Introduction to Machine Learning

Introduction to Data Science

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-II

Outline

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Key Concepts in Machine Learning

Machine Learning

- **Machine learning** is a method of data analysis that automates analytical model building.
- It is a **branch** of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with `minimal human intervention`.

- **Supervised Learning**: The model is trained on labeled data.
- **Unsupervised Learning**: The model is trained on unlabeled data.
- **Reinforcement Learning**: The model learns by interacting with an environment.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Key Concepts in Machine Learning

Machine Learning

- **Machine learning** is a method of data analysis that automates analytical model building.
- It is a **branch** of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with `minimal human intervention`.

- **Supervised Learning**: The model is trained on labeled data.
- **Unsupervised Learning**: The model is trained on unlabeled data.
- **Reinforcement Learning**: The model learns by interacting with an environment.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Key Concepts in Machine Learning

Machine Learning

- **Machine learning** is a method of data analysis that automates analytical model building.
- It is a **branch** of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with `minimal human intervention`.

- **Supervised Learning**: The model is trained on labeled data.
- **Unsupervised Learning**: The model is trained on unlabeled data.
- **Reinforcement Learning**: The model learns by interacting with an environment.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Python Tools for Machine Learning

Python Tools

- **NumPy**: A library for numerical computing.
- **Pandas**: A library for data manipulation and analysis.
- **Matplotlib**: A library for data visualization.
- **Scikit-learn**: A library for machine learning.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Typical Machine Learning Problems

- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying relationships between variables.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

The Machine Learning Workflow

- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.
- **Model Deployment**: Putting the model into production.

Examining the Data

- **Data Exploration**: Understanding the data.
- **Data Cleaning**: Preparing the data.
- **Feature Engineering**: Creating new features.
- **Feature Selection**: Selecting the most important features.
- **Data Preprocessing**: Preparing the data for modeling.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Examining the Data

- **Data Exploration**: Understanding the data.
- **Data Cleaning**: Preparing the data.
- **Feature Engineering**: Creating new features.
- **Feature Selection**: Selecting the most important features.
- **Data Preprocessing**: Preparing the data for modeling.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Examining the Data

- **Data Exploration**: Understanding the data.
- **Data Cleaning**: Preparing the data.
- **Feature Engineering**: Creating new features.
- **Feature Selection**: Selecting the most important features.
- **Data Preprocessing**: Preparing the data for modeling.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Examining the Data

- **Data Exploration**: Understanding the data.
- **Data Cleaning**: Preparing the data.
- **Feature Engineering**: Creating new features.
- **Feature Selection**: Selecting the most important features.
- **Data Preprocessing**: Preparing the data for modeling.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Examining the Data

- **Data Exploration**: Understanding the data.
- **Data Cleaning**: Preparing the data.
- **Feature Engineering**: Creating new features.
- **Feature Selection**: Selecting the most important features.
- **Data Preprocessing**: Preparing the data for modeling.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

K-Nearest Neighbors Classification

- **K-Nearest Neighbors** is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
- It is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation.

Algorithmic Bias

- **Algorithmic bias** is a systematic error in a model that results in unfair outcomes.
- It can be caused by biased training data, biased algorithms, or biased decision-making.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

Introduction to Supervised Machine Learning

Definition

- **Supervised learning** is a type of machine learning where the model is trained on labeled data.
- It involves training a model to map input data to output data based on example input-output pairs.

Overfitting and Underfitting

Overfitting

Overfitting occurs when a model learns the training data too well and performs poorly on new data.

Underfitting

Underfitting occurs when a model is too simple to capture the underlying structure of the data.

Overfitting and Underfitting

Overfitting

Overfitting occurs when a model learns the training data too well and performs poorly on new data.

Underfitting

Underfitting occurs when a model is too simple to capture the underlying structure of the data.

Supervised Learning Datasets

- **Training Dataset**: The data used to train the model.
- **Validation Dataset**: The data used to tune the model hyperparameters.
- **Test Dataset**: The data used to evaluate the model performance.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Supervised Learning Datasets

- **Training Dataset**: The data used to train the model.
- **Validation Dataset**: The data used to tune the model hyperparameters.
- **Test Dataset**: The data used to evaluate the model performance.

Supervised Learning Datasets

- **Training Dataset**: The data used to train the model.
- **Validation Dataset**: The data used to tune the model hyperparameters.
- **Test Dataset**: The data used to evaluate the model performance.

K-Nearest Neighbors: Classification and Regression

- **K-Nearest Neighbors (KNN)** is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
- It can be used for both classification and regression tasks.
- For classification, the output is the class label of the majority of the k-nearest neighbors.
- For regression, the output is the average of the k-nearest neighbors.

