Chap1: Machine Learning in Security: An Overview #2

January 17, 2023

Devesh C Jinwala,
Professor, SVNIT and Adjunct Prof., CSE, IIT Jammu
Department of Computer Science and Engineering,
Sardar Vallabhhai National Institute of Technology, SURAT

Chap 1: An Overview of Machine Learning in Security: Topics

- Introduction to the Course Contents, Review of the Basic Machine Learning Concepts. Foundations of Machine Learning for Security: Artificial Intelligence and Machine Learning.
  Review of the ML techniques. Machine Learning problems viz. Classification, Regression, Clustering, Association rule learning, Structured output, Ranking. Linear Regression. Logistics Regression and Bayesian Classification. Support Vector Machines, Decision Tree and Random Forest, Neural Networks, DNNs , Ensemble learning. Principal Components Analysis. Un-supervised learning algorithms: K-means for clustering problems, K-NN (k nearest neighbours). Apriori algorithm for association rule learning problems. Generative vs Discriminative learning. [4 hours]

An Overview of ML tasks

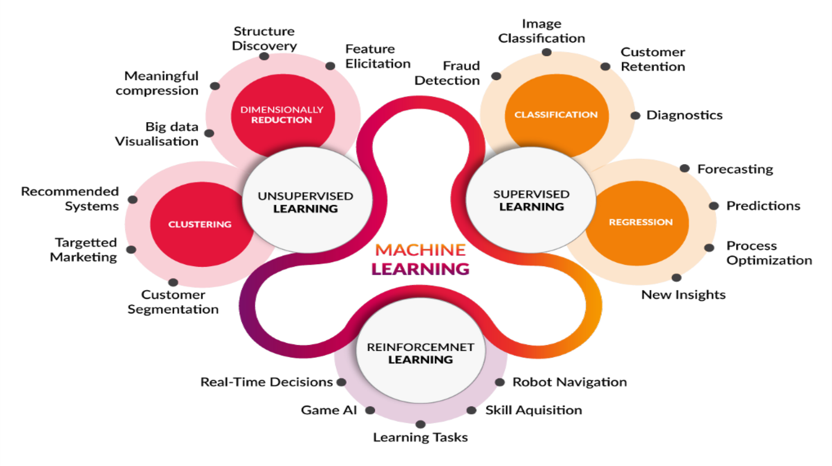
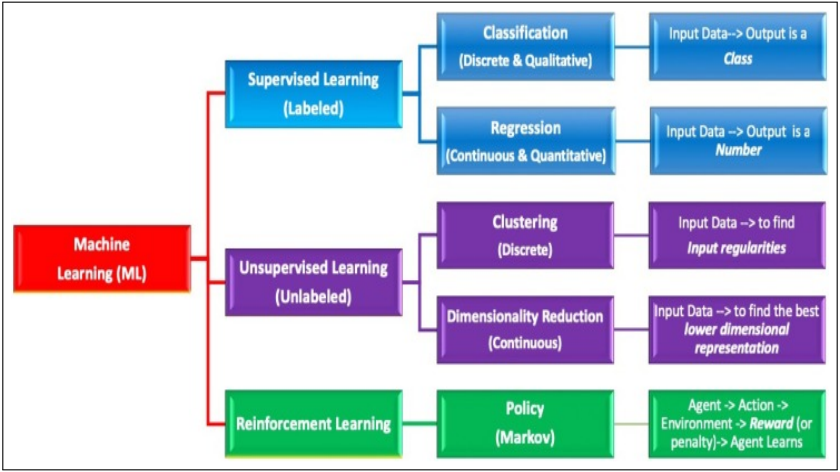


Figure: Machine Learning

Machine Learning...



1 Figure: Machine Learning Techniques w r to input and output

[1] Hooman Rashidi: Academic Pathology, Sept 2019

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.
- Thus, the machine must learn how to differentiate between different classes using these features.

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.
- Thus, the machine must learn how to differentiate between different classes using these features.
- How does the machine classify ?

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.
- Thus, the machine must learn how to differentiate between different classes using these features.
- How does the machine classify ?
  - by pre-labeling data so that the machine knows which class each item belongs to.....thus, must use supervised learning

An Overview of ML tasks: Classification

Classification is simply related to predicting a category of data (discrete variables).

- a type of machine learning task that involves identifying which group or category an item belongs to, based on certain features or characteristics.
- one of the most common types of supervised learning techniques
- used for everything from banking fraud detection to face recognition.
- the process of sorting items into two or more mutually exclusive groups, often called classes.
- used with the goal to correctly assign a given item to the right class based on its features.
- Thus, the machine must learn how to differentiate between different classes using these features.
- How does the machine classify ?
  - by pre-labeling data so that the machine knows which class each item belongs to.....thus, must use supervised learning
- Thus, a classification task results in the model which, given a new individual, determines which class that individual belongs to.

In classification......

- a closely related task is scoring or class probability estimation.

...continued

An Overview of ML tasks: Classification ...

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.

...*continued*

An Overview of ML tasks: Classification ...

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.
- What could be probable use cases/applications of this scenario?

...*continued*

An Overview of ML tasks: Classification ...

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.
- What could be probable use cases/applications of this scenario?
  - in predicting whether or not an email if spam or ham.

...continued

An Overview of ML tasks: Classification ...

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.
- What could be probable use cases/applications of this scenario?
  - in predicting whether or not an email if spam or ham.
  - in finance, determining whether a transaction is a fraud or not.

...continued

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.
- What could be probable use cases/applications of this scenario?
  - in predicting whether or not an email if spam or ham.
  - in finance, determining whether a transaction is a fraud or not.
  - in healthcare, predicting whether a person is suffering from a particular disease or not.

...continued

In classification......

