An introduction to Machine Learning with Python (STK353)

Mahdi Salehi
salehi2sms@gmail.com

16 October 2023

Section 1

Initial concepts

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are

- **Tabular data**: Sales transactions, customer information, financial records, stock prices.
- **Sensor data**: Temperature readings, GPS coordinates, machine sensor data.

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are

- **Tabular data**: Sales transactions, customer information, financial records, stock prices.
- **Sensor data**: Temperature readings, GPS coordinates, machine sensor data.

2. Unstructured Data, in contrast, does not have a predefined structure or format. It lacks a consistent organization and can be more challenging to process and analyze using traditional methods. Examples of unstructured data include:

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are
   - **Tabular data**: Sales transactions, customer information, financial records, stock prices.
   - **Sensor data**: Temperature readings, GPS coordinates, machine sensor data.
2. Unstructured Data, in contrast, does not have a predefined structure or format. It lacks a consistent organization and can be more challenging to process and analyze using traditional methods. Examples of unstructured data include:
   - **Text data**: Emails, social media posts, news articles, customer reviews, medical records.

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are

- **Tabular data**: Sales transactions, customer information, financial records, stock prices.
- **Sensor data**: Temperature readings, GPS coordinates, machine sensor data.

2. Unstructured Data, in contrast, does not have a predefined structure or format. It lacks a consistent organization and can be more challenging to process and analyze using traditional methods. Examples of unstructured data include:

- **Text data**: Emails, social media posts, news articles, customer reviews, medical records.
- **Multimedia data**: Images, videos, audio recordings, satellite images.

Structured data vs unstructured data

1. Structured data is highly organized and follows a predefined format. It is typically stored in databases or spreadsheets, where each data element has a specific data type and is organized into rows and columns. Some examples of structured data are

- **Tabular data**: Sales transactions, customer information, financial records, stock prices.
- **Sensor data**: Temperature readings, GPS coordinates, machine sensor data.

2. Unstructured Data, in contrast, does not have a predefined structure or format. It lacks a consistent organization and can be more challenging to process and analyze using traditional methods. Examples of unstructured data include:

- **Text data**: Emails, social media posts, news articles, customer reviews, medical records.
- **Multimedia data**: Images, videos, audio recordings, satellite images.
- **Web data**: Web pages, HTML documents, web logs and etc.

What is a data set?

A data set is a collection of data objects and their attributes.

- Attribute is also known as variable, field, characteristic, or feature.
- Object is also known as record, point, case, sample, entity, or instance.

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

Figure 1: A dataset.

Explanatory variables vs. Response variable

Explanatory variables (also known as input variables, features, or independent variables) and response variables (also known as output variables or dependent variables) play distinct roles in the modeling process. Here's an explanation of each:

1. Explanatory Variables:
   - They are the input or independent variables used to predict or explain the behavior of the response variable.

Explanatory variables vs. Response variable

Explanatory variables (also known as input variables, features, or independent variables) and response variables (also known as output variables or dependent variables) play distinct roles in the modeling process. Here's an explanation of each:

1. Explanatory Variables:

- They are the input or independent variables used to predict or explain the behavior of the response variable.
- Explanatory variables can be of different types, such as numerical (e.g., age, temperature), categorical (e.g., gender, color), or even more complex types like text or images.

Explanatory variables vs. Response variable

Explanatory variables (also known as input variables, features, or independent variables) and response variables (also known as output variables or dependent variables) play distinct roles in the modeling process. Here's an explanation of each:

1. Explanatory Variables:

- They are the input or independent variables used to predict or explain the behavior of the response variable.
- Explanatory variables can be of different types, such as numerical (e.g., age, temperature), categorical (e.g., gender, color), or even more complex types like text or images.

2. Response Variable:

- The response variable is the output or dependent variable that the machine learning model aims to predict or explain based on the values of the explanatory variables.

2. Response Variable:

- The response variable is the output or dependent variable that the machine learning model aims to predict or explain based on the values of the explanatory variables.
- The response variable can also be of different types, depending on the nature of the problem. For example, it could be a continuous variable (e.g., predicting house prices) or a categorical variable (e.g., classifying images into different categories).

2. Response Variable:
    - The response variable is the output or dependent variable that the machine learning model aims to predict or explain based on the values of the explanatory variables.
    - The response variable can also be of different types, depending on the nature of the problem. For example, it could be a continuous variable (e.g., predicting house prices) or a categorical variable (e.g., classifying images into different categories).

The goal of the machine learning model is to learn a mapping or relationship between the explanatory variables and the response variable, enabling it to make predictions or classifications on new data.

Section 2

Machine learning: What, why & when

What is Machine Learning?

The foundation of machine learning lies in the idea that computers can learn patterns and make predictions based on data.

Coined by Samuel in 1959, the term machine learning (ML) was given to the field of study of the development of algorithms and models that can automatically learn patterns and make predictions or decisions based on data. It is a subset of artificial intelligence (AI) that aims to enable machines to improve their performance over time by using experience.

