Machine Learning Basics

1. Definition of ML
2. Types of ML
3. Challenges of ML

Definition of Machine Learning

“[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed”

- Arthur Samuel, 1959

Definition of Machine Learning

- "A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**"
    - Tom Mitchell, 1997

Mathematical Formulation of ML

- Given some input **X** (say age, mileage of a car), we want to find the output **y** (say price of a car)
  - **Y = f(X)**
- We do not know what the function **f** is, and we use machine learning to help find that function

Machine Learning Terms

- The training data is the **X** and **y** that we use to find **f**
- A feature / predictor is a measurable characteristic we use to make predictions
  - For example, if we are predicting car prices, our features may be age, gas mileage
- The sample size is the number of training instances we use to train our model

Types of ML

Categories of Machine Learning

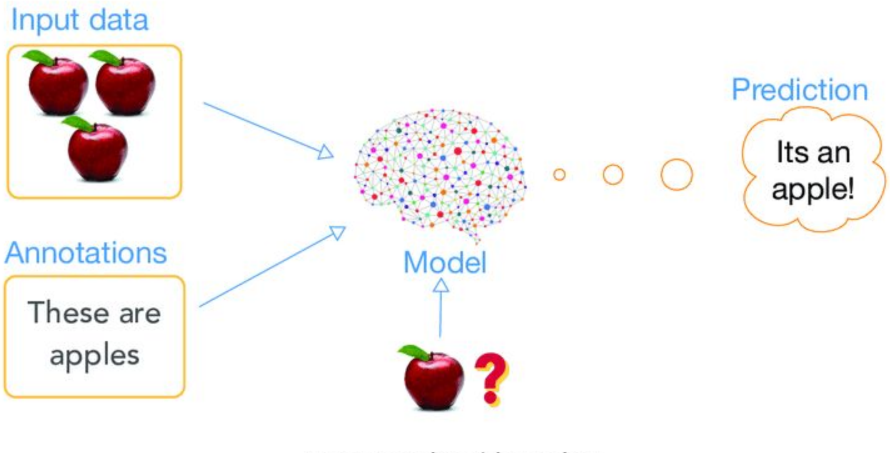
- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised Learning

- In supervised learning, the training data you feed to your ML algorithm includes both the inputs and the desired solutions, called the labels
- Is used for classification, regression