K-Nearest Neighbors: Classification and Regression

- **K-Nearest Neighbors (KNN)** is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
- It can be used for both classification and regression tasks.
- For classification, the output is the class label of the majority of the k-nearest neighbors.
- For regression, the output is the average of the k-nearest neighbors.

Linear Regression with Least Squares

Linear Regression

- **Linear regression** is a type of regression analysis used for predicting the value of a continuous dependent variable.
- It works by finding the line that best fits the data.

Least Squares

Least squares is a method for finding the best-fitting line by minimizing the sum of the squared differences between the predicted and actual values.

Linear Regression with Least Squares

Linear Regression

- **Linear regression** is a type of regression analysis used for predicting the value of a continuous dependent variable.
- It works by finding the line that best fits the data.

Least Squares

Least squares is a method for finding the best-fitting line by minimizing the sum of the squared differences between the predicted and actual values.

Ridge & Lasso

Ridge Regression

Ridge regression is a type of linear regression that includes a penalty term to prevent overfitting. It works by adding a regularization term to the least squares objective function.

Lasso Regression

Lasso regression is a type of linear regression that includes a penalty term to prevent overfitting. It works by adding a regularization term to the least squares objective function.

Ridge & Lasso

Ridge Regression

Ridge regression is a type of linear regression that includes a penalty term to prevent overfitting. It works by adding a regularization term to the least squares objective function.

Lasso Regression

Lasso regression is a type of linear regression that includes a penalty term to prevent overfitting. It works by adding a regularization term to the least squares objective function.

Polynomial Regression

Polynomial Regression

- **Polynomial regression** is a type of regression analysis that models the relationship between the independent and dependent variables as an nth-degree polynomial.
- It can capture non-linear relationships between the variables.

Logistic Regression

Logistic Regression

- **Logistic regression** is a type of regression analysis used for predicting the outcome of a categorical dependent variable.
- It is used for binary classification tasks, where the output is a probability between 0 and 1.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Cross-Validation

- **Cross-validation** is a technique for assessing the performance of a model.
- It involves splitting the data into multiple subsets, training the model on some subsets, and evaluating it on others.
- Common cross-validation **techniques** include k-fold cross-validation and leave-one-out cross-validation.
- Cross-validation helps to reduce overfitting and provides a more accurate estimate of the model's performance.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

One-Hot Encoding

One-Hot Encoding

- **One-hot encoding** is a technique for converting categorical variables into numerical variables.
- It creates a binary vector for each `category`, with a 1 for the *category* and 0s for all other categories.

Data Leakage

- **Data leakage** occurs when information from the test set is inadvertently used to train the model.
- It can lead to overfitting and inflated performance metrics.
- Common sources of **data leakage** include target leakage, train-test contamination, and information leakage.
- To prevent **data leakage**, it is important to carefully separate the training and test data and avoid using information from the test set during training.

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Support Vector Machines

- **Support vector machines** are a type of machine learning model that can be used for both classification and regression tasks.
- They work by finding the hyperplane that best separates the data into different classes.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Decision Trees

- **Decision trees** are a type of machine learning model that can be used for both classification and regression tasks.
- They work by recursively partitioning the data into subsets based on the values of the features.

Naive Bayes Classifier

- The **naive Bayes classifier** is a simple probabilistic classifier based on Bayes' theorem.
- It assumes that the features are conditionally independent given the class label.

Random Forest

- **Random forest** is an ensemble learning method that combines multiple decision trees to create a `strong predictive model`.
- It works by building multiple trees and averaging their predictions to reduce overfitting.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Gradient Boosted Decision Trees

- **Gradient boosted decision trees** are an ensemble learning method that combines multiple decision trees and gradient descedent optimization to create a `strong predictive model`.
- They work by building trees sequentially, with each tree correcting the errors of the previous trees.

Neural Networks

- **Neural networks** are a type of machine learning model inspired by the human brain.
- They consist of layers of interconnected nodes that process input data and produce output data.

Outline

Model Evaluation & Selection

- **Model Evaluation**: Assessing the performance of a model.
- **Model Selection**: Choosing the best model for the task.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Confusion Matrices

Definition

- A **confusion matrix** is a table that summarizes the performance of a classification model.
- It shows the number of true positives, true negatives, false positives, and false negatives.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Basic Evaluation Metrics

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of true positives among all positive predictions.
- **Recall**: The proportion of true positives among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Basic Evaluation Metrics

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of true positives among all positive predictions.
- **Recall**: The proportion of true positives among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall.