It involves the following key components:

What is Machine Learning?

The foundation of machine learning lies in the idea that computers can learn patterns and make predictions based on data.

Coined by Samuel in 1959, the term machine learning (ML) was given to the field of study of the development of algorithms and models that can automatically learn patterns and make predictions or decisions based on data. It is a subset of artificial intelligence (AI) that aims to enable machines to improve their performance over time by using experience.

It involves the following key components:

- Data: Machine learning algorithms require data to learn from. This data can be labeled (with known outcomes) or unlabeled (without known outcomes).

What is Machine Learning?

The foundation of machine learning lies in the idea that computers can learn patterns and make predictions based on data.

Coined by Samuel in 1959, the term machine learning (ML) was given to the field of study of the development of algorithms and models that can automatically learn patterns and make predictions or decisions based on data. It is a subset of artificial intelligence (AI) that aims to enable machines to improve their performance over time by using experience.

It involves the following key components:

- Data: Machine learning algorithms require data to learn from. This data can be labeled (with known outcomes) or unlabeled (without known outcomes).
- Algorithms: Machine learning algorithms are used to analyze and process the data, extract patterns, and make predictions or decisions. These algorithms can be categorized into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

What is Machine Learning?

The foundation of machine learning lies in the idea that computers can learn patterns and make predictions based on data.

Coined by Samuel in 1959, the term machine learning (ML) was given to the field of study of the development of algorithms and models that can automatically learn patterns and make predictions or decisions based on data. It is a subset of artificial intelligence (AI) that aims to enable machines to improve their performance over time by using experience.

It involves the following key components:

- Data: Machine learning algorithms require data to learn from. This data can be labeled (with known outcomes) or unlabeled (without known outcomes).
- Algorithms: Machine learning algorithms are used to analyze and process the data, extract patterns, and make predictions or decisions. These algorithms can be categorized into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.
- Models: Machine learning models are the result of training algorithms on data. These models can be used to make predictions or classify new data.

Why do we use Machine learning?

The goal of machine learning is to build models that can generalize well to new dataset and make accurate predictions or decisions.

Why do we use Machine learning?

The goal of machine learning is to build models that can generalize well to new dataset and make accurate predictions or decisions.

Machine learning offers several benefits and applications:

Why do we use Machine learning?

The goal of machine learning is to build models that can generalize well to new dataset and make accurate predictions or decisions.

Machine learning offers several benefits and applications:

- **Complex Data Analysis:** Machine learning algorithms can handle large and complex datasets, extracting valuable insights and patterns that may not be apparent to humans. They can process vast amounts of data quickly and efficiently.

Why do we use Machine learning?

The goal of machine learning is to build models that can generalize well to new dataset and make accurate predictions or decisions.

Machine learning offers several benefits and applications:

- **Complex Data Analysis:** Machine learning algorithms can handle large and complex datasets, extracting valuable insights and patterns that may not be apparent to humans. They can process vast amounts of data quickly and efficiently.
- **Automation and Efficiency:** By automating tasks that would otherwise require manual effort, machine learning can save time and resources. It can perform repetitive tasks at scale, freeing up human resources for more complex tasks.

Why do we use Machine learning?

The goal of machine learning is to build models that can generalize well to new dataset and make accurate predictions or decisions.

Machine learning offers several benefits and applications:

- **Complex Data Analysis:** Machine learning algorithms can handle large and complex datasets, extracting valuable insights and patterns that may not be apparent to humans. They can process vast amounts of data quickly and efficiently.
- **Automation and Efficiency:** By automating tasks that would otherwise require manual effort, machine learning can save time and resources. It can perform repetitive tasks at scale, freeing up human resources for more complex tasks.

- **Anomaly Detection and Fraud Prevention:** Machine learning algorithms can detect anomalies and patterns indicative of fraudulent activities, such as credit card fraud or cybersecurity threats. They can identify deviations from normal behavior and raise alerts.

- **Anomaly Detection and Fraud Prevention:** Machine learning algorithms can detect anomalies and patterns indicative of fraudulent activities, such as credit card fraud or cybersecurity threats. They can identify deviations from normal behavior and raise alerts.
- **Natural Language Processing:** Machine learning techniques are used in natural language processing tasks like speech recognition, sentiment analysis, language translation, and chatbots. They enable machines to understand and generate human language.

- **Anomaly Detection and Fraud Prevention:** Machine learning algorithms can detect anomalies and patterns indicative of fraudulent activities, such as credit card fraud or cybersecurity threats. They can identify deviations from normal behavior and raise alerts.
- **Natural Language Processing:** Machine learning techniques are used in natural language processing tasks like speech recognition, sentiment analysis, language translation, and chatbots. They enable machines to understand and generate human language.
- **Medical Diagnosis and Healthcare:** Machine learning models can assist in medical diagnosis by analyzing patient data and identifying patterns that may indicate diseases or conditions. They can aid in early detection and personalized treatment plans.

