Machine Learning (CS4102)

Books:

1. Machine Learning by Tom Mitchell (ISBN 0070428077)
2. Machine Learning: A Probabilistic Perspective by Kevin P. Murphy
3. A Course in Machine Learning by Hal Daumé III
4. Pattern Recognition and Machine Learning by Chris Bishop (ISBN 0387310738)
5. Deep Learning (Adaptive Computation and Machine Learning series) by Ian Goodfellow, Yoshua Bengio, Aaron Courville, Francis Bach, 2017
6. Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei.
7. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar

Syllabus: Machine Learning (CS4102)

- Module 1: Overview: Motivation, Introduction of Machine Learning, Types of Machine Learning (3 hrs)
- Module 2: Mathematics: Probability, Bayes' rule, conditional probability, likelihood, hyperparameters, Bayesian method, Polynomial Models, Linear algebra (6 hrs)
- Module 3: Supervised learning: Regression Models, Support vector machines, Generative/discriminative learning, parametric/nonparametric learning, Multilayer Perceptron neural models, Gradient descent, Backpropagation, Batch processing, Learning rate, Cross validation, Overfitting, Regularization, Radial Basis function neural network, Principal component analysis (10 hrs)
- Module 4: Unsupervised learning: clustering, Kohonen Network, SOFM, dimensionality reduction, kernel methods; Advanced discussion on clustering and Gaussian Mixture Models, Expectation Maximization (6 hrs)
- Module 5: Reinforcement learning: Mathematical Formulation, Markov decision process (3 hrs)
- Module 6: Deep Learning Models: RBM, Autoencoder, CNN, Transfer Feature Learning of CNN, RNN, LSTM, GRU, GAN, Different types of GANs, Ensemble methods (10 hrs)
- Module 7: Intelligent Machines (2 hrs)

Motivation

- Machine learning is driven by several key motivations that make it a powerful and sought-after field:

1. **Automation**:
- Machine learning allows tasks (that traditionally require human intelligence)to be automated.
- This includes everything from recognizing objects in images to understanding and generating natural language.

2. Insight Extraction:
- Machine learning algorithms can analyze large amounts of data to uncover patterns, trends, and insights that are not immediately apparent to humans.
- This is particularly useful in fields such as finance, healthcare, and marketing.

3. Prediction and Forecasting:
- Machine learning models can make predictions and forecasts based on historical data.
- This capability is used in weather forecasting, stock market prediction, and many other applications where future outcomes need to be estimated.

Motivation

4. Personalization:

- Many modern applications use machine learning to personalize user experiences.
- For example, recommendation systems on platforms like Netflix and Amazon use machine learning to suggest content based on a user's past behavior.

5. Optimization:

- Machine learning can optimize complex systems by learning from data and improving its performance over time.
- This is used in areas such as supply chain management, resource allocation, and logistics.

6. Pattern Recognition:

- Machine learning excels at recognizing patterns in data that may be too complex for humans to discern.
- This includes medical diagnostics, fraud detection, and quality control in manufacturing.

Motivation

7. Scalability:

- Machine learning algorithms can be applied to large datasets and problems that would be impractical for humans to handle manually.
- This scalability is crucial in fields like image and speech recognition, where massive amounts of data are processed.

8. Adaptability:

- Machine learning models can adapt to new data and changing circumstances, making them versatile in dynamic environments.
- This adaptability is valuable in applications like autonomous vehicles and real-time language translation.

Motivation

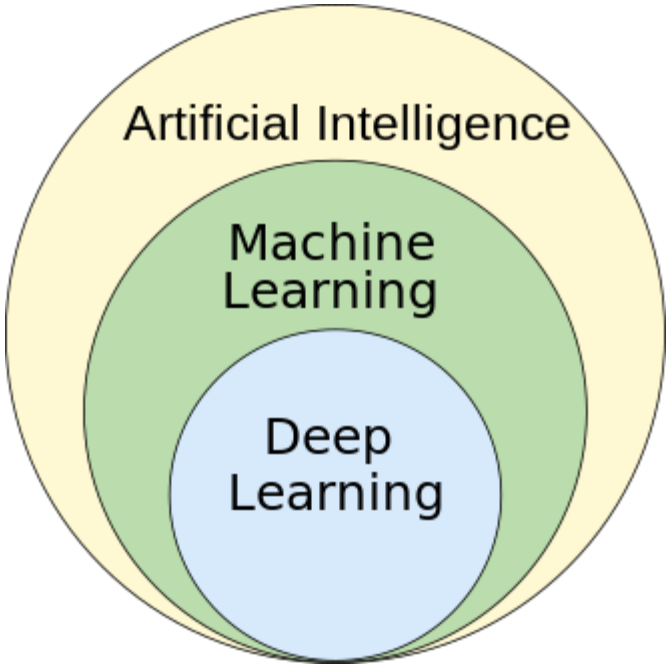
- Overall, the motivation behind machine learning is to leverage data-driven algorithms to solve complex problems more efficiently and effectively than traditional methods allow.
- Its applications span across industries and continue to grow as technology advances and datasets become larger and more accessible.

Machine Learning (ML)

- Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term “Machine Learning ” in 1959 while at IBM.
- He defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed “.
- However, there is no universally accepted definition for machine learning.

What is Machine Learning?

- Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.
- A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, as measured by P, improves with experience E.