- a closely related task is scoring or class probability estimation.
- a scoring model applied to an individual produces, a score representing the probability (or some other quantification of likelihood) that, that individual belongs to each class.
- What could be probable use cases/applications of this scenario?
  - in predicting whether or not an email if spam or ham.
  - in finance, determining whether a transaction is a fraud or not.
  - in healthcare, predicting whether a person is suffering from a particular disease or not.
- What are the ML algorithms/methods applied to solve classification tasks?

...continued

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees
- Boosted trees

An Overview of ML tasks: Classification ...

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees
- Boosted trees
- Logistic regression

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees
- Boosted trees
- Logistic regression
- Naive Bayes

...continued

What are the ML methods applied to solve classification tasks?

- Kernel discriminant analysis (Higher accuracy)
- K-Nearest Neighbors (Higher accuracy)
- Artificial neural networks (ANN) (Higher accuracy)
- Support vector machine (SVM) (Higher accuracy)
- Random forests (Higher accuracy)
- Decision trees
- Boosted trees
- Logistic regression
- Naive Bayes
- Deep learning

An Overview of ML tasks: Classification ...

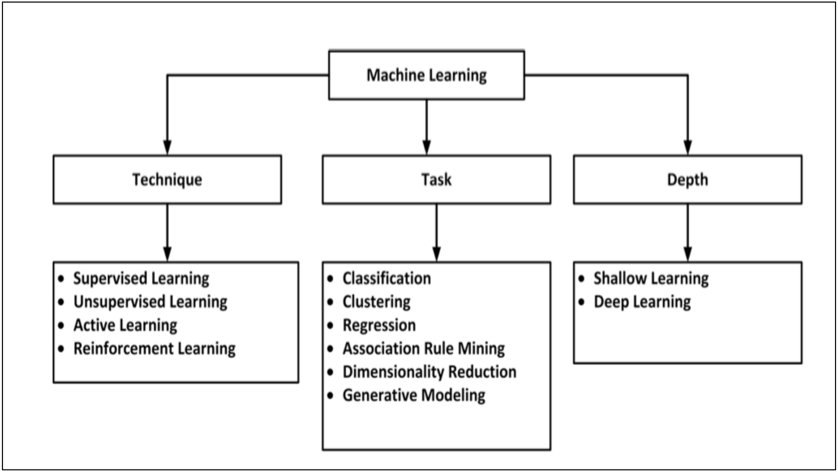


Figure: Machine Learning - Tasks, Techniques & Depth

1

[1] Olakunle Ibitoye et al

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from
  - customer segmentation,

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from
  - customer segmentation,
  - market segmentation,

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from
  - customer segmentation,
  - market segmentation,
  - image segmentation,

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from
  - customer segmentation,
  - market segmentation,
  - image segmentation,
  - document classification, and more.

An Overview of ML tasks: Clustering

Clustering is a commonly used ML task in which data points are grouped into clusters, i.e. groups of closely related data points. It

- i.e. partitioning of a set of objects into distinct groups such that
  - the elements in each group are similar to each other whereas
  - those belonging to different groups are very dissimilar.
- is an unsupervised approach that does not require labeled data
- can be used to identify patterns or similarities within a dataset.
- has many applications ranging from
  - customer segmentation,
  - market segmentation,
  - image segmentation,
  - document classification, and more.
- What are the ML methods applied to solve classification tasks?

An Overview of ML tasks: Clustering

The following are four different type of clustering algorithms:

- Prototype based clustering (K-means)

An Overview of ML tasks: Clustering

The following are four different type of clustering algorithms:

- Prototype based clustering (K-means)
- Hierarchical clustering

An Overview of ML tasks: Clustering

The following are four different type of clustering algorithms:

- Prototype based clustering (K-means)
- Hierarchical clustering
- DBSCAN (Density based spatial clustering of applications with noise)

An Overview of ML tasks: Clustering

The following are four different type of clustering algorithms:

- Prototype based clustering (K-means)
- Hierarchical clustering
- DBSCAN (Density based spatial clustering of applications with noise)
- Distribution based clustering

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.
- more specifically, using two variables viz. an independent variable, and a dependent variable.

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.
- more specifically, using two variables viz. an independent variable, and a dependent variable.
- is commonly used for predictive analysis that is to examine two things.

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.
- more specifically, using two variables viz. an independent variable, and a dependent variable.
- is commonly used for predictive analysis that is to examine two things.
    - first, does a set of predictor variables do a good job in predicting an outcome (dependent) variable?

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.
- more specifically, using two variables viz. an independent variable, and a dependent variable.
- is commonly used for predictive analysis that is to examine two things.
  - first, does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - second, which variables are significant predictors of the outcome variable?

An Overview of ML tasks: Regression

Regression tasks mainly deal with the estimation of numerical values i.e. that of continuous variables. Regression task in ML

- is a supervised ML task used to predict the values of a given target variable using the results that we already know
- thus is used to predict the relationship between two variables by applying a linear equation to observed data.
- more specifically, using two variables viz. an independent variable, and a dependent variable.
- is commonly used for predictive analysis that is to examine two things.
  - first, does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - second, which variables are significant predictors of the outcome variable?
- involves the task of fitting a mathematical model to observed data points, with the objective to minimize the sum of squared errors between the observed data and the predicted values.