Machine learning vs. Statistical learning

While machine learning and statistical learning share common techniques and goals, they differ in their focus, approach, and emphasis. Machine learning emphasizes practical applications, scalability, and prediction accuracy, while statistical learning focuses on understanding relationships, making inferences, and assessing uncertainty. Both fields have their strengths and are often used in complementary ways to solve a wide range of problems.

Machine learning vs. Statistical learning

While machine learning and statistical learning share common techniques and goals, they differ in their focus, approach, and emphasis. Machine learning emphasizes practical applications, scalability, and prediction accuracy, while statistical learning focuses on understanding relationships, making inferences, and assessing uncertainty. Both fields have their strengths and are often used in complementary ways to solve a wide range of problems.

1. **Focus and Purpose:**

Machine learning vs. Statistical learning

While machine learning and statistical learning share common techniques and goals, they differ in their focus, approach, and emphasis. Machine learning emphasizes practical applications, scalability, and prediction accuracy, while statistical learning focuses on understanding relationships, making inferences, and assessing uncertainty. Both fields have their strengths and are often used in complementary ways to solve a wide range of problems.

1. **Focus and Purpose:**
   - Machine Learning: Machine learning models are designed to make the most accurate predictions possible. It emphasizes the development of practical, scalable algorithms that can handle large datasets and complex problems. Machine learning is often used in real-world applications such as image recognition, natural language processing, and recommendation systems.

Machine learning vs. Statistical learning

While machine learning and statistical learning share common techniques and goals, they differ in their focus, approach, and emphasis. Machine learning emphasizes practical applications, scalability, and prediction accuracy, while statistical learning focuses on understanding relationships, making inferences, and assessing uncertainty. Both fields have their strengths and are often used in complementary ways to solve a wide range of problems.

1. **Focus and Purpose:**
   - Machine Learning: Machine learning models are designed to make the most accurate predictions possible. It emphasizes the development of practical, scalable algorithms that can handle large datasets and complex problems. Machine learning is often used in real-world applications such as image recognition, natural language processing, and recommendation systems.
   - Statistical Learning: Statistical models, on the other hand, are designed for inference about the relationships between variables. Many statistical models can make predictions, but predictive accuracy is not their strength. Statistical learning focuses on understanding and modeling complex relationships in data using statistical techniques. It aims to make inferences and draw conclusions about the underlying processes that generate the data.

2. **Approach and Methodology:**

2. **Approach and Methodology:**

- Machine Learning: Machine learning often employs a more algorithmic and computational approach, utilizing techniques such as neural networks, decision trees, support vector machines, and deep learning. It emphasizes optimization, generalization, and scalability. Machine learning algorithms are designed to automatically learn patterns and relationships from data, often with less emphasis on the underlying statistical assumptions.

2. **Approach and Methodology:**

- Machine Learning: Machine learning often employs a more algorithmic and computational approach, utilizing techniques such as neural networks, decision trees, support vector machines, and deep learning. It emphasizes optimization, generalization, and scalability. Machine learning algorithms are designed to automatically learn patterns and relationships from data, often with less emphasis on the underlying statistical assumptions.
- Statistical Learning: Statistical learning focuses on understanding the underlying statistical properties of the data and making inferences based on those properties. It utilizes techniques such as linear regression, logistic regression, Bayesian methods, and hypothesis testing. Statistical learning often involves making assumptions about the data distribution and relies on statistical theory to draw conclusions and make predictions.

3. **Data and Assumptions:**

3. **Data and Assumptions:**
   - Machine Learning: Machine learning algorithms can handle a wide range of data types, including structured, unstructured, and high-dimensional data. They often require large amounts of labeled data for training and can handle noisy or incomplete data. Machine learning algorithms are generally more flexible and can adapt to different data distributions without strict assumptions.

3. **Data and Assumptions:**

- Machine Learning: Machine learning algorithms can handle a wide range of data types, including structured, unstructured, and high-dimensional data. They often require large amounts of labeled data for training and can handle noisy or incomplete data. Machine learning algorithms are generally more flexible and can adapt to different data distributions without strict assumptions.
- Statistical Learning: Statistical learning techniques often assume that the data follows a specific probability distribution or statistical model. They are more suited for structured data and often require assumptions about the data's distribution, independence, and linearity. Statistical learning methods may struggle with high-dimensional or unstructured data and may require more careful preprocessing and feature engineering.

4. **Emphasis on Prediction vs. Inference:**

4. **Emphasis on Prediction vs. Inference:**

- Machine Learning: Machine learning places a stronger emphasis on prediction accuracy and generalization to new data. It focuses on developing models that can make accurate predictions or decisions on new instances. Machine learning algorithms often prioritize optimizing performance metrics such as accuracy.