Handwriting recognition learning problem

- Task T : Recognizing and classifying handwritten words within images
- Performance P : Percent of words correctly classified
- Training experience E : A dataset of handwritten words with given classifications

A robot driving learning problem

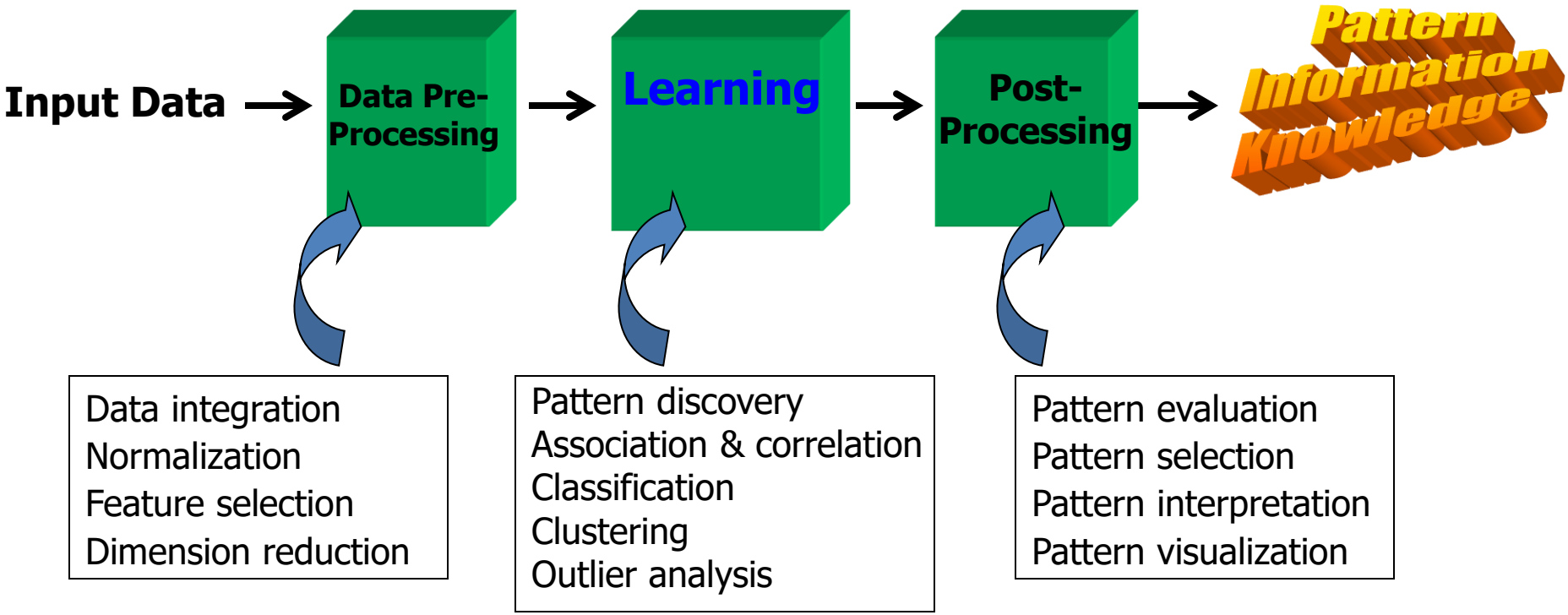
- Task T : Driving on highways using vision sensors
- Performance P : Average distance traveled before an error
- Training experience E : A sequence of images and steering commands recorded while observing a human driver

ML-Definition

- A computer program which learns from experience is called a machine learning program or simply a learning program.

- Experience => Dataset
- Learning =>knowledge discovery from data
- Knowledge discovery => Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

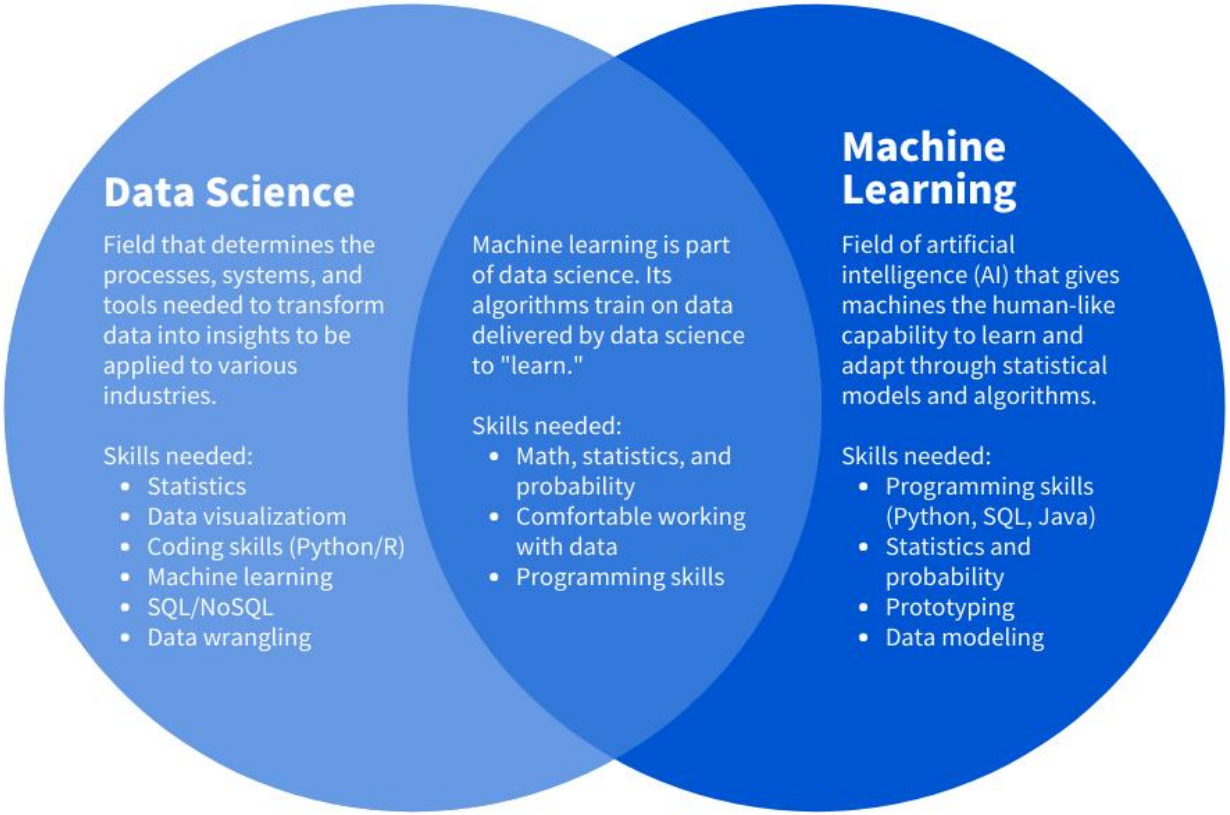
A Typical View of ML



This is a view from typical machine learning and statistics communities

Is Machine Learning and Data Science same?

