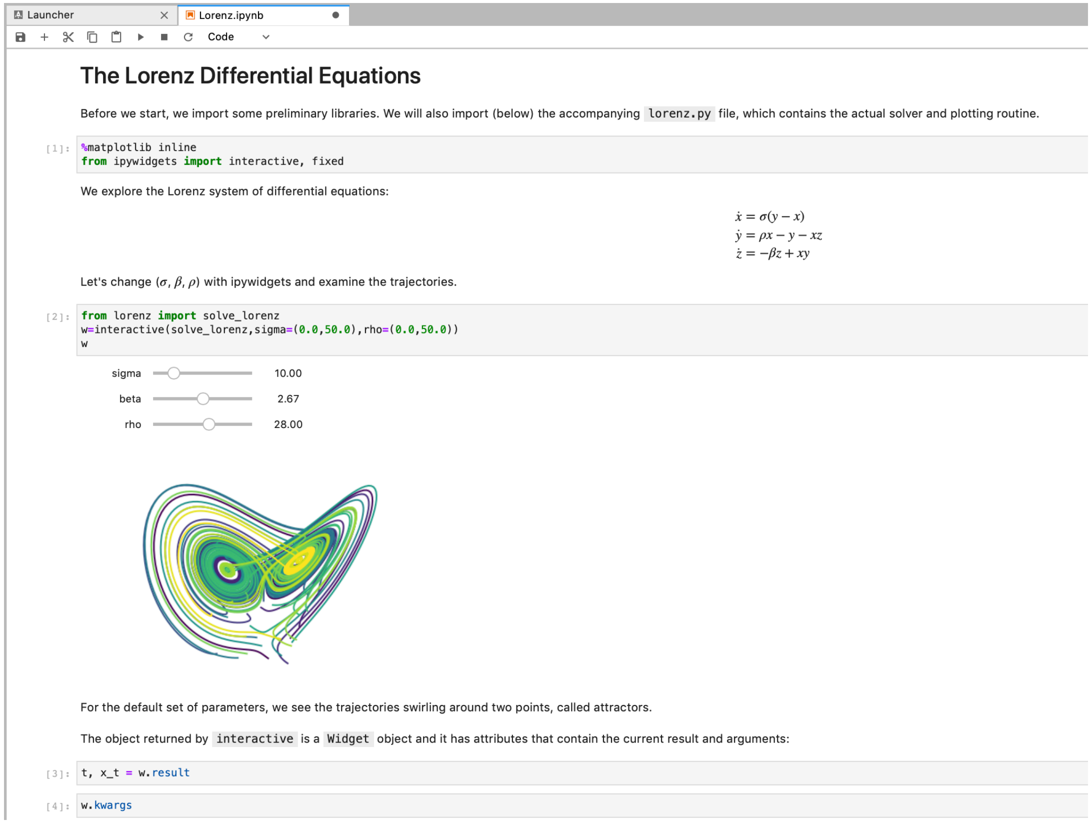
1. Data Processing Tools

- Weka
- R+RStudio
- **MATLAB**
- Python+Jupyter+Scikit



1. Data Processing Tools

- Weka
- R+RStudio
- MATLAB
- **Python+Jupyter +Scikit**

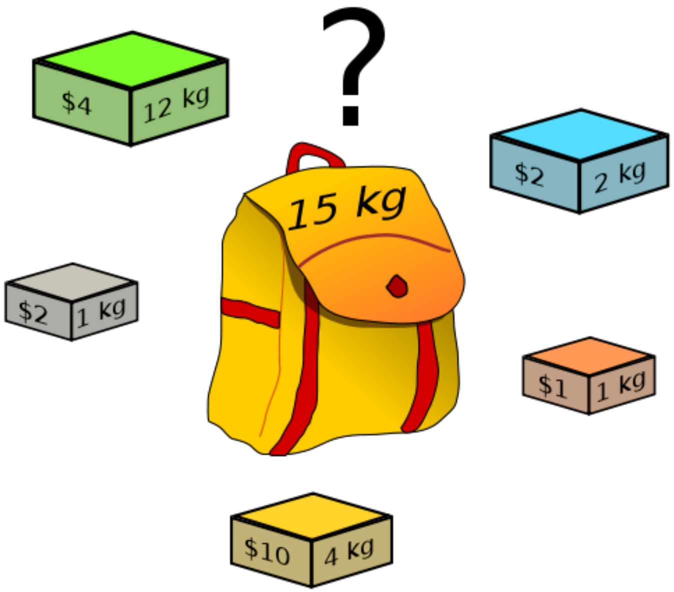


2. Dimensionality Reduction Using Python and Scikit

2.1. Feature Selection

2.1.1 The Knapsack Problem

The feature selection problem can be understood as the Knapsack problem:



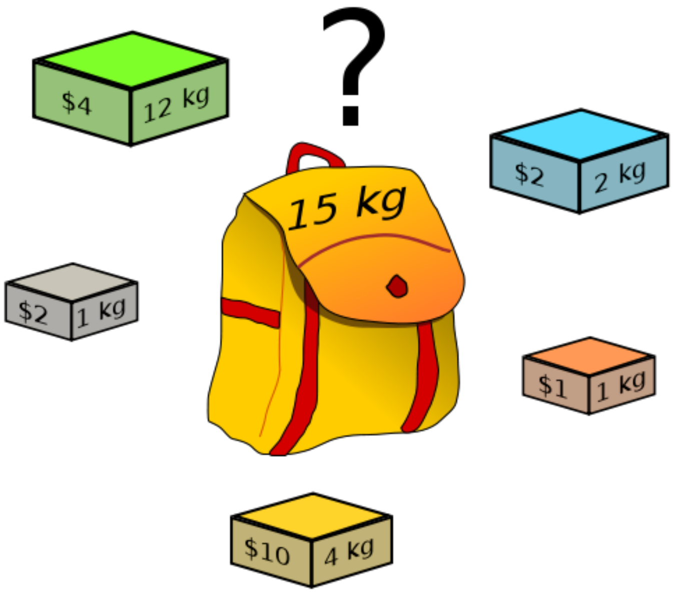
$$maximize \sum_{i=1}^{n} R(v_i x_i)$$

$$s.t. \sum_{i=1}^{n} |v_i|_0 \leq l$$

2.1. Feature Selection

2.1.1 The Knapsack Problem

The feature selection problem can be understood as the Knapsack problem:



Problem performance

features

$$maximize \sum_{i=1}^{n} R(v_i x_i)$$

Binary array

$$s.t. \sum_{i=1}^{n} |v_i|_0 \leq l$$