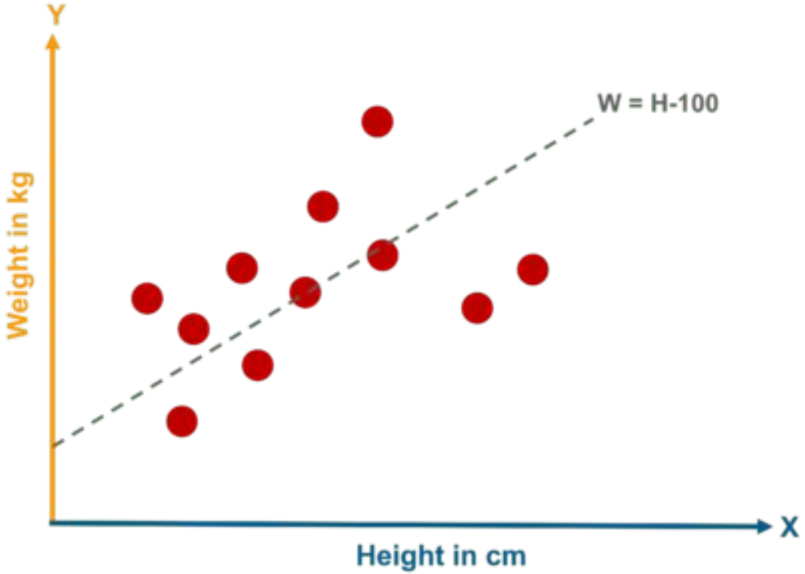
Classification

- In a classification task, we want to predict a discrete item



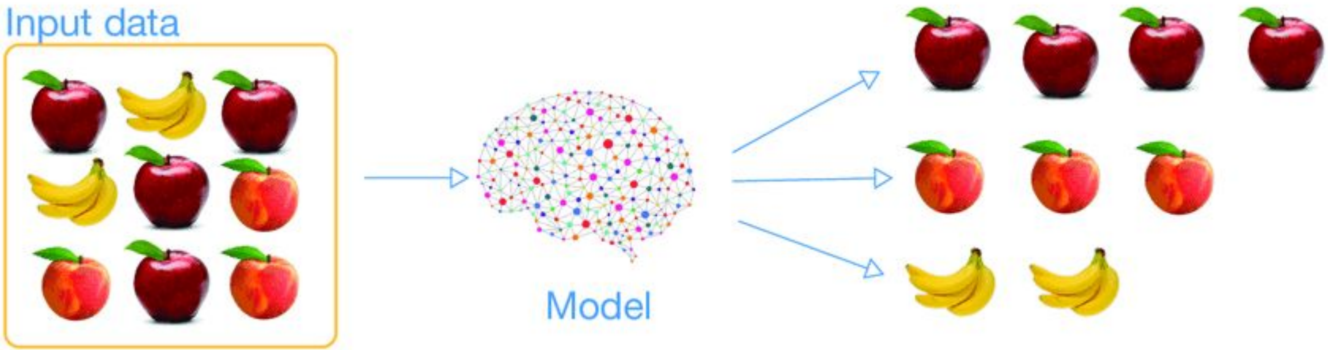
Regression

- In a regression task, we want to predict a real valued



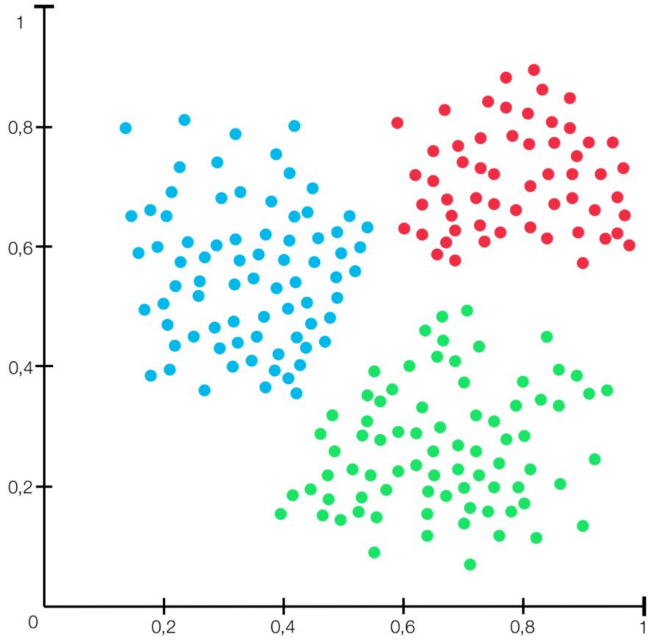
Unsupervised Learning

- In unsupervised learning, we do not have any labels. The system learns without a teacher.