4. **Emphasis on Prediction vs. Inference:**

- Machine Learning: Machine learning places a stronger emphasis on prediction accuracy and generalization to new data. It focuses on developing models that can make accurate predictions or decisions on new instances. Machine learning algorithms often prioritize optimizing performance metrics such as accuracy.
- Statistical Learning: Statistical learning places a stronger emphasis on understanding the underlying relationships and making inferences about the data. It aims to draw conclusions about the significance and effect of variables, estimate parameters, and assess uncertainty. Statistical learning methods often focus on hypothesis testing, confidence intervals, and interpreting the coefficients of the model.

5. **Higher accuracy in prediction vs. interpretability**

5. **Higher accuracy in prediction vs. interpretability**
   - Statistical learning models, such as linear regression or logistic regression, often have a clear and interpretable structure. They provide coefficients or parameter estimates that can be directly interpreted to understand the relationship between input variables and the target variable.
   - These models are often based on well-established statistical principles and assumptions, allowing researchers to make inferences about the significance and effect of variables.
   - Statistical models are commonly used in scientific research, where interpretability is crucial for understanding the underlying processes and drawing meaningful conclusions.
   - Machine learning models, such as deep neural networks or ensemble methods, are often more complex and have a larger number of parameters. This complexity can make them less interpretable compared to statistical models.
   - The focus on accuracy may lead to models that are more black-box in nature, making it challenging to understand how the model arrives at its predictions.

The first trade-off: Interpretability and Accuracy:

- There is often a trade-off between interpretability and accuracy in machine learning. As models become more complex and flexible, they can achieve higher accuracy but at the cost of reduced interpretability.

The first trade-off: Interpretability and Accuracy:

- There is often a trade-off between interpretability and accuracy in machine learning. As models become more complex and flexible, they can achieve higher accuracy but at the cost of reduced interpretability.
- In some cases, interpretability may be less important, such as when the primary goal is to achieve the highest possible accuracy in tasks like image recognition or natural language processing.

The first trade-off: Interpretability and Accuracy:

- There is often a trade-off between interpretability and accuracy in machine learning. As models become more complex and flexible, they can achieve higher accuracy but at the cost of reduced interpretability.
- In some cases, interpretability may be less important, such as when the primary goal is to achieve the highest possible accuracy in tasks like image recognition or natural language processing.
- However, in domains where interpretability is crucial, such as healthcare or finance, simpler and more interpretable models like decision trees or logistic regression may be preferred, even if they sacrifice some accuracy.

The first trade-off: Interpretability and Accuracy:

- There is often a trade-off between interpretability and accuracy in machine learning. As models become more complex and flexible, they can achieve higher accuracy but at the cost of reduced interpretability.
- In some cases, interpretability may be less important, such as when the primary goal is to achieve the highest possible accuracy in tasks like image recognition or natural language processing.
- However, in domains where interpretability is crucial, such as healthcare or finance, simpler and more interpretable models like decision trees or logistic regression may be preferred, even if they sacrifice some accuracy.

There is also another trade-off that should be take care of. You will see it in the sequel.

Various types of machine learning algorithms

Machine learning algorithms can be broadly categorized into three types:

Various types of machine learning algorithms

Machine learning algorithms can be broadly categorized into three types:

1. **Supervised Learning:** Here we have a clearly defined response variable (or variables) Y and a set of explanatory variables $\mathbf{X} = (X_1, \ldots, X_d)$ and the goal is to find the relationship between Y and $\mathbf{X}$ in order to perform statistical inference, prediction, etc. The response variable is like a teacher in this type of problems and it is used to supervise us and show us our mistakes until we get good at what we do. In supervised learning, when the response Y is a quantitative variable (i.e. it takes numerical values), we are dealing with a regression problem. And when the response Y is a qualitative or categorical variable (i.e. it takes finitely many discrete values, in other words, its values belongs to one of finitely many classes or categories), we are dealing with a classification problem.

2. **Unsupervised Learning:** Unsupervised learning involves training models on unlabeled data, where the algorithm learns to find patterns and structures in the data without any predefined labels. Indeed, here, there is no clearly defined objective and we often do not have a predefined response variable Y (we have only observed the input variables). So, nothing is there to supervise us and the ultimate goal is to explore the data and look for interesting patterns in the data. In these settings we may for example be interested in detecting patterns in the data. Usually, this is the starting point to perform a subsequent supervised learning. Clustering and dimensionality reduction (e.g. PCA) are common tasks in unsupervised learning.

3. **Reinforcement Learning:** Here the ultimate goal is to develop learning system that receives a reward signal and tries to learn to maximize the reward signal. It has been applied to problems such as game playing (playing chess with the computer), robotics, and autonomous driving. We do not consider this type of learning in this course but feel free to consult with the book Reinforcement Learning: An Introduction by Sutton and Barto (2018).