max features

num of selected features

2.1. Feature Selection

2.1.2. Statistical Approach (Filters)

A score is assigned to each feature according to a statistical measure, for example:

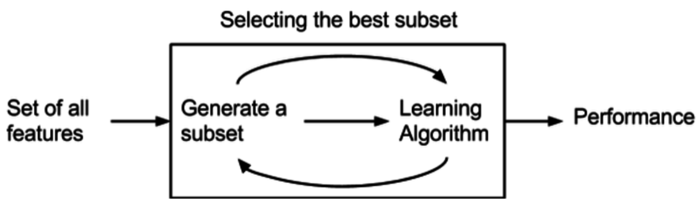
- Variance σ^2
- Chi - Squared χ^2
- Anova.
- Entropy.
- etc.

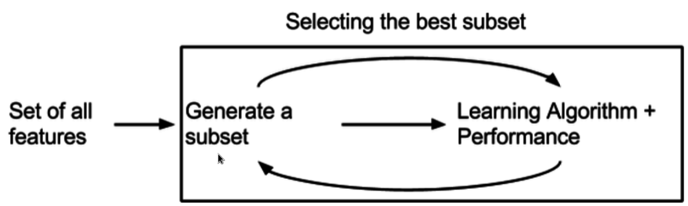
2.1. Feature Selection

2.1.3. Wrappers and Embedded Models

In simple words both approaches consists in:

- **Wrappers**: A predictive model is used to evaluate different combinations of features using a search method, for example, SFS, BFS, etc. Once, some combinations a tried, best subset of features is selected.
- **Embedded Models**: Use two predictive models, one to evaluate subset performance and one to LEARN which subset is best.





