Coordinating and integRating state-of-the-art Earth Observation Activities in the regions of North Africa, Middle East and Balkans and Developing Links with GEO related intiatives toward GEOSS

A brief introduction to Machine Learning (ML)

Capacities and Skills: Towards the provision of EO services in the Balkans & the MENA region

T4.2 – Improved Food Security and Water Extremes Management

June 14th – A webinar by i-BEC and TAU

I-BEC INTERBALKAN ENVIRONMENT CENTER

European Commission

GEO GROUP ON EARTH OBSERVATIONS

Table of Contents

1. What is ML?
    1. Basic definitions
    2. Classification
    3. Regression
2. On the validation of models
    1. Performance measures
    2. Overfitting
    3. Properly validating the model
3. Big Data
    1. The curse of dimensionality
    2. PCA

Table of Contents

1. What is ML?
   1. Basic definitions
   2. Classification
   3. Regression
2. On the validation of models
   1. Performance measures
   2. Overfitting
   3. Properly validating the model
3. Big Data
   1. The curse of dimensionality
   2. PCA

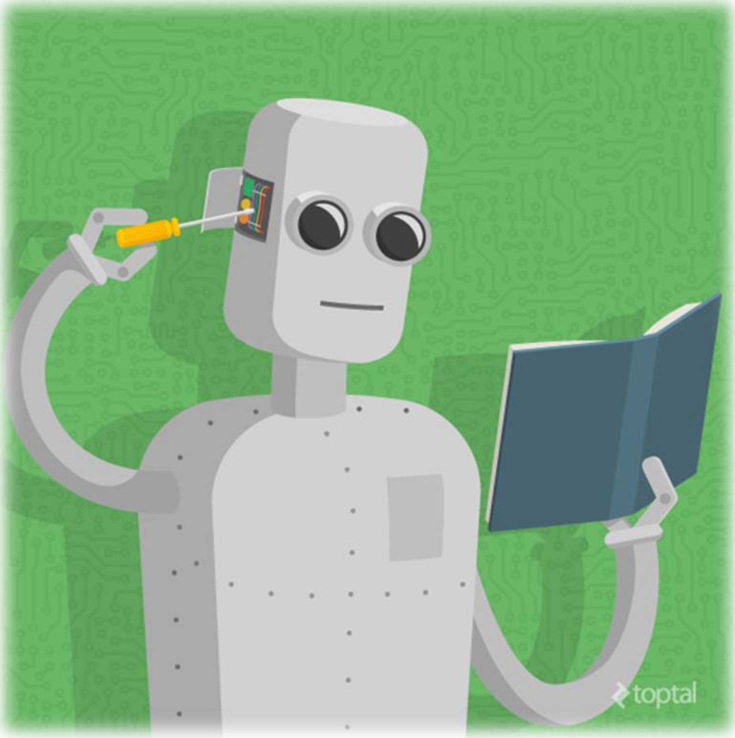
Definition of ML (I)

“The field of study that gives computers the ability to learn without being explicitly programmed”

Arthur Samuel, 1959

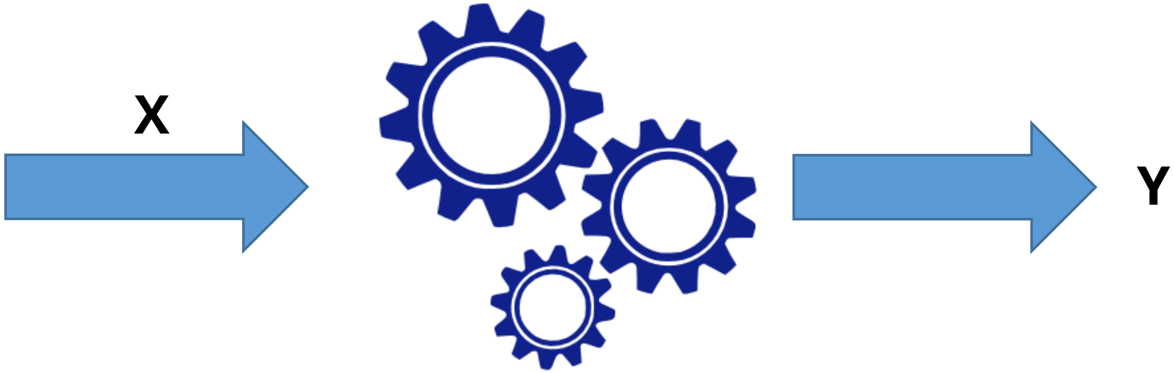
“A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”