Machine Learning Tasks Dimensions... reviewed again

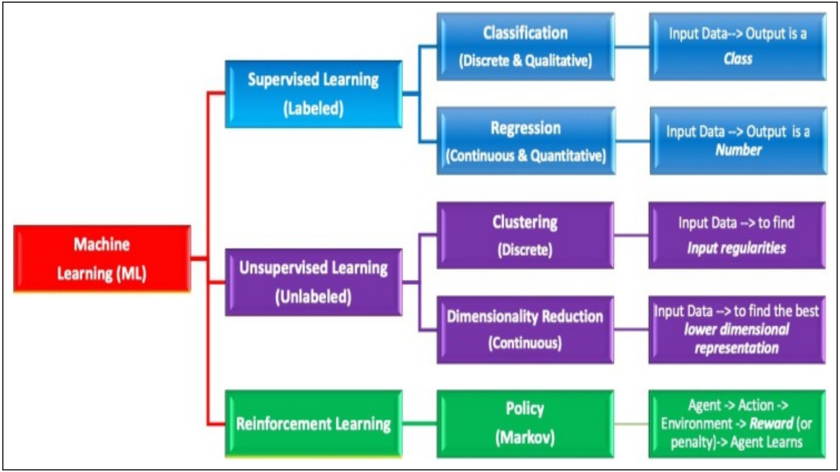


Figure: Machine Learning : Tasks, Techniques & Depth

1

[1] Hooman Rashidi: Academic Pathology, Sept 2019

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
    - have a basic assumption that there exists a linear relationship between the input and output variables

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.
    - similarly, predicting anemia based on the pale color of face, tiredness and underweightedness of a person.

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.
    - similarly, predicting anemia based on the pale color of face, tiredness and underweightedness of a person.
- Non-linear models

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.
    - similarly, predicting anemia based on the pale color of face, tiredness and underweightedness of a person.
- Non-linear models
  - do not rely on any assumptions of a linear relationship between the input and output variables.

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.
    - similarly, predicting anemia based on the pale color of face, tiredness and underweightedness of a person.
- Non-linear models
  - do not rely on any assumptions of a linear relationship between the input and output variables.
  - the goal is to try to identify complex relationships within the dataset.

An Overview of ML tasks: Regression...

Linear and Non-linear Regression tasks in ML

- could use either linear OR non-linear models to build predictive models.
- Linear models
  - have a basic assumption that there exists a linear relationship between the input and output variables
  - the goal here is to find the best fit line for data.
    - e.g. find the relation between the weight of the person and his/her height.
    - similarly, predicting anemia based on the pale color of face, tiredness and underweightedness of a person.
- Non-linear models
  - do not rely on any assumptions of a linear relationship between the input and output variables.
  - the goal is to try to identify complex relationships within the dataset.
    - e.g. determine the impact of gold prices, prices of crude oil etc on the inflation. Similarly, the analysis in sectors like insurance, agriculture, finance, investing.

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees
- Linear regression

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees
- Linear regression
- Support vector regression

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees
- Linear regression
- Support vector regression
- LASSO / Ridge

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees
- Linear regression
- Support vector regression
- LASSO / Ridge
- Deep learning

An Overview of ML tasks: Regression

Some of the examples include estimation of housing price, product price, stock price etc. Some of the following ML methods could be used for solving regressions problems:

- Kernel regression (Higher accuracy)
- Gaussian process regression (Higher accuracy)
- Regression trees
- Linear regression
- Support vector regression
- LASSO / Ridge
- Deep learning
- Random forests

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)
- this relationship shall then be used to make predictions.

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)
- this relationship shall then be used to make predictions.
- this can be represented as a linear equation of the form $y = a + bx$.

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)
- this relationship shall then be used to make predictions.
- this can be represented as a linear equation of the form $y = a + bx$.
- the ML linear regression algorithm aims to find the best possible values for the coefficients a and b by performing calculations on the data provided.

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)
- this relationship shall then be used to make predictions.
- this can be represented as a linear equation of the form $y = a + bx$.
- the ML linear regression algorithm aims to find the best possible values for the coefficients a and b by performing calculations on the data provided.
- once the calculations are done, the linear regression algorithm returns a model, i.e. the values of a and b.

Linear Regression as an ML task: Another example

- Assume that task is to establish a linear relationship between an independent variable (x) and a dependent variable (y)
- this relationship shall then be used to make predictions.
- this can be represented as a linear equation of the form $y = a + bx$.
- the ML linear regression algorithm aims to find the best possible values for the coefficients a and b by performing calculations on the data provided.
- once the calculations are done, the linear regression algorithm returns a model, i.e. the values of a and b.
- Then, the ML can use the equation $y = a + bx$. (with known values of a and b) to make predictions.

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,
- then using that information as training data so that the machine can make accurate predictions when presented with new data points.

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,
- then using that information as training data so that the machine can make accurate predictions when presented with new data points.
- e.g. providing recommendations or helping with search engine optimization e.g.

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,
- then using that information as training data so that the machine can make accurate predictions when presented with new data points.
- e.g. providing recommendations or helping with search engine optimization e.g.
  - if one is looking for a particular product online but couldn't find it through traditional search methods

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,
- then using that information as training data so that the machine can make accurate predictions when presented with new data points.
- e.g. providing recommendations or helping with search engine optimization e.g.
  - if one is looking for a particular product online but couldn't find it through traditional search methods
  - similarity matching could present other products

An Overview of ML tasks: Similarity matching

Similarity matching task in ML is a task in which machines are trained to match items based on their similarity. Similarity matching

- can be used for a wide range of applications, such as natural language processing, image recognition and recommendation systems.
- requires that the system needs to learn how to distinguish between similar and dissimilar items.
- can be done by creating feature vectors from examples of known data points,
- then using that information as training data so that the machine can make accurate predictions when presented with new data points.
- e.g. providing recommendations or helping with search engine optimization e.g.
  - if one is looking for a particular product online but couldn't find it through traditional search methods
  - similarity matching could present other products
    - that would be such as those that closely match the desired item based on their features and characteristics.

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Co-occurrence grouping

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.

1

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Co-occurrence grouping

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
  - e.g. what items are commonly purchased together?

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
    - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
    - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?
- the difference is based on the way the similarity of objects is found...

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
  - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?
- the difference is based on the way the similarity of objects is found...
  - in clustering, it is found based on the objects' attributes whereas

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Co-occurrence grouping

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
    - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?
- the difference is based on the way the similarity of objects is found...
    - in clustering, it is found based on the objects' attributes whereas
    - in co-occurrence grouping, it is found based on them appearing together in transactions.

