Machine Learning

Data & Features

Dr. Sherif Saad

Learning Objectives

Introduce the students to machine learning concepts

Explain the main three types of learning and ML terminology

Understand the building blocks for successfully designing machine learning systems.

How to apply machine learning algorithms (not really how to create them)

Outlines

- Data and Features Basic Terminologies
- Feature Selection
- Data Preprocessing

Data & Features Basic Terminologies

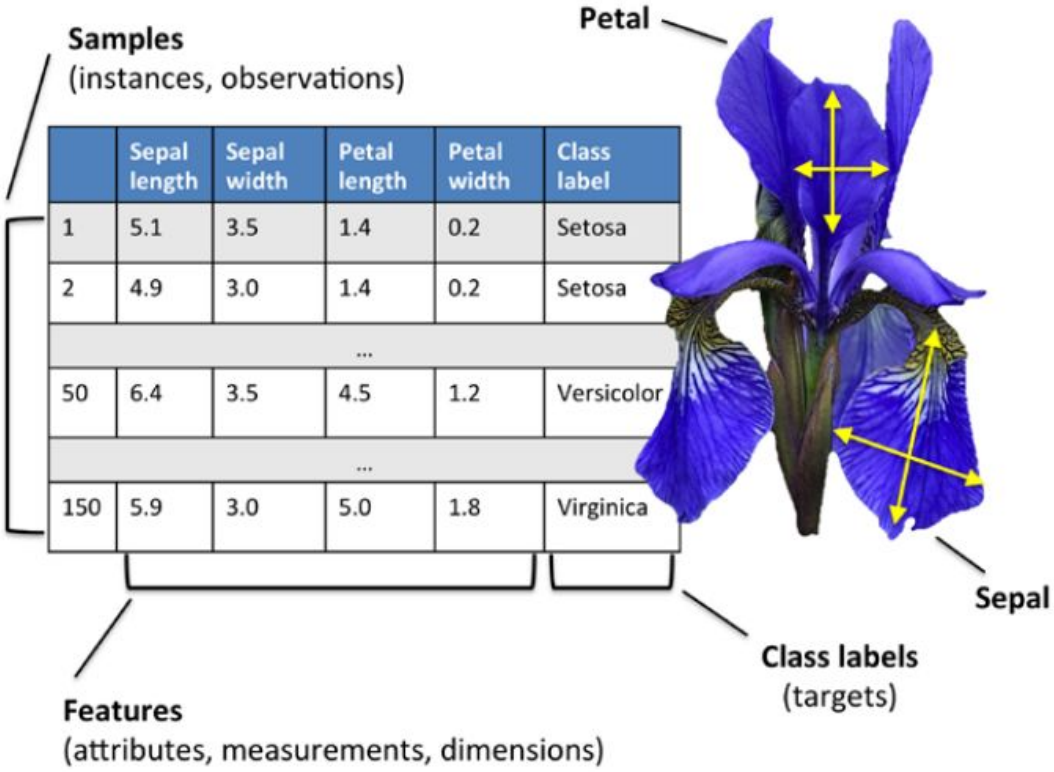
- Data in the context of machine learning is a collection of samples and each sample is a collection of features.
- The most common structure of the data is a matrix, where each row represent a sample and each column represent a feature.
- A sample is any phenomenon you can describe with quantitative traits.
- Features are those quantitative traits that describe your samples.
- A Dataset of Movies: each movie in this dataset is a sample, information about the movie like the movie length, release date, budget, cast, director, and any other information that describe | define the movie sample are called features.

Data & Features Basic Terminologies

Example:

- The Iris dataset, which is a classic example in the field of machine learning.
- The Iris dataset contains the measurements of 150 iris flowers from three different species: Setosa, Versicolor, and Virginica.
- Each flower sample represents one row in our data set, and the flower measurements in centimeters are stored as columns, which we also call the features of the dataset

Data & Features Basic Terminologies



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Data & Features Basic Terminologies

Features (aka)

- *Attribute* - Features are a quantitative attributes of the samples being observed
- *Axis* - Features are orthogonal axes of their *feature space*, if they are linearly independent
- *Column* - Features are represented as columns in your dataset
- *Dimension* - A dataset's features, grouped together can be treated as a *n-dimensional* coordinate space

Data & Features Basic Terminologies

Features (aka)

- *Input* - Feature values are the input of data-driven, machine learning algorithms
- *Predictor* - Features used to predict other attributes are called predictors
- *View* - Each feature conveys a quantitative trait or perspective about the sample being observed
- *Independent Variable* - Autonomous features used to calculate others are like independent variables in algebraic equations

Data & Features Basic Terminologies

Samples (aka)

- Instances
- Observations
- Signal

Class (aka)