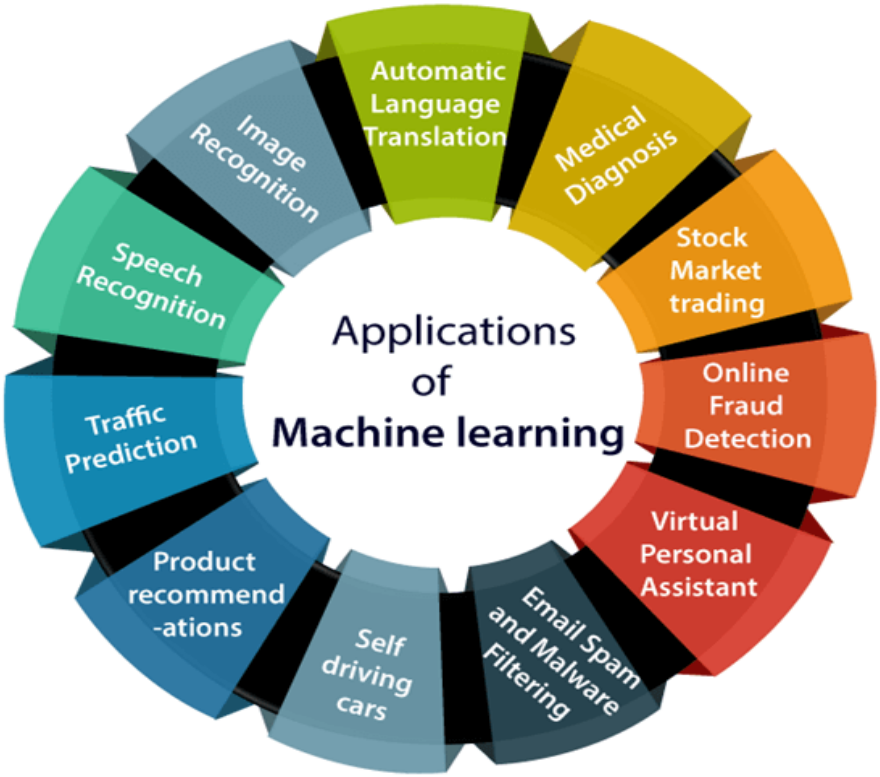
- While data science and machine learning are related, they are very different fields.
- The two relate to each other in a similar way that squares are rectangles, but rectangles are not squares. Data science is the all-encompassing rectangle, while machine learning is a square that is its own entity.
- In a nutshell, data science brings structure to big data while machine learning focuses on learning from the data itself.
- Data science focuses on managing, processing, and interpreting big data to effectively inform decision-making. Machine learning leverages algorithms to analyze data, learn from it, and forecast trends.



- In recent years, machine learning and artificial intelligence (AI) have dominated parts of data science, playing a critical role in data analytics and business intelligence.
- Data Science and Machine Learning are often used by data scientists in their work and are rapidly being adopted by nearly every industry.

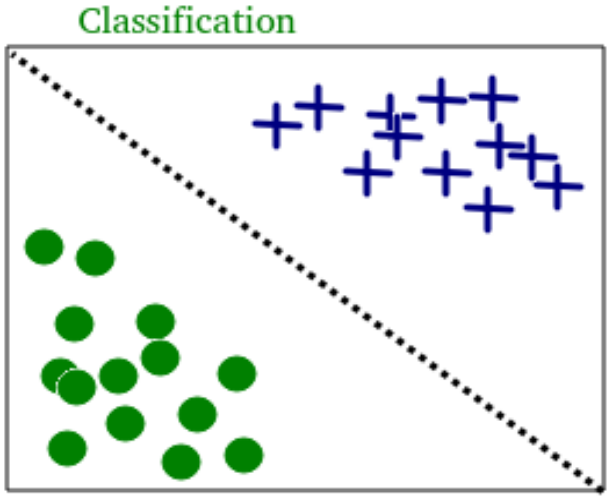
Applications of Machine Learning

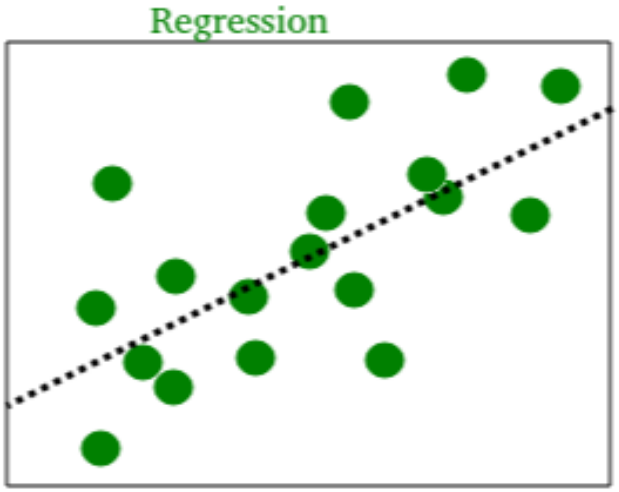
- Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day.
- We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc.
- Below are some most trending real-world applications of Machine Learning:



Types of machine learning problems

1. **Supervised learning**: This type of ML involves supervision, where machines are trained on labeled datasets and enabled to predict outputs based on the provided training.
- Supervised machine learning is further classified into two broad categories: (i) Classification, and (ii) Regression





Types of machine learning problems

(i) **Classification**: These refer to algorithms that address classification problems where the output variable is categorical; for example, yes or no, true or false, male or female, etc.

- Real-world applications of this category are evident in spam detection and email filtering.
- Some known classification algorithms include the Decision Tree Algorithm, Logistic Regression Algorithm, Support Vector Machine Algorithm, and Random Forest Algorithm.

Types of machine learning problems

(ii) Regression: Regression algorithms handle regression problems where input and output variables have a linear/non-linear relationship.

- These are known to predict continuous output variables.
- Examples include weather prediction, market trend analysis, etc.
- Popular regression algorithms include the Simple Linear Regression Algorithm, Multivariate Regression Algorithm, Decision Tree Algorithm, Lasso Regression, etc.