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Co-occurrence grouping

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
    - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?
- the difference is based on the way the similarity of objects is found...
    - in clustering, it is found based on the objects' attributes whereas
    - in co-occurrence grouping, it is found based on them appearing together in transactions.
        - e.g. the purchase records from a supermarket may uncover the association that bread is purchased together with eggs much more frequently than expected.

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Co-occurrence grouping

Co-occurrence grouping tasks

- are also called frequent itemset mining, association rule discovery, and market-basket analysis tasks
- association between entities are found based on transactions involving them.
  - e.g. what items are commonly purchased together?
- What is the difference between clustering and co-occurrence grouping— ?
- the difference is based on the way the similarity of objects is found...
  - in clustering, it is found based on the objects' attributes whereas
  - in co-occurrence grouping, it is found based on them appearing together in transactions.
    - e.g. the purchase records from a supermarket may uncover the association that bread is purchased together with eggs much more frequently than expected.
- Assignment: With the help of an example, explain the differences between similarity matching, classification, co-occurrence grouping and multi variate querying

[1]

[1] https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

An Overview of ML tasks: Multivariate querying

- Multivariate querying is about querying or finding similar objects.

An Overview of ML tasks: Multivariate querying

- Multivariate querying is about querying or finding similar objects.
- Some of the ML methods used for such problems are as follows

An Overview of ML tasks: Multivariate querying

- Multivariate querying is about querying or finding similar objects.
- Some of the ML methods used for such problems are as follows
    - Nearest neighbors

An Overview of ML tasks: Multivariate querying

- Multivariate querying is about querying or finding similar objects.
- Some of the ML methods used for such problems are as follows
  - Nearest neighbors
  - Range search

An Overview of ML tasks: Multivariate querying

- Multivariate querying is about querying or finding similar objects.
- Some of the ML methods used for such problems are as follows
  - Nearest neighbors
  - Range search
  - Farthest neighbors

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.
- density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function.

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.
- density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function.
- Some of the ML methods used for solving density estimation tasks are as follows

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.
- density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function.
- Some of the ML methods used for solving density estimation tasks are as follows
  - Kernel density estimation (Higher accuracy)

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.
- density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function.
- Some of the ML methods used for solving density estimation tasks are as follows
  - Kernel density estimation (Higher accuracy)
  - Mixture of Gaussians

Probability density and mass function estimation problems

- are related to finding the likelihood or frequency of objects.
- density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function.
- Some of the ML methods used for solving density estimation tasks are as follows
    - Kernel density estimation (Higher accuracy)
    - Mixture of Gaussians
    - Density estimation tree

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and
  - machine translation of web pages.

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and
  - machine translation of web pages.
- Deep learning models have achieved state-of-the-art results on many machine translation tasks e.g. used to machine translate with close to human-level accuracy

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and
  - machine translation of web pages.
- Deep learning models have achieved state-of-the-art results on many machine translation tasks e.g. used to machine translate with close to human-level accuracy
  - Web pages from English to Chinese

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and
  - machine translation of web pages.
- Deep learning models have achieved state-of-the-art results on many machine translation tasks e.g. used to machine translate with close to human-level accuracy
  - Web pages from English to Chinese
  - speech from English to French

An Overview of ML tasks: Machine translation

Machine translation

- is the process of translating text from one language to another using ML algorithms.
- There are many different machine translation tasks
  - machine translation of documents,
  - machine translation of the speech, and
  - machine translation of web pages.
- Deep learning models have achieved state-of-the-art results on many machine translation tasks e.g. used to machine translate with close to human-level accuracy
  - Web pages from English to Chinese
  - speech from English to French
  - and so on....

An Overview of ML tasks: Transcription

Transcription tasks are

- those that involve converting audio or video recordings or images having text into written text.

An Overview of ML tasks: Transcription

Transcription tasks are

- those that involve converting audio or video recordings or images having text into written text.
- commonly used in fields such as journalism, academia, and medicine

An Overview of ML tasks: Transcription

Transcription tasks are

- those that involve converting audio or video recordings or images having text into written text.
- commonly used in fields such as journalism, academia, and medicine
- have been automated to some degree using ML algorithms, recently.

An Overview of ML tasks: Transcription

Transcription tasks are

- those that involve converting audio or video recordings or images having text into written text.
- commonly used in fields such as journalism, academia, and medicine
- have been automated to some degree using ML algorithms, recently.
- however, the DL models used require a large amount of data to train on, and they often struggle with background noise and accents....

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.
- is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science.

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.
- is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science.
- here, data is used to make inferences about the relationships between variables.

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.
- is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science.
- here, data is used to make inferences about the relationships between variables.
- the goal is to identify which variables are causing certain outcomes and how they are related.

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.
- is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science.
- here, data is used to make inferences about the relationships between variables.
- the goal is to identify which variables are causing certain outcomes and how they are related.
- For example, the question "Is smoking cigarettes related to lung cancer?".....may necessitate

An Overview of ML tasks: Causal modeling

Causal modeling

- is a type of ML task that aims to infer the causes and effects of certain conditions or variables.
- is an important tool for researchers in fields such as epidemiology, economics, psychology, marketing, and political science.
- here, data is used to make inferences about the relationships between variables.
- the goal is to identify which variables are causing certain outcomes and how they are related.
- For example, the question "Is smoking cigarettes related to lung cancer?".....may necessitate
  - a researcher to determine the causal relationship between smoking cigarettes and lung cancer.

Causal modeling

- there are two aspects of Causal Modeling:

An Overview of ML tasks: Causal modeling...

Causal modeling

- there are two aspects of Causal Modeling:
    - Causal Discovery: these algorithms try to derive causal relations from observational data. Given a set of data, a causal discovery algorithm returns a set of statements regarding the causal interactions between the measured variables.

An Overview of ML tasks: Causal modeling...

Causal modeling

- there are two aspects of Causal Modeling:
  - Causal Discovery: these algorithms try to derive causal relations from observational data. Given a set of data, a causal discovery algorithm returns a set of statements regarding the causal interactions between the measured variables.
  - Causal Inference. is the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect

Causal modeling in Cyber Security

- in this research attempt[1], the characteristics of the VERIS Community Database (VCDB) were studied

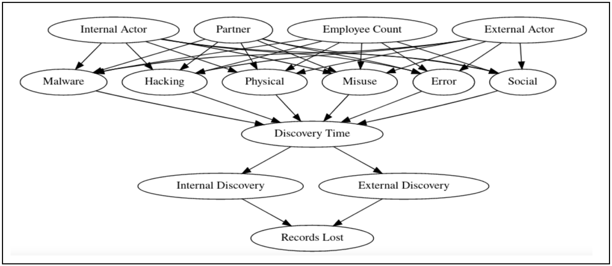


Figure: A Causal Model for Cybersecurity (on VCDB)

[1] https://dx.doi.org/10.25046/aj050349

Causal modeling in Cyber Security

- in this research attempt[1], the characteristics of the VERIS Community Database (VCDB) were studied
- this was done to to evaluate the risks of data breach of cyber-security incidents

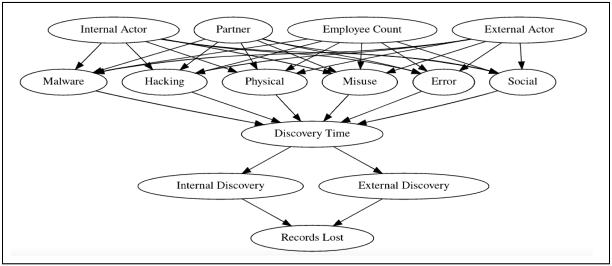


Figure: A Causal Model for Cybersecurity (on VCDB)

1 https://dx.doi.org/10.25046/aj050349

Causal modeling in Cyber Security

- in this research attempt[1], the characteristics of the VERIS Community Database (VCDB) were studied
- this was done to to evaluate the risks of data breach of cyber-security incidents
  - VCDB is an open-source dataset dataset of cyber-security incidents - containing a breadth of information regarding data breaches.

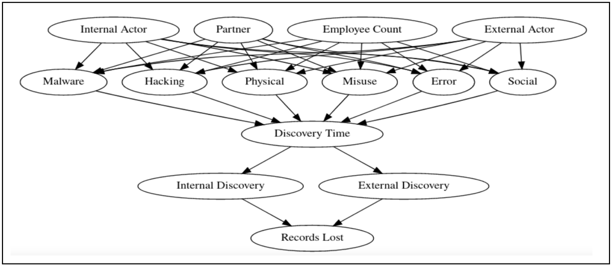


Figure: A Causal Model for Cybersecurity (on VCDB)

[1] https://dx.doi.org/10.25046/aj050349

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
    - Manifold learning/KPCA (Higher accuracy)

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis
  - Independent component analysis

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis
  - Independent component analysis
  - Gaussian graphical models

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis
  - Independent component analysis
  - Gaussian graphical models
  - Non-negative matrix factorization

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis
  - Independent component analysis
  - Gaussian graphical models
  - Non-negative matrix factorization
  - Compressed sensing

An Overview of ML tasks: Dimensionality reduction (feature extraction)

Dimensionality reduction or Feature extraction

- is the process of reducing the number of random variables under consideration
- can be divided into feature selection and feature extraction.
- Following are some ML methods that could be used for dimension reduction:
  - Manifold learning/KPCA (Higher accuracy)
  - Principal component analysis
  - Independent component analysis
  - Gaussian graphical models
  - Non-negative matrix factorization
  - Compressed sensing
- research is done with dimensionality reduction analysis (DRA) to cyber security where relevant features for threat detection are ranked and identified - reduced to improve the response time.

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.

An Overview of ML tasks: Link prediction

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.
  - but it can also be applied to other types of data such as customer transactions data or scientific research papers.

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.
  - but it can also be applied to other types of data such as customer transactions data or scientific research papers.
- LP models are also often used in recommendation systems to recommend items to customers based on their past behavior or preferences.

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.
  - but it can also be applied to other types of data such as customer transactions data or scientific research papers.
- LP models are also often used in recommendation systems to recommend items to customers based on their past behavior or preferences.
- can also estimate the strength of a link. That is,

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.
  - but it can also be applied to other types of data such as customer transactions data or scientific research papers.
- LP models are also often used in recommendation systems to recommend items to customers based on their past behavior or preferences.
- can also estimate the strength of a link. That is,
  - for recommending movies to customers, can be used to create a graph between customers and the movies they've watched or rated.

Link prediction

- is a task that focuses on identifying potential connections between entities that are not yet connected.
- is used to predict relationships between entities, such as customers, products, authors, and more.
- has a goal to build a model that can accurately identify connections between entities in a dataset.
- in the past, has primarily been used for social network analysis
  - to suggest friends or followers for users of a social network platform.
  - but it can also be applied to other types of data such as customer transactions data or scientific research papers.
- LP models are also often used in recommendation systems to recommend items to customers based on their past behavior or preferences.
- can also estimate the strength of a link. That is,
  - for recommending movies to customers, can be used to create a graph between customers and the movies they've watched or rated.
  - within the graph, those potential links (strong link) are searched that should exist between customers and movies.

An Overview of ML tasks: Link prediction in security

- Graph link prediction is an important task in cyber-security

An Overview of ML tasks: Link prediction in security

- Graph link prediction is an important task in cyber-security
  - relationships between entities within a computer network, such as users interacting with computers, or system libraries and the corresponding processes that use them,

An Overview of ML tasks: Link prediction in security

- Graph link prediction is an important task in cyber-security
  - relationships between entities within a computer network, such as users interacting with computers, or system libraries and the corresponding processes that use them,
  - this insight can provide key insights into adversary behavior

An Overview of ML tasks: Link prediction in security

- Graph link prediction is an important task in cyber-security
  - relationships between entities within a computer network, such as users interacting with computers, or system libraries and the corresponding processes that use them,
  - this insight can provide key insights into adversary behavior
  - Poisson matrix factorization (PMF) is a popular model for link prediction in large networks,

An Overview of ML tasks: Link prediction in security

- Graph link prediction is an important task in cyber-security
  - relationships between entities within a computer network, such as users interacting with computers, or system libraries and the corresponding processes that use them,
  - this insight can provide key insights into adversary behavior
  - Poisson matrix factorization (PMF) is a popular model for link prediction in large networks,
  - PMF is extended to include scenarios that are commonly encountered in cyber-security applications.

An Overview of ML tasks: Synthesis & sampling

Synthesis and sampling

- are used to generate new data from existing data or to select a representative subset of data for further analysis

An Overview of ML tasks: Synthesis & sampling

Synthesis and sampling

- are used to generate new data from existing data or to select a representative subset of data for further analysis
- are often used together, in order to create a more diverse and representative dataset

An Overview of ML tasks: Synthesis & sampling

Synthesis and sampling

- are used to generate new data from existing data or to select a representative subset of data for further analysis
- are often used together, in order to create a more diverse and representative dataset

An Overview of ML tasks: Synthesis & sampling

Synthesis and sampling

- are used to generate new data from existing data or to select a representative subset of data for further analysis
- are often used together, in order to create a more diverse and representative dataset

Synthesis can be used

- to generate new data points, by extrapolating from existing data points.
- For example, if we have a dataset of images of animals, we can use synthesis to generate new images of animals that are similar to the ones in the dataset.

An Overview of ML tasks: Synthesis & sampling

Synthesis and sampling

- are used to generate new data from existing data or to select a representative subset of data for further analysis
- are often used together, in order to create a more diverse and representative dataset

Synthesis can be used

- to generate new data points, by extrapolating from existing data points.
- For example, if we have a dataset of images of animals, we can use synthesis to generate new images of animals that are similar to the ones in the dataset.

Sampling can be used

- to select a subset of data that is representative of the entire dataset.
- For example, if we have a dataset of images of animals, we can use sampling to select a subset of images that represents all the different animal types in the dataset.

An Overview of ML tasks: Anomaly detection

Anomaly detection

- is the process of identifying unusual patterns in data that do not conform to expected behavior.

An Overview of ML tasks: Anomaly detection

Anomaly detection

- is the process of identifying unusual patterns in data that do not conform to expected behavior.
- is often used in a wide range of applications, such as detecting fraudulent activity in financial data, detecting malicious behavior in network traffic data, and identifying equipment malfunctions in sensor data.

Anomaly detection

- is the process of identifying unusual patterns in data that do not conform to expected behavior.
- is often used in a wide range of applications, such as detecting fraudulent activity in financial data, detecting malicious behavior in network traffic data, and identifying equipment malfunctions in sensor data.
- can be performed using a variety of ML models, such as density-based methods, cluster-based methods, and rule-based methods. Therefore, it is is important to select the right model for the particular application

An Overview of development phases in ML models

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.
2. Examine the data: examination of data involves examination to get a general feel for the dataset. For example, for a simple CSV file-based dataset

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.
2. Examine the data: examination of data involves examination to get a general feel for the dataset. For example, for a simple CSV file-based dataset
   - one can look at the dataset's shape (i.e., the number of rows and columns in the dataset).

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.
2. Examine the data: examination of data involves examination to get a general feel for the dataset. For example, for a simple CSV file-based dataset
   - one can look at the dataset's shape (i.e., the number of rows and columns in the dataset).
   - one can also peek inside the dataset by looking at its first 10 or 20 rows.

....continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.
2. Examine the data: examination of data involves examination to get a general feel for the dataset. For example, for a simple CSV file-based dataset
   - one can look at the dataset's shape (i.e., the number of rows and columns in the dataset).
   - one can also peek inside the dataset by looking at its first 10 or 20 rows.
   - one can perform fundamental analysis on the data to generate some descriptive statistical measures (such as the mean, standard deviation, minimum and maximum values).

...continued

Life Cycle of a Machine Learning Project

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

1. Loading the data: may need different libraries to load the respective data. For example,
   - for loading CSV files, one needs the pandas library.
   - for loading 2D images, one can use the Pillow or OpenCV library.
   - and so on.
2. Examine the data: examination of data involves examination to get a general feel for the dataset. For example, for a simple CSV file-based dataset
   - one can look at the dataset's shape (i.e., the number of rows and columns in the dataset).
   - one can also peek inside the dataset by looking at its first 10 or 20 rows.
   - one can perform fundamental analysis on the data to generate some descriptive statistical measures (such as the mean, standard deviation, minimum and maximum values).
   - one can check if the dataset contains missing data and the ways to handle those.

...continued

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
    - this is done before we handle missing values or do any form of computation on our dataset

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
    - this is done before we handle missing values or do any form of computation on our dataset
    - a common practice is to use 80% of the dataset for training and 20% for testing.

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
    - this is done before we handle missing values or do any form of computation on our dataset
    - a common practice is to use 80% of the dataset for training and 20% for testing.
    - the training subset is the actual dataset used for training the model.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
    - this is done before we handle missing values or do any form of computation on our dataset
    - a common practice is to use 80% of the dataset for training and 20% for testing.
    - the training subset is the actual dataset used for training the model.
    - after the training process is complete, we can use the test subset to evaluate how well the model generalizes to unseen data (i.e., data not used to train the model).

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
    - this is done before we handle missing values or do any form of computation on our dataset
    - a common practice is to use 80% of the dataset for training and 20% for testing.
    - the training subset is the actual dataset used for training the model.
    - after the training process is complete, we can use the test subset to evaluate how well the model generalizes to unseen data (i.e., data not used to train the model).
2. Visualizing the data: after splitting the dataset, for further investigation.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
   - this is done before we handle missing values or do any form of computation on our dataset
   - a common practice is to use 80% of the dataset for training and 20% for testing.
   - the training subset is the actual dataset used for training the model.
   - after the training process is complete, we can use the test subset to evaluate how well the model generalizes to unseen data (i.e., data not used to train the model).
2. Visualizing the data: after splitting the dataset, for further investigation.
   - we can plot some graphs to better understand the data we are investigating.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Split the Dataset: ....into training and test subsets
   - this is done before we handle missing values or do any form of computation on our dataset
   - a common practice is to use 80% of the dataset for training and 20% for testing.
   - the training subset is the actual dataset used for training the model.
   - after the training process is complete, we can use the test subset to evaluate how well the model generalizes to unseen data (i.e., data not used to train the model).
2. Visualizing the data: after splitting the dataset, for further investigation.
   - we can plot some graphs to better understand the data we are investigating.
   - For instance, we can plot scatter plots to investigate the relationships between the features and the target variable.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data
   - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use)

....continued

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
    - some problems with the dataset include missing values, textual and categorical data
    - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
    - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data
   - e.g., “Red”, “Green”, and “Blue” for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
   - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.
2. Lastly, Train and Evaluate the model:

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
    - some problems with the dataset include missing values, textual and categorical data
    - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
    - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.
2. Lastly, Train and Evaluate the model:
    - Based on the previous steps of analyzing the dataset, one can narrow down appropriate ML algorithms

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data
   - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
   - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.
2. Lastly, Train and Evaluate the model:
   - Based on the previous steps of analyzing the dataset, one can narrow down appropriate ML algorithms
   - then, build models using those algorithms.

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data
   - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
   - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.
2. Lastly, Train and Evaluate the model:
   - Based on the previous steps of analyzing the dataset, one can narrow down appropriate ML algorithms
   - then, build models using those algorithms.
3. after building the models, one needs to evaluate models using different metrics

Life Cycle of a Machine Learning Project...

Most ML projects follow a typical life cycle that includes some (or all) of the steps, as follows: (we assume that that the data is already collected and available for use) *....continued*

1. Data Preprocessing ...because the data that we receive is not ready to be used immediately.
   - some problems with the dataset include missing values, textual and categorical data
   - e.g., "Red", "Green", and "Blue" for color, or range of features that differ too much (such as a feature with a range of 0 to 10,000 and another with a range of 0 to 5).
   - most machine learning algorithms do not perform well when any of the issues above exist in the dataset.
2. Lastly, Train and Evaluate the model:
   - Based on the previous steps of analyzing the dataset, one can narrow down appropriate ML algorithms
   - then, build models using those algorithms.
3. after building the models, one needs to evaluate models using different metrics
4. select the best-performing model for deployment.

An Overview ML models development phases: Data Gathering

- Any ML problem requires a lot of data for training/testing purposes.
- Identifying the right data sources and gathering data from these data sources is the first step in ML development cycle
- Data could be found from databases, external agencies, the internet, etc.

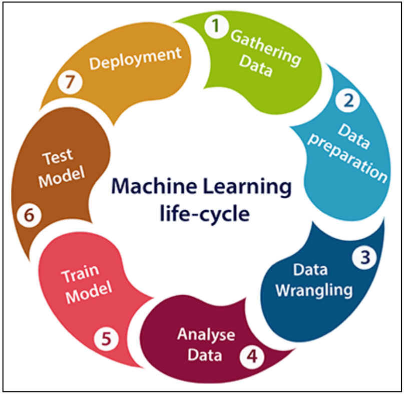


Figure: ML Development Life Cycle

- Before starting training the models, it is of utmost importance to prepare data appropriately.
- As part of data preprocessing, some of the following tasks may be required
  - Data cleaning requires one to identify attributes having not enough data or attributes which are not have variance. These data (rows and columns) need to be removed from the training data set.
  - Missing data imputation using data imputation techniques such as replacing missing data with mean, median, or mode.

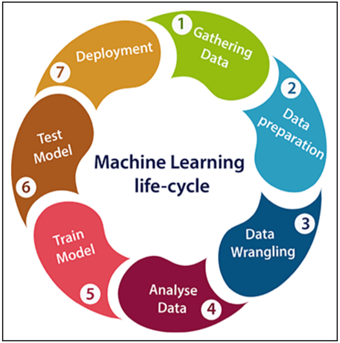


Figure: ML Development Life Cycle

An Overview ML models development phases: Exploratory Data Analysis (EDA)

- Once data is preprocessed, the next step is to perform exploratory data analysis.
- this is done to understand data distribution and relationships between/within the data.
- Some of the following are performed as part of EDA:
  - Correlation analysis
  - Multicollinearity analysis
  - Data distribution analysis

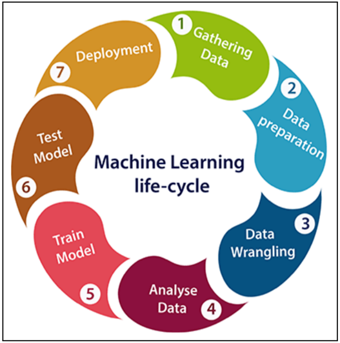


Figure: ML Development Life Cycle

- is one of the most critical tasks when building machine learning models
- this is so because not only would it help build models of higher accuracy but also help achieve objectives related to building simpler models, reducing overfitting etc.
- includes tasks such as
  - deriving features from raw features,
  - identifying important features,
  - feature extraction and feature selection.
- some of the techniques used for feature selection are enlisted on the next slide.

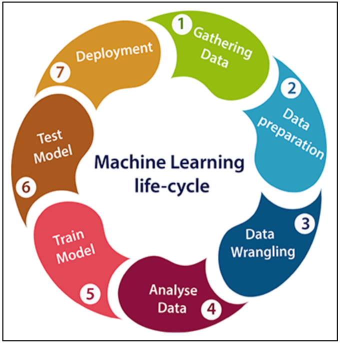


Figure: ML Development Life Cycle

An Overview ML models development phases: Feature Engineering...

The following are some of the statistical tests used in feature engineering:

- Pearson's correlation
- Linear discriminant analysis (LDA)
- Analysis of Variance (ANOVA)
- Chi-square tests
- Wrapper methods that use a subset of features.

The following are some of the algorithms used for Wrapper methods that help in feature selection by using a subset of features

- Forward selection
- Backward elimination
- Recursive feature elimination

The following are some of the algorithms used for Regularization techniques that penalize one or more features appropriately to come up with most important features.

- Regularization with classification algorithms such as Logistic regression, SVM, etc.
- Elastic net regularization
- LASSO (L1) regularization
- Ridge (L2) regularization

An Overview ML models development phases: Training Models

- is the step to be followed once the features are determined.
- one of the methods followed is as follows: start with random initial parameter values and then gradually update them by taking small steps until we reach an optimal solution.
- the iterative process helps us reduce error rates over time and ultimately provide better predictions for our target variable.

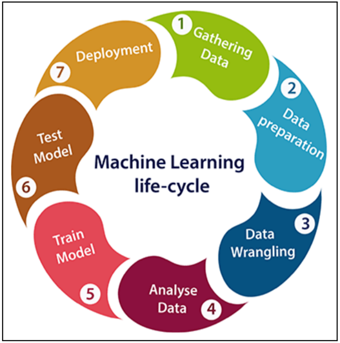


Figure: ML Development Life Cycle

- often, there are multiple models trained using different algorithms.
- hence, it is an important task is to select the most optimal models for deploying them in production.
- Hyperparameter tuning is the most common task performed as part of model selection.
- Also, if there are two models trained using different algorithms which have similar performance, then one also needs to perform algorithm selection.

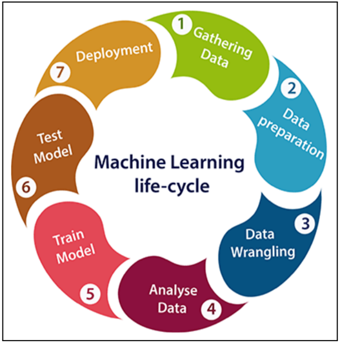


Figure: ML Development Life Cycle

An Overview ML models development phases: Testing and matching

- Testing and matching tasks relate to comparing data sets.
- Following are some of the methods that could be used for such kinds of problems:
    - Minimum spanning tree
    - Bipartite cross-matching
    - N-point correlation

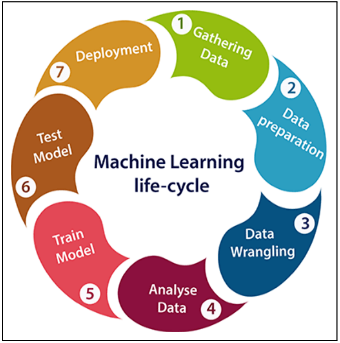


Figure: ML Development Life Cycle

An Overview ML models development phases: Model monitoring

- Once the models are trained and deployed, they require to be monitored at regular intervals.
- Monitoring models require the processing actual values and predicted values and measuring the model performance based on appropriate metrics.

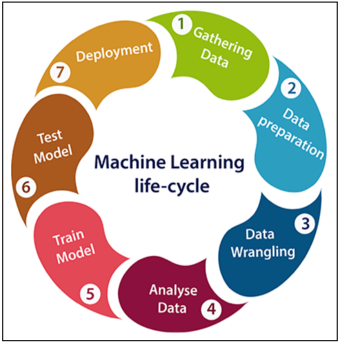


Figure: ML Development Life Cycle

- This is to be applied in case, the model performance degrades,
- the models subsequently required to be retrained.
- The following gets done as part of model retraining:
  - New features get determined
  - New algorithms can be used
  - Hyperparameters can get tuned
  - Model ensembles may get deployed

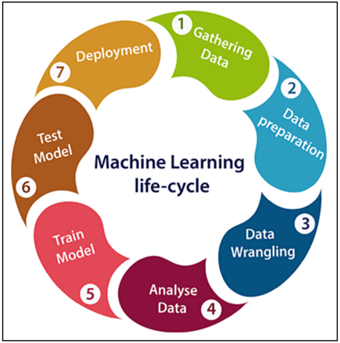


Figure: ML Development Life Cycle

- is another popular term that is commonly conflated with machine learning.

Deep Learning, Neural Networks

- is another popular term that is commonly conflated with machine learning.
- is a strict subset of machine learning referring to a specific class of multilayered models that use layers of simpler statistical components to learn representations of data.

Deep Learning, Neural Networks

- is another popular term that is commonly conflated with machine learning.
- is a strict subset of machine learning referring to a specific class of multilayered models that use layers of simpler statistical components to learn representations of data.

Deep Learning, Neural Networks

- is another popular term that is commonly conflated with machine learning.
- is a strict subset of machine learning referring to a specific class of multilayered models that use layers of simpler statistical components to learn representations of data.

Neural network

- is a more general term for this type of layered statistical learning architecture that might or might not be *deep* (i.e., have many layers).

B l a n k

B l a n k

B l a n k

Blank

B l a n k

B l a n k

B l a n k

B l a n k