Section 3

Experimental design in machine learning

It refers to the process of setting up and conducting experiments to evaluate the performance of machine learning models. It involves various steps, including splitting the dataset into training and test sets, selecting appropriate performance metrics, and utilizing techniques like cross-validation. Let's explore each of these components:

It refers to the process of setting up and conducting experiments to evaluate the performance of machine learning models. It involves various steps, including splitting the dataset into training and test sets, selecting appropriate performance metrics, and utilizing techniques like cross-validation. Let's explore each of these components:

1. **Splitting the dataset:** To evaluate the performance of a machine learning model, it is crucial to divide the available dataset into two separate sets: the training set and the test set. The training set is used to train the model, while the test set is used to assess its performance on unseen data. The typical split is around 70-80% for training and 20-30% for testing, but this can vary depending on the size of the dataset.

It refers to the process of setting up and conducting experiments to evaluate the performance of machine learning models. It involves various steps, including splitting the dataset into training and test sets, selecting appropriate performance metrics, and utilizing techniques like cross-validation. Let's explore each of these components:

1. **Splitting the dataset:** To evaluate the performance of a machine learning model, it is crucial to divide the available dataset into two separate sets: the training set and the test set. The training set is used to train the model, while the test set is used to assess its performance on unseen data. The typical split is around 70-80% for training and 20-30% for testing, but this can vary depending on the size of the dataset.
2. **Performance metrics:** Performance metrics are used to measure how well a machine learning model performs on the test set. The choice of metrics depends on the specific task and the nature of the data. For classification tasks, common metrics include accuracy, precision, recall, F1 score, and area under the ROC curve. For regression tasks, metrics like mean squared error (MSE), mean absolute error (MAE), and R-squared are commonly used.

Cross-validation:

Cross-validation is a technique used to assess the performance of a model and mitigate the potential bias introduced by a single train-test split. It involves dividing the dataset into multiple subsets or "folds." The model is trained and evaluated multiple times, with each fold serving as the test set while the remaining folds are used for training. This helps to obtain a more robust estimate of the model's performance.

Cross-validation:

Cross-validation is a technique used to assess the performance of a model and mitigate the potential bias introduced by a single train-test split. It involves dividing the dataset into multiple subsets or "folds." The model is trained and evaluated multiple times, with each fold serving as the test set while the remaining folds are used for training. This helps to obtain a more robust estimate of the model's performance.

- **k-fold cross-validation:** In k-fold cross-validation, the dataset is divided into k equal-sized folds. The model is trained and evaluated k times, with each fold serving as the test set once. The performance metrics are then averaged across the k iterations to obtain a more reliable estimate of the model's performance.

Cross-validation:

Cross-validation is a technique used to assess the performance of a model and mitigate the potential bias introduced by a single train-test split. It involves dividing the dataset into multiple subsets or "folds." The model is trained and evaluated multiple times, with each fold serving as the test set while the remaining folds are used for training. This helps to obtain a more robust estimate of the model's performance.

- **k-fold cross-validation:** In k-fold cross-validation, the dataset is divided into k equal-sized folds. The model is trained and evaluated k times, with each fold serving as the test set once. The performance metrics are then averaged across the k iterations to obtain a more reliable estimate of the model's performance.
- **Leave-one-out cross-validation:** Leave-one-out cross-validation is a special case of k-fold cross-validation where k is equal to the number of samples in the dataset. Each sample is used as the test set once, while the remaining samples are used for training. It can be computationally expensive but provides an unbiased estimate of the model's performance.

The second trade-off: Overfiting vs. Underfiting vs Good fit

- Overfitting: Overfitting is like a student who memorizes the exact answers to specific practice questions but fails to understand the underlying concepts. When faced with new test questions that require applying those concepts in a different way, the student struggles. Similarly, in machine learning, an overfit model memorizes the training data too closely but fails to generalize well to new data.

- Underfitting: Underfitting is like a student who doesn't study enough and lacks a solid understanding of the subject. They may perform poorly on both practice questions and new test questions because they haven't grasped the fundamental concepts. In machine learning, an underfit model is too simplistic and fails to capture the underlying patterns in the data, resulting in poor performance on both the training data and new data.

- Good fit: A good fit is like a student who strikes the right balance between memorization and understanding. They study the materials, grasp the core concepts, and can apply them to a variety of questions. This student performs well on both practice questions and new test questions because they have a solid foundation. Similarly, in machine learning, a good fit model captures the underlying patterns in the data without being too specific or too simplistic, allowing it to generalize well to new data.

Section 4

More details on Supervised learning

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.
3. **Naive Bayes**: A probabilistic classifier that assumes independence between features given the class label. It is commonly used for text classification tasks.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.
3. **Naive Bayes**: A probabilistic classifier that assumes independence between features given the class label. It is commonly used for text classification tasks.
4. **Decision Trees**: A versatile algorithm that builds a tree-like model to make decisions based on feature values. It can be used for both classification and regression tasks.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.
3. **Naive Bayes**: A probabilistic classifier that assumes independence between features given the class label. It is commonly used for text classification tasks.
4. **Decision Trees**: A versatile algorithm that builds a tree-like model to make decisions based on feature values. It can be used for both classification and regression tasks.
5. **Support Vector Machines (SVM)**: A powerful algorithm that finds a hyperplane to separate different classes in the feature space. It can be used for both classification and regression tasks.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.
3. **Naive Bayes**: A probabilistic classifier that assumes independence between features given the class label. It is commonly used for text classification tasks.
4. **Decision Trees**: A versatile algorithm that builds a tree-like model to make decisions based on feature values. It can be used for both classification and regression tasks.
5. **Support Vector Machines (SVM)**: A powerful algorithm that finds a hyperplane to separate different classes in the feature space. It can be used for both classification and regression tasks.
6. **K-Nearest Neighbors (KNN)**: An algorithm that classifies new instances based on the majority vote of their k nearest neighbors in the feature space.

Some well-known algorithms

1. **Linear Regression**: A simple algorithm used for regression tasks that models the relationship between the input features and the continuous target variable.
2. **Logistic Regression**: A popular algorithm for binary classification tasks that models the relationship between the input features and the probability of belonging to a particular class.
3. **Naive Bayes**: A probabilistic classifier that assumes independence between features given the class label. It is commonly used for text classification tasks.
4. **Decision Trees**: A versatile algorithm that builds a tree-like model to make decisions based on feature values. It can be used for both classification and regression tasks.
5. **Support Vector Machines (SVM)**: A powerful algorithm that finds a hyperplane to separate different classes in the feature space. It can be used for both classification and regression tasks.
6. **K-Nearest Neighbors (KNN)**: An algorithm that classifies new instances based on the majority vote of their k nearest neighbors in the feature space.
7. **Neural Networks**: Deep learning algorithms that consist of interconnected layers of artificial neurons. They can handle complex patterns and are used for various tasks, including image recognition, natural language processing, and speech recognition.

Linear Regression

Linear regression: What, Why and When?

Linear Regression

Linear regression: What, Why and When?

- Linear regression is a statistical technique used to model the linear relationship between a **response (dependent) variable** and one or more explanatory (independent) variables. The explanatory variables here are also called **predictors**.

Linear Regression

Linear regression: What, Why and When?

- Linear regression is a statistical technique used to model the linear relationship between a **response (dependent) variable** and one or more explanatory (independent) variables. The explanatory variables here are also called **predictors**.
- The response variable as well as the predictors here must all be continuous variables.

Linear Regression

Linear regression: What, Why and When?

- Linear regression is a statistical technique used to model the linear relationship between a **response (dependent) variable** and one or more explanatory (independent) variables. The explanatory variables here are also called **predictors**.
- The response variable as well as the predictors here must all be continuous variables.
- It assumes a linear relationship between the variables, where the dependent variable can be predicted based on the values of the independent variables.

Linear Regression

Linear regression: What, Why and When?

- Linear regression is a statistical technique used to model the linear relationship between a **response (dependent) variable** and one or more explanatory (independent) variables. The explanatory variables here are also called **predictors**.
- The response variable as well as the predictors here must all be continuous variables.
- It assumes a linear relationship between the variables, where the dependent variable can be predicted based on the values of the independent variables.
- The goal of linear regression is to find the best-fit line that minimizes the difference between the predicted values and the actual values of the dependent variable. The most well-known method used to estimate the parameters of a regression models is "ordinary least square (OLS)". This is the reason that sometimes the regression models is called the OLS model.

Simple linear regression

When a linear regression model has a single predictor (explanatory variable), it is said to be a *simple linear regression*. More specifically, the relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1, are constants called the regression coefficients or parameters, and ϵ is a random disturbance or error.

Simple linear regression

When a linear regression model has a single predictor (explanatory variable), it is said to be a *simple linear regression*. More specifically, the relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1, are constants called the regression coefficients or parameters, and ϵ is a random disturbance or error.

- The coefficient β_1, called the slope, may be interpreted as the change in Y for a unit change in X.

Simple linear regression

When a linear regression model has a single predictor (explanatory variable), it is said to be a *simple linear regression*. More specifically, the relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1, are constants called the regression coefficients or parameters, and ϵ is a random disturbance or error.

- The coefficient β_1, called the slope, may be interpreted as the change in Y for a unit change in X.
- The coefficient β_0, called the *constant coefficient* or *intercept*, is the predicted value of Y when $X = 0$.

Simple linear regression

When a linear regression model has a single predictor (explanatory variable), it is said to be a *simple linear regression*. More specifically, the relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where β_0 and β_1, are constants called the regression coefficients or parameters, and ϵ is a random disturbance or error.