Types of machine learning problems

2. Unsupervised machine learning:

- Unsupervised learning refers to a learning technique that’s devoid of supervision.
- Here, the machine is trained using an unlabeled dataset and is enabled to predict the output without any supervision.
- An unsupervised learning algorithm aims to group the unsorted dataset based on the input’s similarities, differences, and patterns.

Types of machine learning problems

- Unsupervised machine learning is further classified into two types:
- (i) **Clustering**: The clustering technique refers to grouping objects into clusters based on parameters such as similarities or differences between objects.
- For example, grouping customers by the products they purchase.
- Some known clustering algorithms include the K-Means Clustering Algorithm, Mean-Shift Algorithm, DBSCAN Algorithm, Principal Component Analysis, and Independent Component Analysis.

Types of machine learning problems

(ii) Association: Association learning refers to identifying typical relations between the variables of a large dataset.

- It determines the dependency of various data items and maps associated variables. Typical applications include web usage mining and market data analysis.
- Popular algorithms obeying association rules include the Apriori Algorithm, Eclat Algorithm, and FP-Growth Algorithm.

Types of machine learning problems

3. Semi-supervised learning comprises characteristics of both supervised and unsupervised machine learning.

- It uses the combination of labeled and unlabeled datasets to train its algorithms.
- Consider an example of a college student.

  - A student learning a concept under a teacher's supervision in college is termed supervised learning.

  - In unsupervised learning, a student self-learns the same concept at home without a teacher's guidance.

  - Meanwhile, a student revising the concept after learning under the direction of a teacher in college is a semi-supervised form of learning.

How Semi-Supervised Learning Works?

- **Learning Process**:

1. **Initial Training**: The model begins with a small set of labeled data and learns from it using standard supervised learning techniques.
2. **Utilizing Unlabeled Data**: The model then incorporates unlabeled data during training to generalize better and improve performance. This is typically done by enforcing consistency across predictions made on unlabeled data.

- **Assumptions**: Semi-supervised learning relies on assumptions like the smoothness assumption, where points close in the input space are likely to have the same output label, and the cluster assumption, where points in the same cluster share the same label.

Motivations for Semi-Supervised Learning

1. **Cost Efficiency**: Labeled data can be expensive and time-consuming to acquire, especially in domains like medical imaging or natural language processing. By using unlabeled data along with a smaller set of labeled data, semi-supervised learning reduces the need for extensive labeling efforts.
2. **Scarcity of Labeled Data**: In many real-world applications, large amounts of unlabeled data are available, but labeling all of them may not be feasible. Semi-supervised learning allows leveraging this abundant unlabeled data effectively.
3. **Performance Improvement**: Incorporating unlabeled data often improves the generalization ability of models. By learning from a larger, more diverse dataset (combining labeled and unlabeled data), models can capture underlying structures and variations better.
4. **Domain Adaptation**: Semi-supervised learning can help adapt models trained on labeled data from one domain to perform well in related but different domains where labeled data is scarce or unavailable.
5. **Ethical Considerations**: In certain applications, labeling data might involve sensitive information or ethical concerns. Semi-supervised learning can potentially reduce the need for extensive labeling, thereby minimizing exposure to sensitive data.

Applications

- **Natural Language Processing**: Sentiment analysis, machine translation, and text classification benefit from semi-supervised techniques due to the abundance of unlabeled text data.
- **Image and Video Analysis**: Object recognition, video classification, and segmentation tasks can use semi-supervised learning to enhance accuracy with limited labeled datasets.
- **Bioinformatics**: Analyzing genetic sequences, drug discovery, and medical imaging often have limited labeled data, making semi-supervised learning beneficial.

- In summary, semi-supervised learning expands the scope of machine learning by effectively utilizing both labeled and unlabeled data to improve model performance, reduce labeling costs, and address challenges posed by limited labeled data in various domains.

Types of machine learning

4. **Reinforcement learning (RL):**

- It is a type of machine learning where an agent learns to make decisions by interacting with an environment.
- It learns from trial and error, receiving feedback in the form of rewards or penalties for actions it takes.

Components of Reinforcement Learning

1. **Agent**: The entity that takes actions in an environment. It perceives the environment through observations and selects actions to maximize cumulative rewards.
2. **Environment**: The external system with which the agent interacts. It responds to the actions of the agent and provides feedback (reward or penalty) based on the actions taken.
3. **Action (a)**: Choices made by the agent at each time step, influencing the environment.
4. **State (s)**: A representation of the current situation of the agent within the environment. It captures relevant information that the agent needs to decide on its next action.
5. **Reward (r)**: A scalar feedback signal from the environment to the agent after each action. It indicates the immediate benefit or detriment of the action taken by the agent.

Components of Reinforcement Learning

6. **Policy (π)**: The strategy or rule that the agent uses to select actions (choices) based on the current state. It maps states to actions and can be stochastic or deterministic.
7. **Value Function (V(s))**: The expected cumulative reward an agent can expect to receive starting from a particular state. It helps the agent evaluate the goodness of states or state-action pairs.
8. **Q-function (Q(s, a))**: Similar to the value function but for state-action pairs. It estimates the expected cumulative reward starting from state s, taking action a, and then following the policy thereafter.

How Reinforcement Learning Works?

- **Initialization**: The agent starts in a particular state within the environment.
- **Action Selection**: Based on the current state and the policy, the agent selects an action to execute.
- **Interaction with Environment**: The agent executes the chosen action, which changes the state of the environment.
- **Observation and Reward**: The agent receives feedback from the environment in the form of a reward signal, indicating the immediate consequence of the action.
- **Learning and Policy Update**: Using this feedback, the agent updates its policy or value function to improve its decision-making ability in subsequent interactions.
- **Exploration vs. Exploitation**: RL algorithms balance between exploring new actions to discover potentially better strategies and exploiting known good actions to maximize immediate rewards.

Applications of Reinforcement Learning

- **Game Playing**: RL has been successfully applied to games like Chess, Go, and video games, where agents learn optimal strategies by playing against themselves or human players.
- **Robotics**: RL enables robots to learn complex tasks such as grasping objects, navigating environments, and even locomotion.
- **Autonomous Systems**: RL is used in autonomous vehicles, drones, and other autonomous agents to make decisions in dynamic environments.
- **Finance**: RL algorithms are used in algorithmic trading to optimize portfolio management and make trading decisions.
- **Healthcare**: RL can optimize treatment strategies, personalized medicine, and resource allocation in healthcare settings.

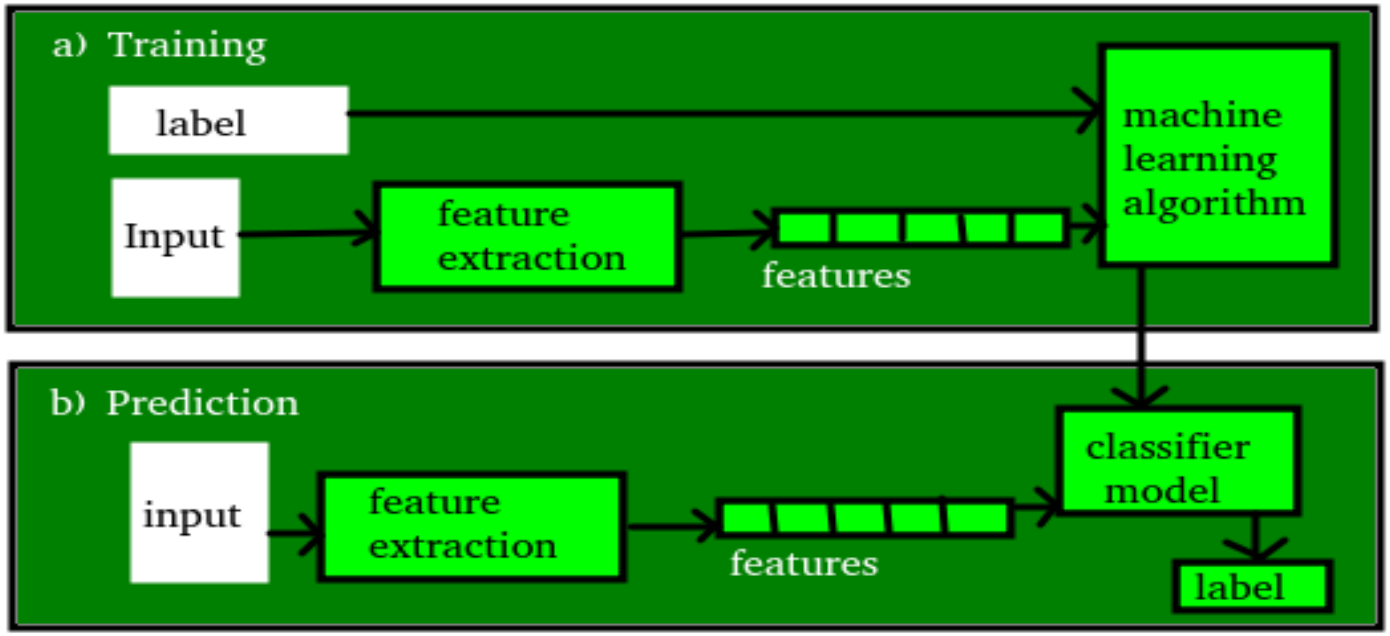
- Reinforcement learning continues to advance with new algorithms and applications, playing a significant role in creating intelligent agents capable of learning and adapting to complex environments.

Terminologies of Machine Learning

- **Model:** A model is a **specific representation** learned from data by applying some machine learning algorithm. A model is also called a **hypothesis**.
- **Feature:** A feature is an individual measurable property of the data. A set of numeric features can be conveniently described by a **feature vector**. Feature vectors are fed as input to the model. For example, in order to predict a fruit, there may be features like color, smell, taste, **etc.**
- **Target (Label):** A target variable or label is the value to be predicted by the model. For example, the label of fruit object be the name of the fruit like apple, orange, banana, etc.

Terminologies of Machine Learning

- **Training:** The idea is to give a set of inputs(features) and its expected outputs(labels), so after training, we will have a model (hypothesis) that will then map new data to one of the categories trained on.
- **Prediction** Once the model is ready, it can be fed a set of inputs to which it will provide a predicted output(label). But make sure if the machine performs well on unseen data, then only we can say the machine performs well.



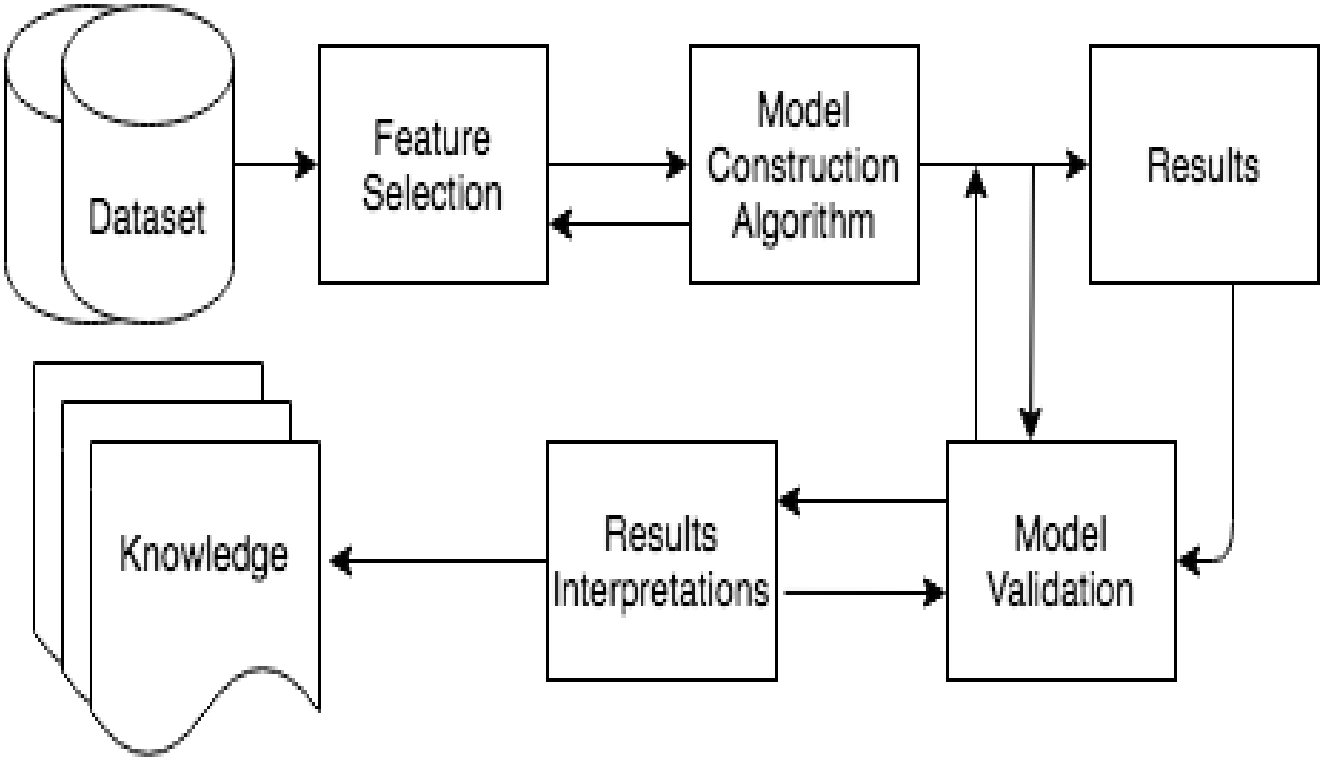
Steps to get started with machine learning

1. ***Define the Problem*:** Identify the problem you want to solve and determine if machine learning can be used to solve it.
2. ***Collect Data*:** Gather and clean the data that you will use to train your model. The quality of your model will depend on the quality of your data.
3. ***Explore the Data:*** Use data visualization and statistical methods to understand the structure and relationships within your data.
4. ***Pre-process the Data*:** Prepare the data for modeling by normalizing, transforming, and cleaning it as necessary.
5. ***Split the Data*:** Divide the data into training and test datasets to validate your model.
6. ***Choose a Model*:** Select a machine learning model that is appropriate for your problem and the data you have collected.

Steps to get started with machine learning

7. ***Train the Model:*** Use the training data to train the model, adjusting its parameters to fit the data as accurately as possible.
8. ***Evaluate the Model:*** Use the test data to evaluate the performance of the model and determine its accuracy.
9. ***Fine-tune the Model:*** Based on the results of the evaluation, fine-tune the model by adjusting its parameters and repeating the training process until the desired level of accuracy is achieved.
10. ***Deploy the Model:*** Integrate the model into your application or system, making it available for use by others.
11. ***Monitor the Model:*** Continuously monitor the performance of the model to ensure that it continues to provide accurate results over time.

Basic Steps of ML



A Machine Learning Application

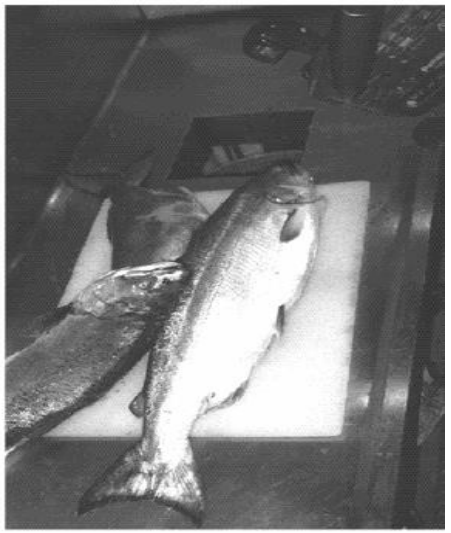
(from *Pattern Classification by* Duda & Hart & Stork – Second Edition, 2001)

- A fish-packing plant wants to automate the process of sorting incoming fish according to species

- As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

Classification

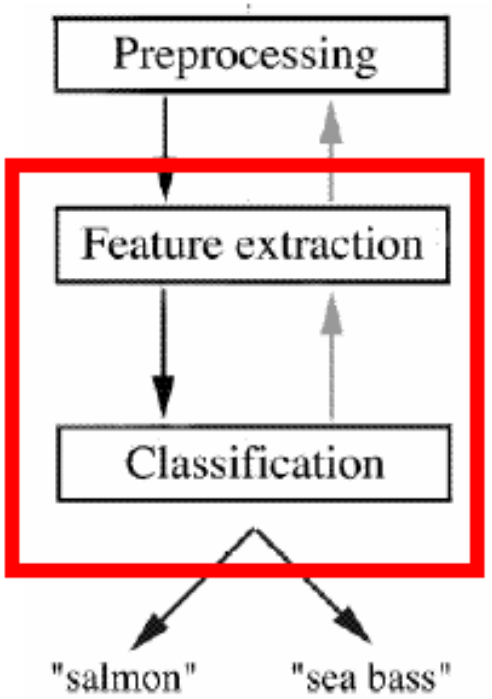
- Features (to distinguish):
  - (i) Length
  - (ii) Lightness
  - (iii) Width
  - (iv) Position of mouth



Classification

- **Preprocessing:** Images of different fishes are isolated from one another
  and from background;

- **Feature extraction:** The information of a single fish is then sent to a feature extractor, that measure certain "features" or "properties";



- **Classification:** The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

Classification

- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature*:* (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier *learns* itself from the training examples, how to distinguish Salmon from Bass based on the *fish_length*

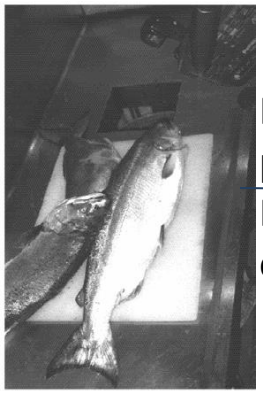
Classification

- Classification model (hypothesis):
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like:

    Rule: If *Length >= l* then sea bass* otherwise *salmon*

  - Here the value of *l** determined by the classifier
- Testing the model
  - Once we get a model out of the classifier, we may use the classifier to test future examples.
  - The test data is provided in the form <fish_length>
  - The classifier outputs <fish_type> by checking *fish_length* against the model

Classification



Training data

Pre-processing, Feature extraction

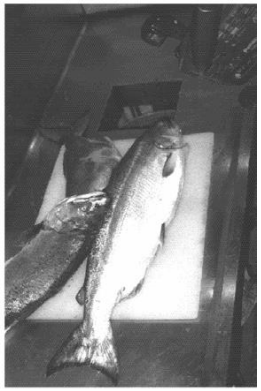
12, salmon
15, sea bass
8, salmon
5, sea bass

Feature vector

Training

If len > 12, ***then*** sea bass ***else*** salmon

Model



Test data

Pre-processing,

Feature extraction

Unlabeled data

Labeled data

15, salmon
10, salmon

18, ?
8, ?