Tom Mitchell, 1997



A general definition

- **Dataset:** A set of N data points, containing input **X** and output **Y**
- **X**: Comprised of M input attributes (or features, or independent variables) for each datum
- **Y:** One labeled output (classification: discrete, regression: continuous)
- **The ML problem:** Find a mapping function *f* that predicts **Y** given **X**



Examples of ML

- Image processing
  - Given a set of images, identify the ones containing dogs
  - Given a photo of book’s page, identify the text
- Text processing
  - Given a tweet, identify if the user is happy or sad
  - Given an e-mail, identify if the mail is spam
- Data mining
  - Given a credit card transaction, identify if it is fraudulent
  - Given a soil spectral curve, identify the soil’s OM %

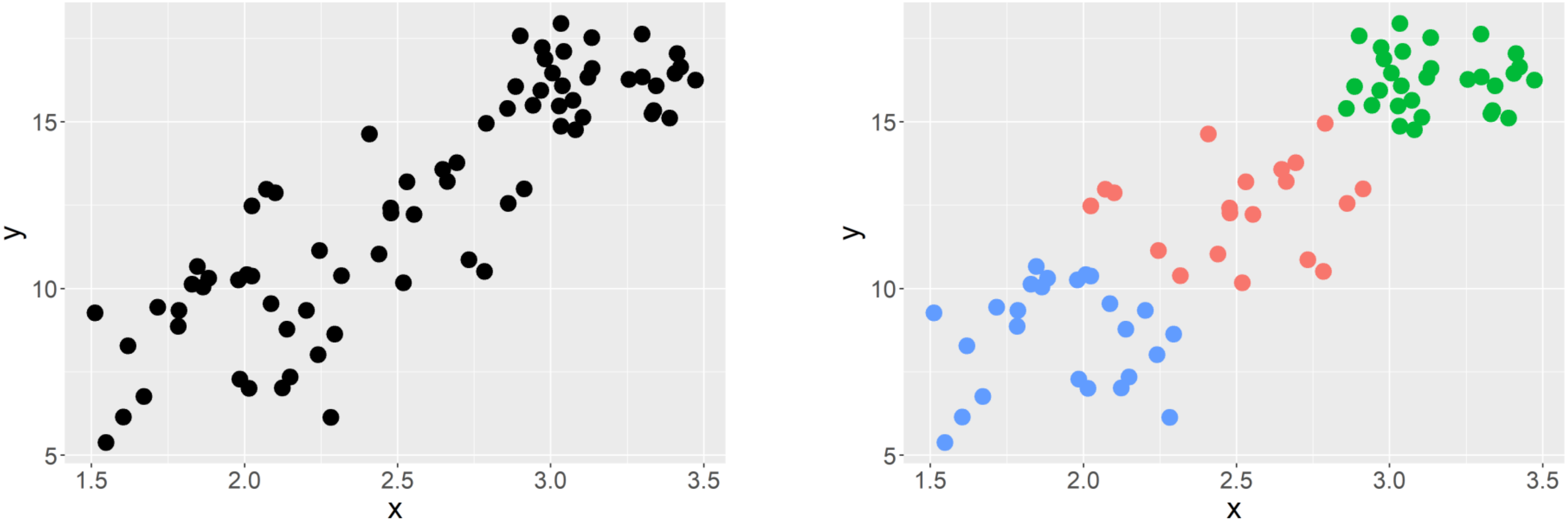
Supervised ML

ML learns from labeled data – a supervisor labels each datum

Image	Does it have a dog?
	Yes
	No
	Yes

Unsupervised ML

ML learns from unlabeled data



Two dimensional data without output label can
be clustered into similar groups

Classification

- Each datum is labeled using a discrete categorical value
- Examples:
  - If an image contains a dog or not (“Yes” or “No”)
  - If a tweet is positive or negative (“Positive” or “Negative”)
  - Class of a soil sample (“Vertisol”, “Inceptisol”, “Entisol”, …)