Basic Evaluation Metrics

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of true positives among all positive predictions.
- **Recall**: The proportion of true positives among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Basic Evaluation Metrics

- **Accuracy**: The proportion of correct predictions.
- **Precision**: The proportion of true positives among all positive predictions.
- **Recall**: The proportion of true positives among all actual positives.
- **F1 Score**: The harmonic mean of precision and recall.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Classifier Decision Metrics

- **ROC Curve**: A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve**: A plot of precision against recall.
- **AUC-ROC**: The area under the ROC curve.
- **AUC-PR**: The area under the precision-recall curve.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Classifier Decision Metrics

- **ROC Curve**: A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve**: A plot of precision against recall.
- **AUC-ROC**: The area under the ROC curve.
- **AUC-PR**: The area under the precision-recall curve.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Classifier Decision Metrics

- **ROC Curve**: A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve**: A plot of precision against recall.
- **AUC-ROC**: The area under the ROC curve.
- **AUC-PR**: The area under the precision-recall curve.

Classifier Decision Metrics

- **ROC Curve**: A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve**: A plot of precision against recall.
- **AUC-ROC**: The area under the ROC curve.
- **AUC-PR**: The area under the precision-recall curve.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Multi-Class Evaluation

- **Macro-Averaging**: The average of the evaluation metrics for each class.
- **Micro-Averaging**: The evaluation metrics calculated on the aggregate confusion matrix.
- **Weighted-Averaging**: The average of the evaluation metrics weighted by the number of samples in each class.
- **One-vs-All**: A strategy for multi-class classification that trains a separate binary classifier for each class.

Multi-Class Evaluation

- **Macro-Averaging**: The average of the evaluation metrics for each class.
- **Micro-Averaging**: The evaluation metrics calculated on the aggregate confusion matrix.
- **Weighted-Averaging**: The average of the evaluation metrics weighted by the number of samples in each class.
- **One-vs-All**: A strategy for multi-class classification that trains a separate binary classifier for each class.

Multi-Class Evaluation

- **Macro-Averaging**: The average of the evaluation metrics for each class.
- **Micro-Averaging**: The evaluation metrics calculated on the aggregate confusion matrix.
- **Weighted-Averaging**: The average of the evaluation metrics weighted by the number of samples in each class.
- **One-vs-All**: A strategy for multi-class classification that trains a separate binary classifier for each class.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Multi-Class Evaluation

- **Macro-Averaging**: The average of the evaluation metrics for each class.
- **Micro-Averaging**: The evaluation metrics calculated on the aggregate confusion matrix.
- **Weighted-Averaging**: The average of the evaluation metrics weighted by the number of samples in each class.
- **One-vs-All**: A strategy for multi-class classification that trains a separate binary classifier for each class.

Regression Evaluation

- **Mean Squared Error**: The average of the squared differences between the predicted and actual values.
- **Mean Absolute Error**: The average of the absolute differences between the predicted and actual values.
- **R-Squared**: The proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared**: A modified version of `R-squared` that adjusts for the number of predictors in the model.
- **Root Mean Squared Error**: The square root of the `mean squared error`.

Regression Evaluation

- **Mean Squared Error**: The average of the squared differences between the predicted and actual values.
- **Mean Absolute Error**: The average of the absolute differences between the predicted and actual values.
- **R-Squared**: The proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared**: A modified version of `R-squared` that adjusts for the number of predictors in the model.
- **Root Mean Squared Error**: The square root of the `mean squared error`.

Regression Evaluation

- **Mean Squared Error**: The average of the squared differences between the predicted and actual values.
- **Mean Absolute Error**: The average of the absolute differences between the predicted and actual values.
- **R-Squared**: The proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared**: A modified version of `R-squared` that adjusts for the number of predictors in the model.
- **Root Mean Squared Error**: The square root of the `mean squared error`.

Regression Evaluation

- **Mean Squared Error**: The average of the squared differences between the predicted and actual values.
- **Mean Absolute Error**: The average of the absolute differences between the predicted and actual values.
- **R-Squared**: The proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared**: A modified version of `R-squared` that adjusts for the number of predictors in the model.
- **Root Mean Squared Error**: The square root of the `mean squared error`.

Regression Evaluation

- **Mean Squared Error**: The average of the squared differences between the predicted and actual values.
- **Mean Absolute Error**: The average of the absolute differences between the predicted and actual values.
- **R-Squared**: The proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared**: A modified version of `R-squared` that adjusts for the number of predictors in the model.
- **Root Mean Squared Error**: The square root of the `mean squared error`.

Model Calibration

- **Model calibration** is the process of adjusting the output of a model to match the true probability distribution.
- It is important for models that *output probabilities*, such as logistic regression and support vector machines.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/data-science-introduction*

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS