Module-04, Python for Machine Learning

Supervised Machine Learning Model

Dostdar Ali
Instructor

Data science and Artificial Intelligence
3-Months Course
at
Karakaroum international Univrsity

January 27, 2024

Table of Contents

Build a model

- In supervised learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used.
- If a model is able to make accurate predictions on unseen data, we say it is able to generalize from the training set to the test set.
- We want to build a model that is able to generalize as accurately as possible.

Build a model

- In supervised learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used.
- If a model is able to make accurate predictions on unseen data, we say it is able to generalize from the training set to the test set.
- We want to build a model that is able to generalize as accurately as possible.

Build a model

- In supervised learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used.
- If a model is able to make accurate predictions on unseen data, we say it is able to generalize from the training set to the test set.
- We want to build a model that is able to generalize as accurately as possible.

Problem formulation

- Convert your business problem into a Statistical problem
- Clearly define the dependent and independent variable
- Identify whether you want to predict or infer

Problem formulation

- Convert your business problem into a Statistical problem
- Clearly define the dependent and independent variable
- Identify whether you want to predict or infer

Problem formulation

- Convert your business problem into a Statistical problem
- Clearly define the dependent and independent variable
- Identify whether you want to predict or infer

Data frame

- Transform collected data into a useable data table format
- Example Pandas data frame

Data frame

- Transform collected data into a useable data table format
- Example Pandas data frame

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Pre-Processing

- Filter data
- Aggregate values
- Missing value treatment
- Outlier treatment
- Variable transformation
- Variable reduction

Train-Test split

Training data is the information used to train an algorithm. The training data includes both input data and the corresponding expected output. Based on this data, the algorithm can learn the relationship between input and output variables.
Testing data includes only input data, not the corresponding expected output. The testing data is used to assess the accuracy of model created or the predictor function created using the training data.

- There should not be any overlap between the two.
- Usually, 70-80 percentage of the available data is used as training data and 20-30 percentage as testing data

Train-Test split

Training data is the information used to train an algorithm. The training data includes both input data and the corresponding expected output. Based on this data, the algorithm can learn the relationship between input and output variables.
Testing data includes only input data, not the corresponding expected output. The testing data is used to assess the accuracy of model created or the predictor function created using the training data.

- There should not be any overlap between the two.
- Usually, 70-80 percentage of the available data is used as training data and 20-30 percentage as testing data

Model Building

Validation and Model Accuracy

Prediction

Great Job
Thank yo