2.2. Feature Extraction

2.2.1. Feature Transformation

Feature transformation consists in obtaining set of features G from a lineal or non-linear mapping of an input set X.

$$G = f(X) = \alpha x_1 + \alpha x_2 + \alpha x_3 + \ldots + \alpha x_n$$

$$s.t. \, |G| < |X|$$

2.2. Feature Extraction

2.2.1. Feature Transformation

Feature transformation consists in obtaining set of features G from a lineal or non-linear mapping of an input set X.

Mapping function

New features

$$G = f(X) = \alpha x_1 + \alpha x_2 + \alpha x_3 + \ldots + \alpha x_n$$

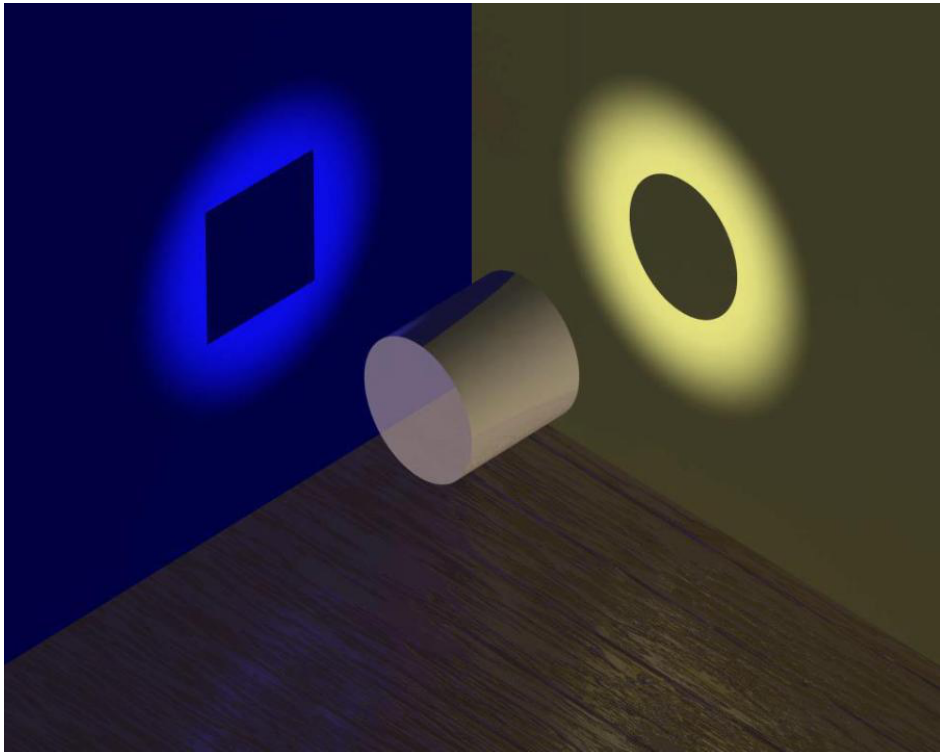
Original features

$$s.t.\ |G| < |X|$$

2.2. Feature Extraction

2.2.2. Principal Component Analysis

PCA consists in projecting a set of data over an hyperspace composed by principal dimensions of data.