- The coefficient β_1, called the slope, may be interpreted as the change in Y for a unit change in X.
- The coefficient β_0, called the *constant coefficient* or *intercept*, is the predicted value of Y when $X = 0$.

Based on the available data, we wish to estimate the parameters β_0 and β_1. This is equivalent to finding the straight line that gives the best fit (representation) of the points in the scatter plot of the response versus the predictor variable. There are some methods for estimating them e.g. the MLE and the ordinary least square (OLS). The latter is more practical here, this is the reason that sometimes we say the **OLS model** rather than the linear regression model.

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \ldots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \ldots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

- In the multiple linear regression, we should use a variable selection method to only retain the significant and influential predictors in the model and exclude the remaining ones. The well-known techniques for this purpose are

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \ldots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

- In the multiple linear regression, we should use a variable selection method to only retain the significant and influential predictors in the model and exclude the remaining ones. The well-known techniques for this purpose are
  - Forward algorithm,

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \ldots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

- In the multiple linear regression, we should use a variable selection method to only retain the significant and influential predictors in the model and exclude the remaining ones. The well-known techniques for this purpose are
  - Forward algorithm,
  - Backward algorithm,

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \ldots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \ldots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

- In the multiple linear regression, we should use a variable selection method to only retain the significant and influential predictors in the model and exclude the remaining ones. The well-known techniques for this purpose are
  - Forward algorithm,
  - Backward algorithm,
  - Step-wise algorithm.

Multiple linear regression

When we have more than one predictors in our data set, we may use the multiple linear regression model rather than the simple one. It is obvious that using more predictors, yields more appropriate regression model.
Assume thay t he data consist of n observations on a dependent or response variable Y and d predictors $\mathbf{X} = (X_1, \dots, X_d)$. The relationship between Y and $\mathbf{X}$ is formulated as the following linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \epsilon$$

where $\beta_0, \beta_1, \dots, \beta_d$ are constants referred to as the model regression coefficients. Similar to the simple linear regression, the OLS method is utilized for the estimation of the coefficients.

- In the multiple linear regression, we should use a variable selection method to only retain the significant and influential predictors in the model and exclude the remaining ones. The well-known techniques for this purpose are
  - Forward algorithm,
  - Backward algorithm,
  - Step-wise algorithm.
- In the reality, we use the multiple linear regression (as a generalization of the simple model) since we usually face to more than one predictors.

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

- a) Our explanatory variables, say $X_1, \ldots, X_k$ are all some categorical random variables.

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

- a) Our explanatory variables, say $X_1, \ldots, X_k$ are all some categorical random variables.
- b) Given $Y = y$, the random variables $X_1, \ldots, X_k$ are independent

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

- a) Our explanatory variables, say $X_1, \ldots, X_k$ are all some categorical random variables.
- b) Given $Y = y$, the random variables $X_1, \ldots, X_k$ are independent
- c) Our response variable takes its values among $\{y_1, \ldots, y_c\}$, as a result, it is a categorical random variable too.

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

- a) Our explanatory variables, say $X_1, \ldots, X_k$ are all some categorical random variables.
- b) Given $Y = y$, the random variables $X_1, \ldots, X_k$ are independent
- c) Our response variable takes its values among $\{y_1, \ldots, y_c\}$, as a result, it is a categorical random variable too.

Now, if we are given a new observation $(X_1 = x_1, \ldots, X_k = x_k)$, then, the Naive Bayes prediction of y will be chosen from the set $\{y_1, \ldots, y_c\}$ such that it maximizes the following conditional probability

$$\Pr(Y = y \mid X_1 = x_1, \ldots, X_k = x_k).$$

Naive Bayes model

The Naive Bayes algorithm is a popular machine learning algorithm used for classification tasks. Thus, despite the linear regression model in the previous section, here, the response variable should be a categorical variable.

More precisely, Suppose that we are going to predict the value of a response (class) variable (say Y) given the observations of some explanatory variables. Suppose further that

- a) Our explanatory variables, say $X_1, \ldots, X_k$ are all some categorical random variables.
- b) Given $Y = y$, the random variables $X_1, \ldots, X_k$ are independent
- c) Our response variable takes its values among $\{y_1, \ldots, y_c\}$, as a result, it is a categorical random variable too.

Now, if we are given a new observation $(X_1 = x_1, \ldots, X_k = x_k)$, then, the Naive Bayes prediction of y will be chosen from the set $\{y_1, \ldots, y_c\}$ such that it maximizes the following conditional probability

$$\Pr(Y = y \mid X_1 = x_1, \ldots, X_k = x_k).$$

The above probability is called the ***posterior probability*** *which refers to the probability of a particular class given the observed evidence or features.*

It can not be solved directly, but using the Bayes formula we can calculate it as (for $i = 1, \ldots, c$)

$$\begin{aligned} \Pr(Y = y_i \mid X_1 = x_1, \ldots, X_k = x_k) &= \frac{\Pr(X_1 = x_1, \ldots, X_k = x_k \mid Y = y_i) \Pr(Y = y_i)}{\Pr(X_1 = x_1, \ldots, X_k = x_k)} \\ &= \frac{\Pr(Y = y_i) \prod_{j=1}^{k} \Pr(X_j = x_j \mid Y = y_i)}{\prod_{j=1}^{k} \Pr(X_j = x_j)}. \qquad (1) \end{aligned}$$

The term $\Pr(Y = y_i)$ *given by the numerator is called the* ***prior probability*** *which is the probability of a particular class or category before considering any evidence or features.*
The term $\Pr(X_1 = x_1, \ldots, X_k = x_k \mid Y = y_i)$ *of the numerator is also called the* ***likelihood*** *which is the probability of observing a particular feature given a specific class.*

Since the denominator of the above fraction is the same for all y_i, $i = 1, \ldots, c$, one can only compute the numerator for every y_i.

Section 5

More details on Unsupervised learning

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)
   - Density-based algorithms (pdfCluster, DBSCAN)

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)
   - Density-based algorithms (pdfCluster, DBSCAN)
   - Model-based clustering (Gaussian mixture models)

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
    - Partition-based algorithms (K-**means**, K-mediods)
    - Distance-based algorithms (hierarchical clustering)
    - Density-based algorithms (pdfCluster, DBSCAN)
    - Model-based clustering (Gaussian mixture models)
2. **Latent Dirichlet Allocation (LDA)**: A probabilistic model used for topic modeling. It discovers latent topics in a collection of documents and assigns topic probabilities to each document.

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
    - Partition-based algorithms (K-**means**, K-mediods)
    - Distance-based algorithms (hierarchical clustering)
    - Density-based algorithms (pdfCluster, DBSCAN)
    - Model-based clustering (Gaussian mixture models)
2. **Latent Dirichlet Allocation (LDA)**: A probabilistic model used for topic modeling. It discovers latent topics in a collection of documents and assigns topic probabilities to each document.
3. **Word2Vec**: A popular word embedding technique that represents words as dense vectors in a continuous space. It captures semantic relationships between words and is often used for tasks like word similarity, document classification, and sentiment analysis.

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)
   - Density-based algorithms (pdfCluster, DBSCAN)
   - Model-based clustering (Gaussian mixture models)
2. **Latent Dirichlet Allocation (LDA)**: A probabilistic model used for topic modeling. It discovers latent topics in a collection of documents and assigns topic probabilities to each document.
3. **Word2Vec**: A popular word embedding technique that represents words as dense vectors in a continuous space. It captures semantic relationships between words and is often used for tasks like word similarity, document classification, and sentiment analysis.
4. **Principle component analysis (PCA)**: A dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important information.

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)
   - Density-based algorithms (pdfCluster, DBSCAN)
   - Model-based clustering (Gaussian mixture models)
2. **Latent Dirichlet Allocation (LDA)**: A probabilistic model used for topic modeling. It discovers latent topics in a collection of documents and assigns topic probabilities to each document.
3. **Word2Vec**: A popular word embedding technique that represents words as dense vectors in a continuous space. It captures semantic relationships between words and is often used for tasks like word similarity, document classification, and sentiment analysis.
4. **Principle component analysis (PCA)**: A dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important information.
5. **Association Rule Mining**: An algorithm that discovers interesting relationships or associations between items in a dataset. It is commonly used in market basket analysis.

Some well-known algorithms

1. **Clustering algorithms**: Data points in one cluster are more similar to one another while data points in separate clusters are less similar to one another.
   - Partition-based algorithms (K-**means**, K-mediods)
   - Distance-based algorithms (hierarchical clustering)
   - Density-based algorithms (pdfCluster, DBSCAN)
   - Model-based clustering (Gaussian mixture models)
2. **Latent Dirichlet Allocation (LDA)**: A probabilistic model used for topic modeling. It discovers latent topics in a collection of documents and assigns topic probabilities to each document.
3. **Word2Vec**: A popular word embedding technique that represents words as dense vectors in a continuous space. It captures semantic relationships between words and is often used for tasks like word similarity, document classification, and sentiment analysis.
4. **Principle component analysis (PCA)**: A dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving the most important information.
5. **Association Rule Mining**: An algorithm that discovers interesting relationships or associations between items in a dataset. It is commonly used in market basket analysis.
6. **Anomaly Detection**: Algorithms that identify unusual or anomalous patterns in data. Techniques like One-Class SVM and Isolation Forest are commonly used for anomaly detection.

References

1. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
2. Chatterjee, S. & Hadi, A. (2006). Regression Analysis by Example (Fourth Edition), Wiley.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
4. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
5. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.
6. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
7. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern Classification. Wiley.
8. Ng, A. (2018). Machine Learning Yearning. deeplearning.ai.