Feature vector

Test/ Classify

sea bass (error!)
salmon (correct)

sea bass
salmon

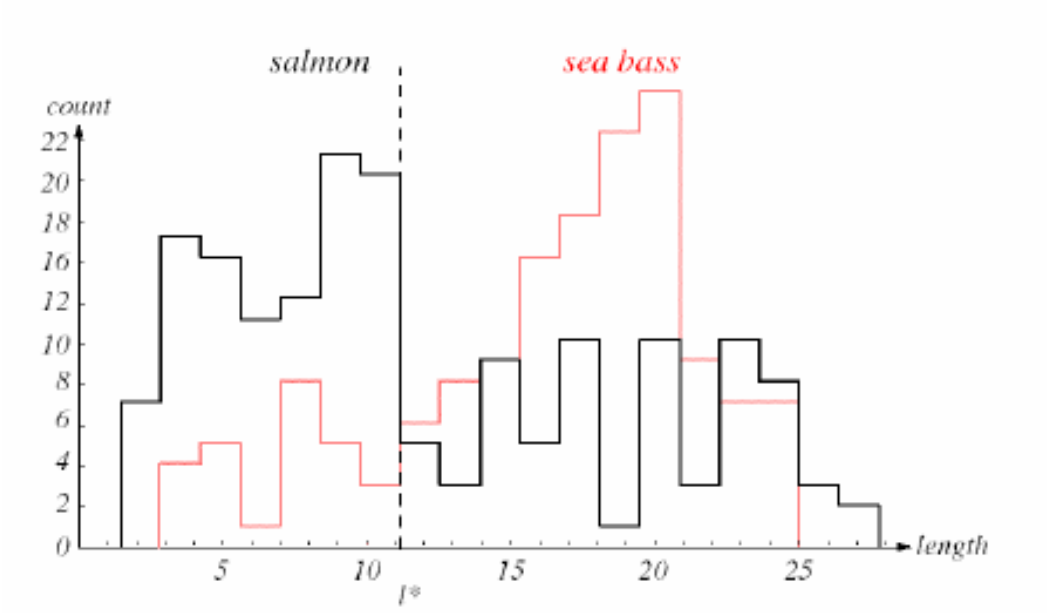
Evaluation/Prediction

Classification

- Why error?
  - Insufficient training data
  - Too few features
  - Too many/irrelevant features
  - Overfitting / specialization

Classification

Histograms of the length feature for the two categories

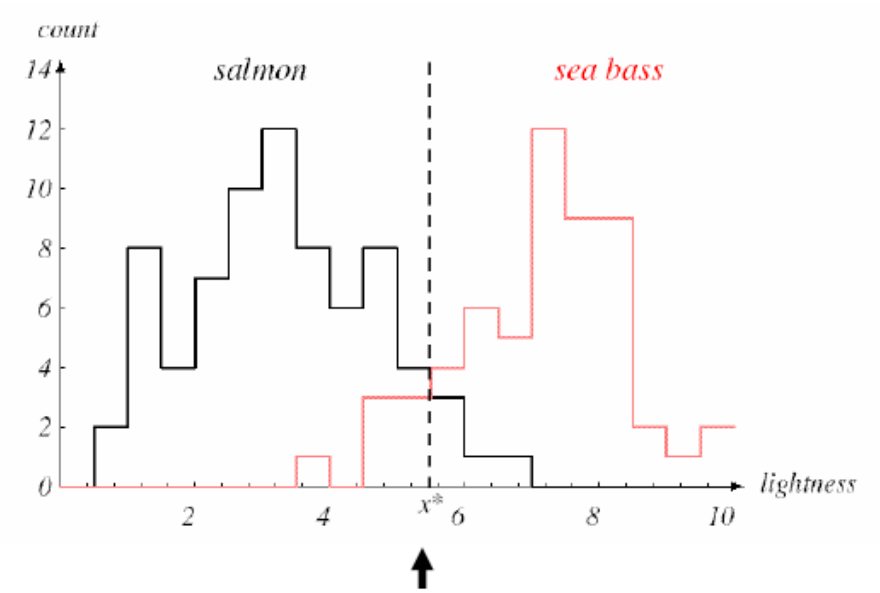


Leads to the smallest number of errors on average

We cannot reliably separate sea bass from salmon by length alone!