Eggs	Legs	Tail	Predator	Animal
No	4	Yes	Yes	Bear
Yes	2	Yes	No	Flamingo
Yes	2	Yes	Yes	Hawk

Regression

- The label of each datum is a continuous value
- Examples:
  - OM of a soil sample using its spectrum
  - Yield of a produce
  - Salary of an employee

Position	Experience	Skill	Salary
Marketing	7 years	1.5	27,000
Developer	3 years	2.0	25,000
Manager	10 years	2.1	35,000

Linear regression

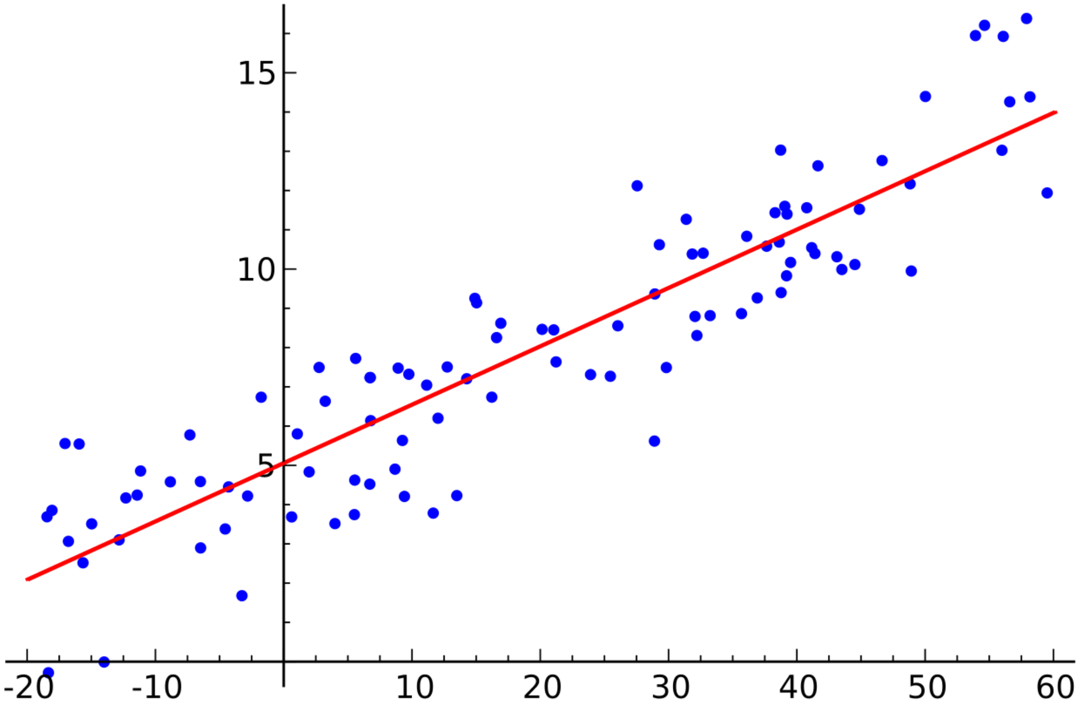
X = input feature
Y = real output

$$\hat{Y} = \lambda X + b$$

The ML algorithm must find λ and b, such that:

$$SSE = \sum_i \left(Y - \hat{Y}\right)^2$$

is minimized



ML models for regression

- Statistical regression
  - Linear regression
  - Ordinary least squares regression (OLSR)
  - Multivariate adaptive regression splines (MARS)
  - Regression + regularization (LASSO, Ridge, Elastic net)
- Instance based algorithms
- Decision Trees
- Support Vector Regression
- Artificial Neural Networks
- … and many more …

Table of Contents

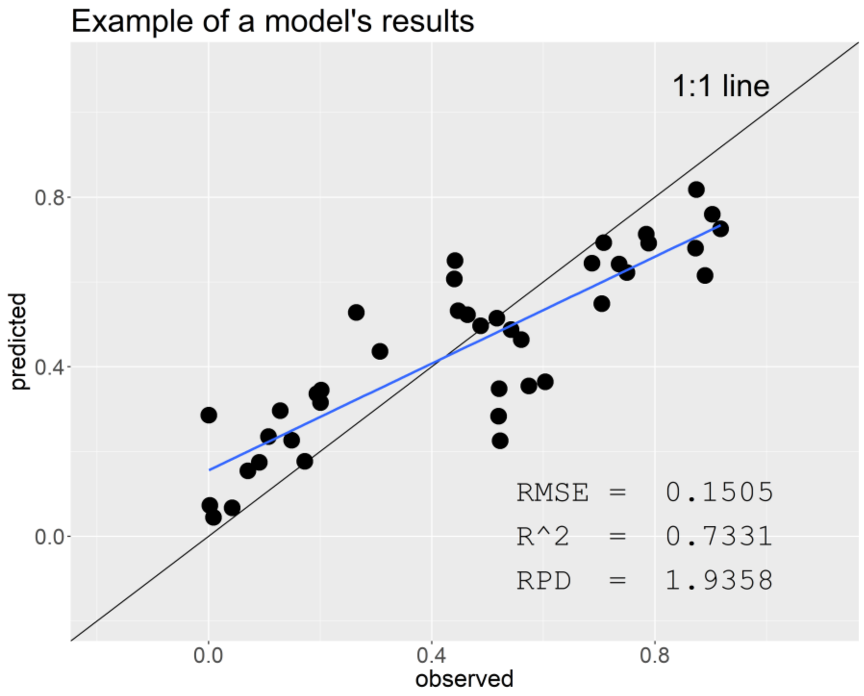
1. What is ML?
   1. Basic definitions
   2. Classification
   3. Regression
2. On the validation of models
   1. Performance measures
   2. Overfitting
   3. Properly validating the model
3. Big Data
   1. The curse of dimensionality
   2. PCA

Performance measures for regression (I)

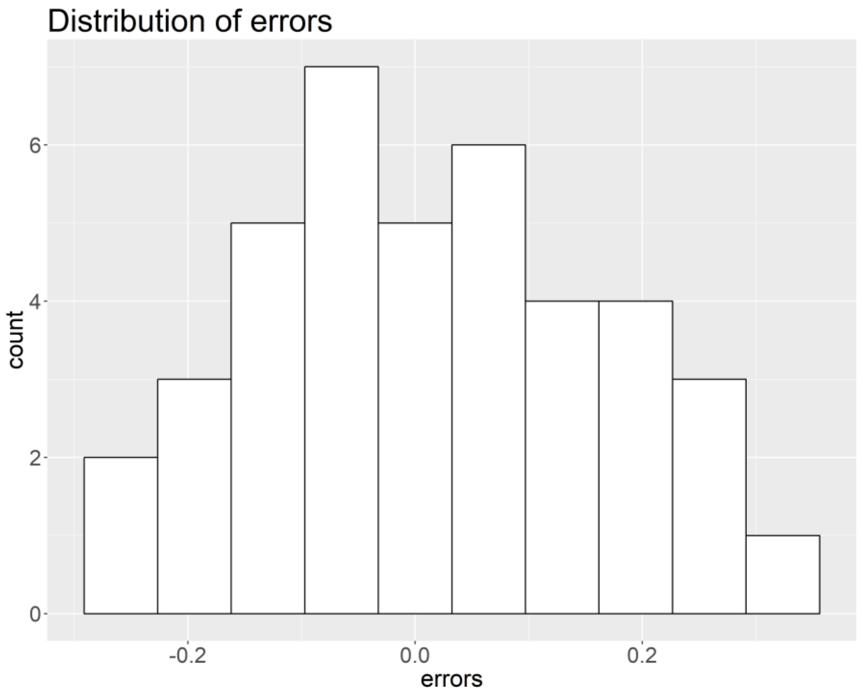
- $\mathrm{SSE} = \sum_i \left(Y - \hat{Y}\right)^2$ [sum of squared errors]
- $\mathrm{MSE} = \frac{\mathrm{SSE}}{N}$ [mean squared error]
- $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$ [root mean squared error]
- $R^2 = 1 - \frac{\mathrm{MSE}}{VARIANCE(Y)}$ [coefficient of determination]

- It is important to always test the model using unseen, independent data! The performance measures between the data used for training ("training data") and the data used for testing ("testing data") might be different!

Performance measures for regression (II)



Example of a model's results



Distribution of errors

Overfitting

- When a ML models learns from data, it is possible to over adjust its parameters to increase the performance on the training set slightly. However, when we test the model on the new (unseen) data, the performance of the model might be poorer than expected.
- We need to ensure that the model is not only accurate in the training data, but also predictive!

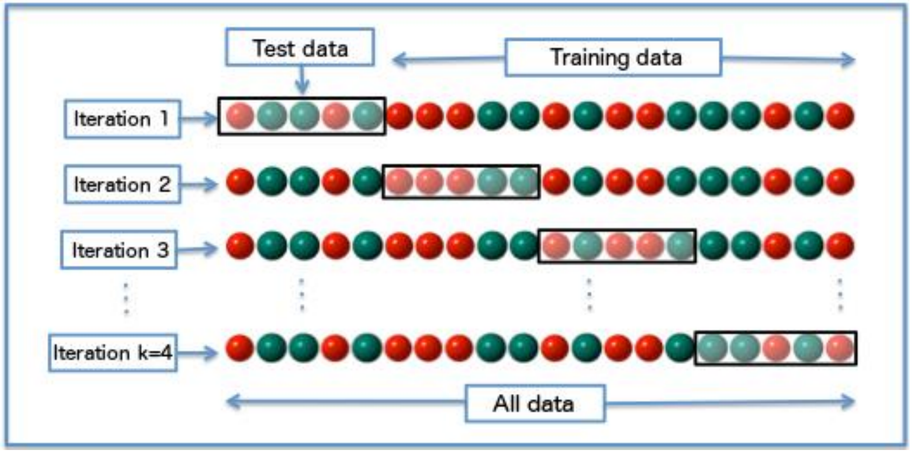
Validating the model

1. Divide the available data into training, validation and test sets
2. Select architecture and/or training parameters
3. Train the model using the training set
4. Evaluate the model using the validation set
5. Repeat steps 2 through 4 using different architectures and/or training parameters
6. Select the best model and train it using data from both training and validation set
7. Assess this final model using the test set

For cross-validation, repeat steps 3 and 4 for each of the k folds

Splitting the dataset

- To split the dataset into training, validation and test sets, you can use different procedures:
  - Randomly split the data
  - Stratify the data
  - Pre-cluster the data
  - Use a sample selection algorithm (e.g. the Kennard-Stone algorithm)
  - Use cross-validation



k-cross-validation with k = 4

Table of Contents

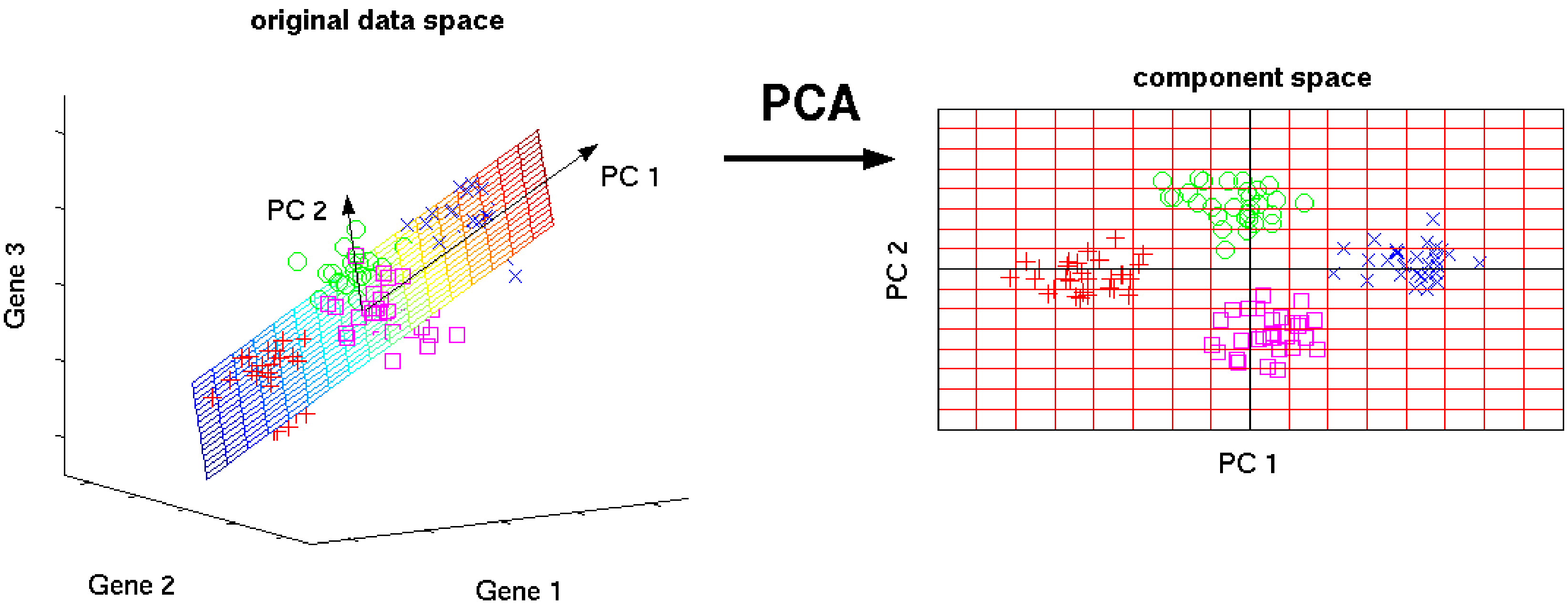
1. What is ML?
   1. Basic definitions
   2. Classification
   3. Regression
2. On the validation of models
   1. Performance measures
   2. Overfitting
   3. Properly validating the model
3. Big Data
   1. The curse of dimensionality
   2. PCA

Operating on Big Data

- **Big Data**: High in Volume, Variety, Variability and Veracity
- The larger the dataset, the higher the training time and (potentially) the more complex the model
- Example: The GEO-CRADLE SSL

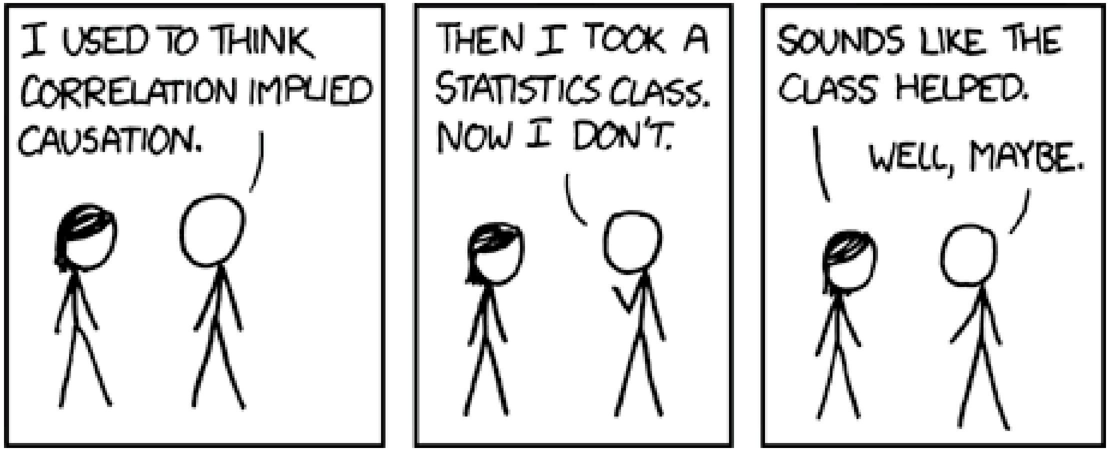
- **The curse of dimensionality**
  - The problem: When a dataset is comprised of data with hundreds of features, the data become sparse and appear to be dissimilar in many ways, which might hinder the performance of some ML methods.
  - The solution: Use a dimensionality reduction technique, which performs feature extraction, or feature selection

Principal Components



Feature Selection

Feature selection attempts to find the best subset of initial features, which correlate the most with the output variable. Common algorithms are the mRMR (minimum redundancy maximum relevance) algorithm, the CFS (correlation feature selection) algorithm, and metaheuristics methods (filter or wrapper methods).


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Questions?