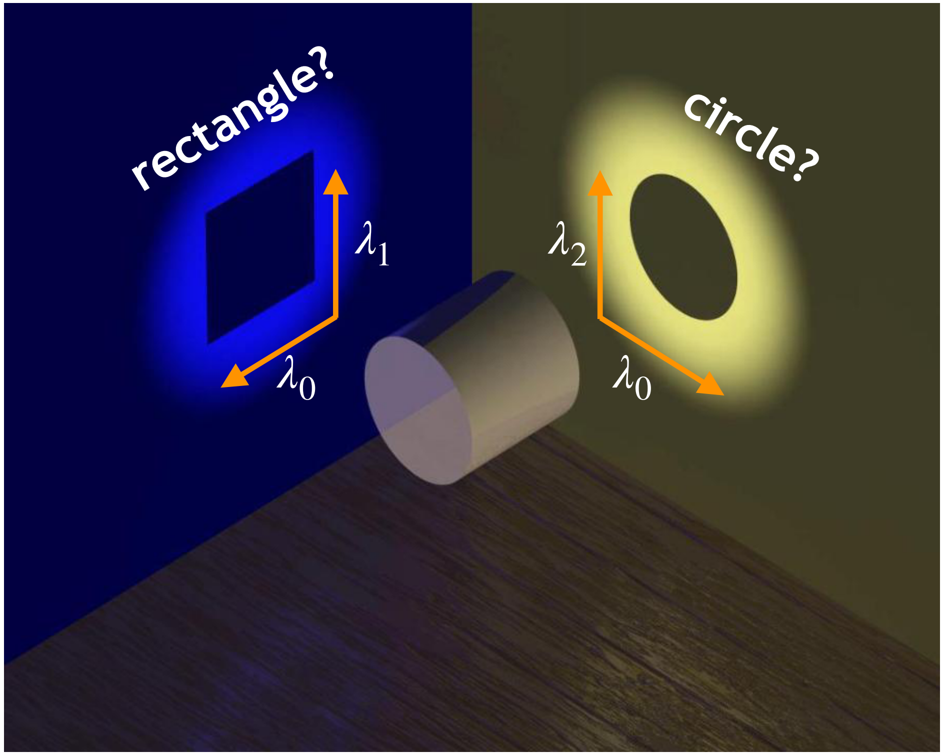


2.2. Feature Extraction

2.2.2. Principal Component Analysis

PCA consists in projecting a set of data over an hyperspace composed by principal dimensions of data.

Which $\lambda_i \in \Lambda$ are best?



2.2. Feature Extraction

2.2.2. Principal Component Analysis

Matrix factorization techniques can be used to extract Eigenvalues or principal components of a dataset. Single Value Decomposition is a matrix factorization technique:

$$X = U \cdot \sum \cdot V^T$$

Principal Components Matrix

2.2. Feature Extraction

2.2.2. Principal Component Analysis

Matrix factorization techniques can be used to extract Eigenvalues or principal components of a dataset. Single Value Decomposition is a matrix factorization technique:

$$X = U \cdot \sum \cdot V^T$$

Principal Components Matrix

Once V is extracted, we can project matrix X over one or more Eigenvalues or principal components and reduce original dimension of X:

$$X_{new} = X \cdot V'$$

Subset of V

3. Conclusions

There are many approaches that can be applied in order to reduce dimensionality which will be dependent of the problem we are trying to solve. Other some popular techniques are:

- Multidimensional Scaling (MDS) —> distance reduction.
- Isomap —> graph connections and geodesic preservation.
- t-SNE —> Embedding.
- LDA —> Classifier that learns discriminative axes.

4. Classwork

Work over next problems, using previous code as a base:

1. Load diabetes regression dataset using **load_diabetes()** from sklearn.datasets
    1. Tabulate data X and y.
    2. Apply a feature selection approach to obtain best 3 features.
    3. Plot using **plt.plot()** 3 plots: (x0, y), (x1, y), (x2, y).

2. Load breast_cancer classification dataset using **load_breast_cancer()** from sklearn.datasets:
    1. Tabulate data X and y.
    2. Apply a feature extraction approach to obtain 2 features.
    3. Plot using **plt.scatter()** to obtain a figure where can be observed elements and classes.

UNIVERSIDAD
NACIONAL
DE COLOMBIA