- New Feature*:*
  - *Average lightness of the fish scales*

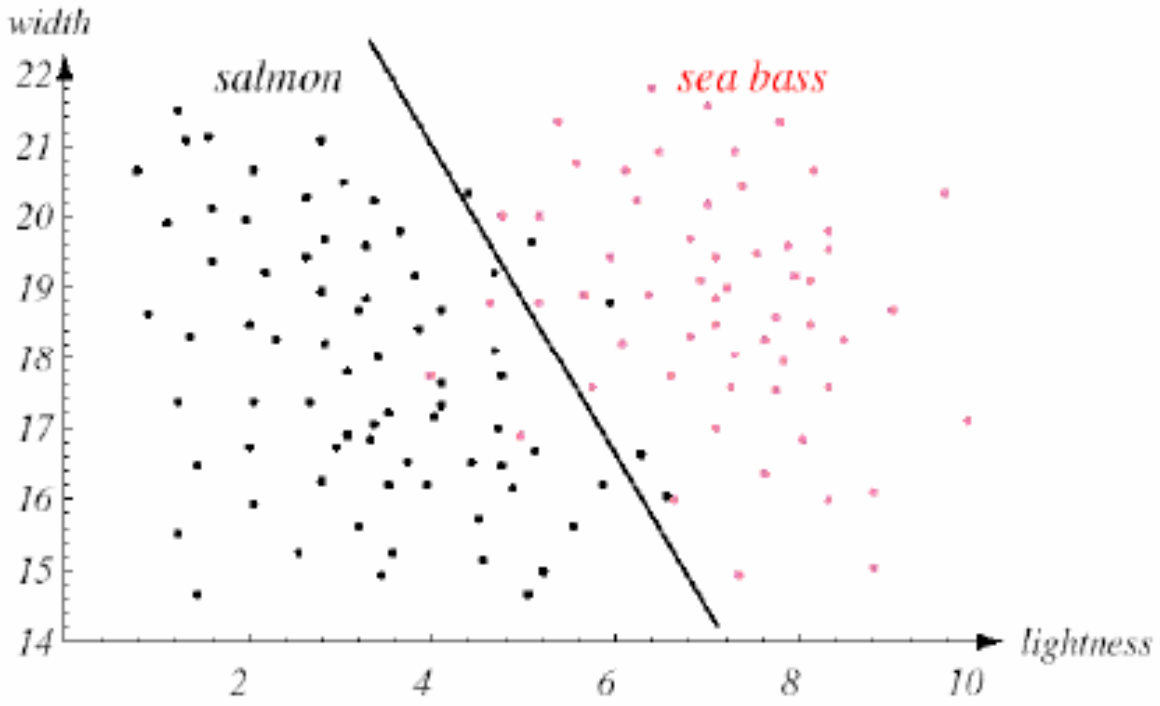
Histograms of the lightness feature for the two categories



Leads to the smallest number of errors on average

The two classes are much better separated!

Classification



Decision rule: Classify the fish as a sea bass if its feature vector falls above the decision boundary shown, and as salmon otherwise

Classification

Training data

Pre-processing,

Feature extraction

12, 4, salmon
15, 8, sea bass
8, 2, salmon
5, 10, sea bass

Feature vector

Training

If ltns > 6 **or** len*5+ltns*2>100 ***then*** sea bass ***else*** salmon

Model

Test data

Pre-processing,

Feature extraction

15, 2, salmon
10, 7, salmon
18, 7, ?
8, 5, ?

Feature vector

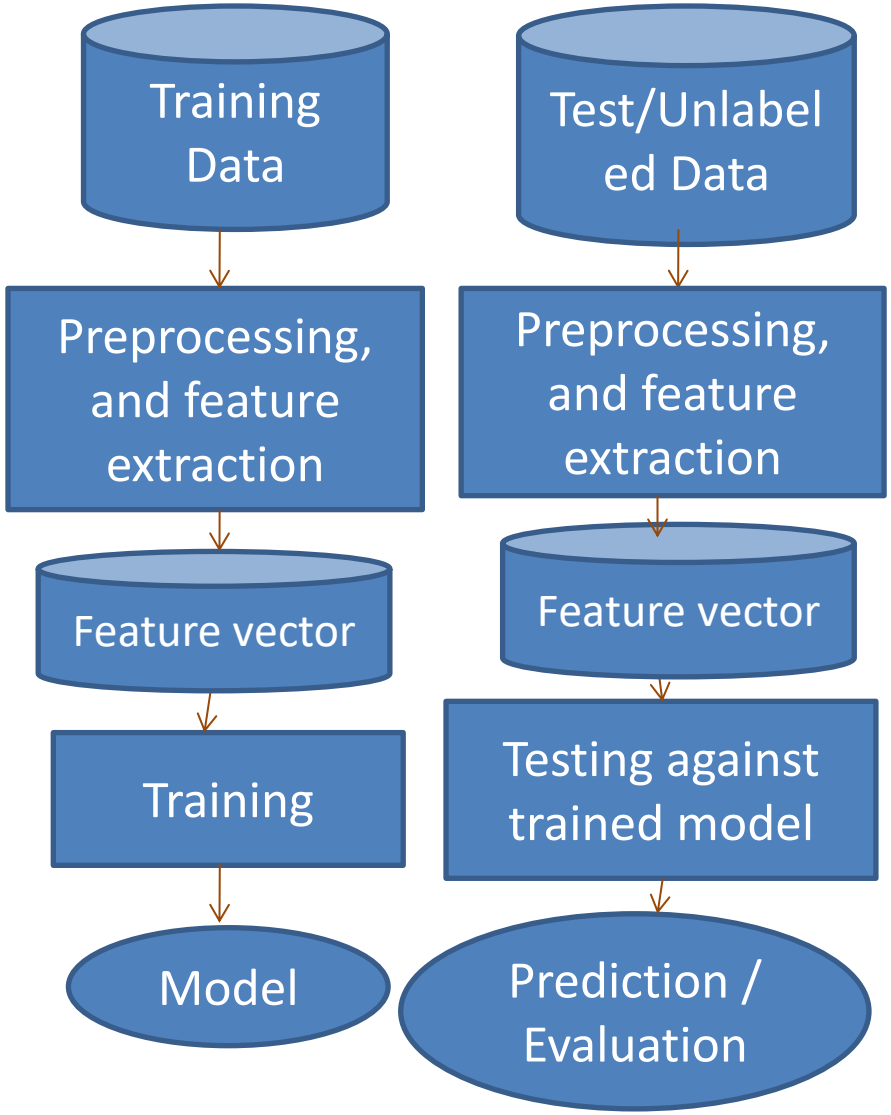
Test/ Classify

salmon (correct)
salmon (correct)
sea bass
salmon

Evaluation/Prediction

Validation

- The overall classification process goes like this →



Example in Python

```
# Load the necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC

# Load the iris dataset
df = pd.read_csv('iris.csv')

# Split the data into features and labels
X = df[['sepal_length', 'sepal_width', 'petal_length', 'petal_width']]
y = df['species']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Create an SVM model and train it
model = SVC()
model.fit(X_train, y_train)

# Evaluate the model on the test data
accuracy = model.score(X_test, y_test)

print('Test accuracy:', accuracy)
```

output:

Test accuracy: 0.9666666666666667

Thank you