- Target
- Labels

Types of Feature

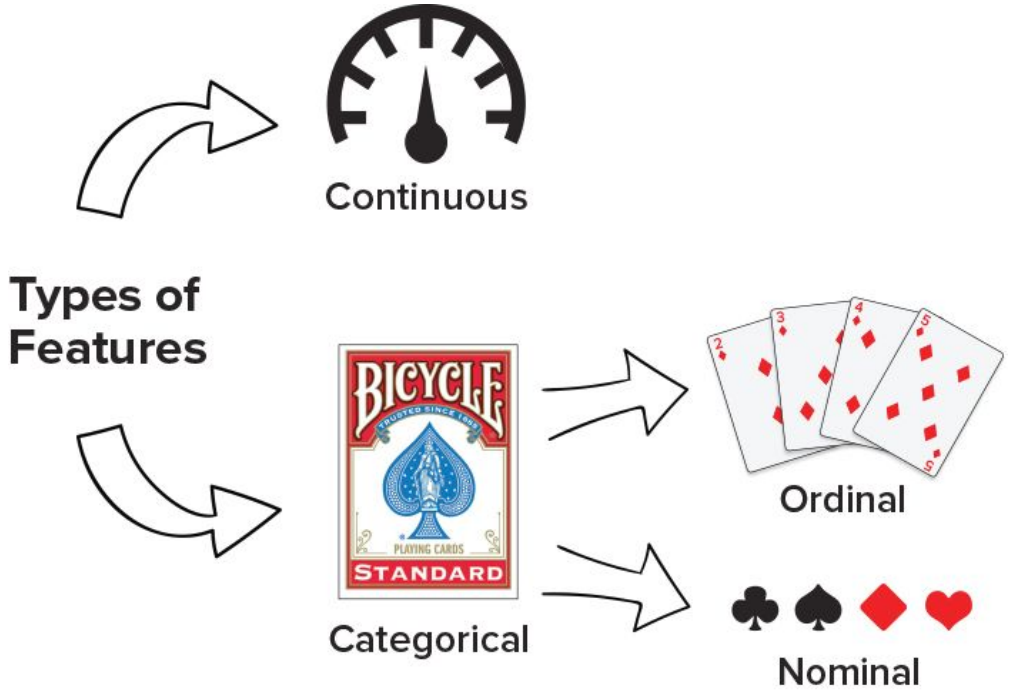


Image source: microsoft data science training course

Types of Features

Features can be generally divided into two main groups; continuous features and categorical features.

Continuous features are easy to measure and compare.

Categorical features are not easy to measure or compare.

If we think again about the movies dataset we can think of the movie length as a continuous feature that we can measure in unit of time (e.g minutes, hours). A categorical feature will be the movie genres. Furthermore, categorical features can be divided into nominal and ordinal features.

Types of Features

Continuous data is almost always represented with numeric features. But just because you have a numeric feature doesn't cause it to be continuous. There are times where you might have numerical categorical data.

Continuous features, there exist a measurable difference between possible feature values.

Categorical features, there is a specified number of discrete, possible feature values.

Types of Features

What happens if your dataset holds the age of 1000 people recorded in years? Should you treat it as continuous or as ordinal?

Do you think the feature encoding or representation affect its type?

Feature	Continuous	Categorical
Colors		
Stock Price		
Time		
Car Models		
High-Medium-Low		
GPA Letter Grade		
Weight		

Feature Selection

Collect as much data and features as you can during the data collection stage.

individual weak features (e.g. features with overlap, or not entirely independent) if combined sometime can form strong feature.

The data and features collection process should be guided by the problem we are trying to solve (e.g. which features are relevant to our problem).

In a disease diagnosis problem we can collect about the name, nationality, race, age, diet, income, occupation. Which of these features are more relevant to the disease diagnosis problem

Feature Selection

Feature selection is also called variable selection or attribute selection. It is mostly an automatic process that aims at selecting the most relevant features (attributes) in the dataset to the predictive modelling problem we are trying to solve.

Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

Should we focus on collecting more samples or more features and why?

Feature Selection

The Curse Dimensionality

Many ML algorithms are implemented as matrix operations, and without a greater than or equal number of samples to features, a fully-formed system of independent equations cannot be made. You can always create *more* features based on your existing ones. But creating *pseudo-samples*, while not impossible, might be a bit more difficult.

Fewer attributes is desirable because it *reduces the complexity of the model*, and a simpler model is simpler to understand and explain.

Feature Selection

Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, where as feature selection methods include and exclude attributes present in the data without changing them.

It is important to consider feature selection a part of the model selection process. If you do not, you may inadvertently introduce bias into your models which can result in overfitting.

Feature Selection Algorithms

Filter Methods:

- Apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset.
- The methods are often univariate and consider the feature independently, or with regard to the dependent variable.
- Chi squared test, information gain and correlation coefficient scores.

Feature Selection Algorithms

Wrapper Methods:

- Consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations
- The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm.
- Example: Recursive feature elimination algorithm

Feature Selection Algorithms

Embedded Methods:

- Learn which features best contribute to the accuracy of the model while the model is being created.
- Regularization algorithms are the most common type of embedded methods
- Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm
- Examples of regularization algorithms are the LASSO, Elastic Net and Ridge Regression

Next Time

Data Preprocessing (Tuesday Tutorial)

Supervised Learning