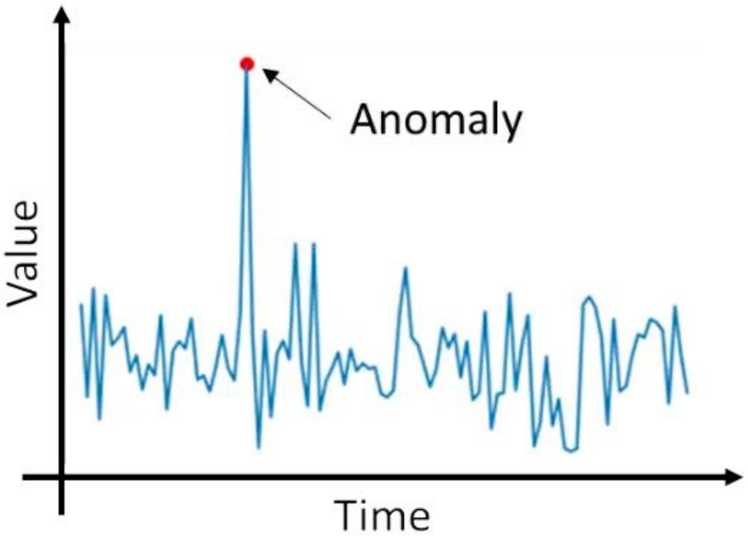
Clustering

- In <u>clustering</u>, we detect groups of similar instances



Anomaly Detection

- In anomaly detection, we try to detect if an instance is normal or an anomaly



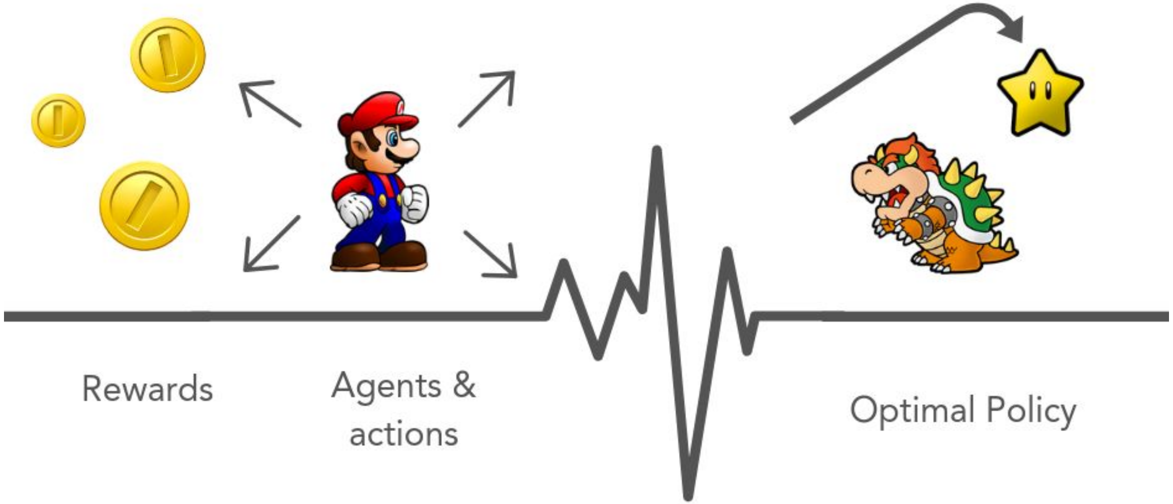
Semi-supervised Learning

- In semi-supervised learning, we learn with some labeled data as well as unlabeled data

Reinforcement Learning

- In reinforcement learning, the learning system (a.k.a the agent) observes the environment, selects and performs actions, and gets rewards in return.
- The agent learns by itself what the best strategy (a.k.a. policy) to get the most reward over time.

Reinforcement Learning



Machine Learning Challenges

Lack of Training Data

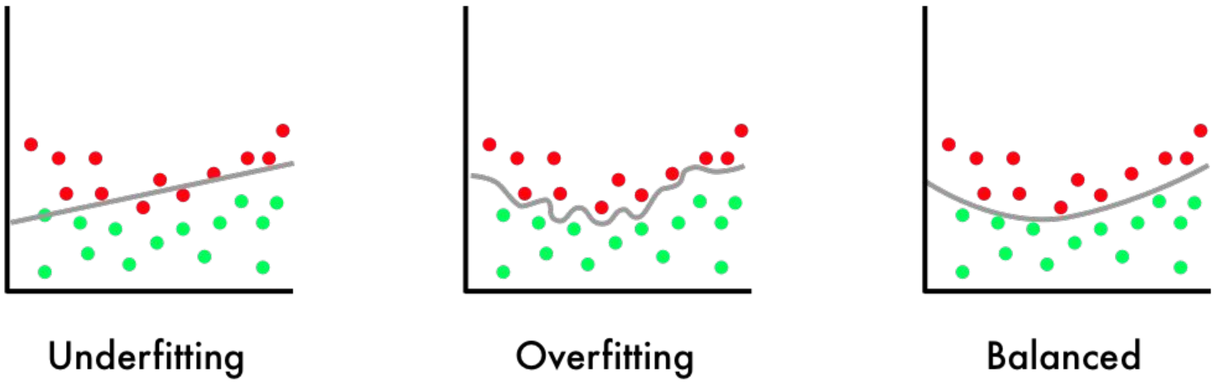
Bad Data


BAD DATA IS...
Duplicate data
Missing data
Inaccurate data
Incorrect data

Overfitting and Underfitting

- Overfitting occurs when your model performs well on the training data but does not generalize well
- Underfitting occurs when the model is too simple to capture the underlying structure of your data

Overfitting and Underfitting

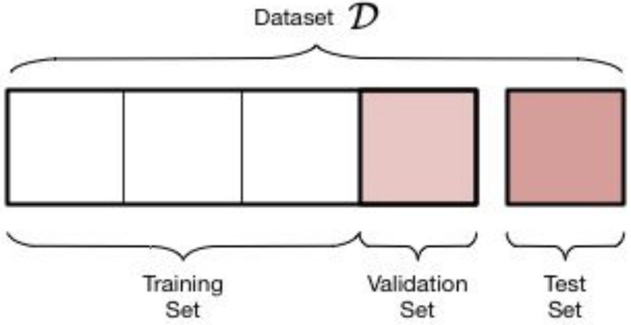


Overfitting and Underfitting Solutions

- Overfitting solutions
  - Simplify or constrain the model
  - Gather more data
  - Reduce the noise in the data
- Underfitting solutions
  - Select a more powerful model
  - Reduce the constraints on the model
  - Feed better features to the learning algorithm

Training, Validating, and Testing

- Generally, we want to split our data into a training, validating, and testing set
- We would train our model on the train set, perform hyperparameter tuning on the validation set, and finally evaluate our model on the test set



No Free Lunch Theorem

- Why can't we use a single framework (i.e. neural networks) for all possible datasets?
- No Free Lunch Theorem (1996): David Wolpert proved that if you make no assumptions about the data, it is impossible to know a priori which model works best

Prediction Accuracy and Model Interpretability Trade-off

- Generally, a model that is more accurate is less interpretable and visa versa

White-box vs Black-box Models

- White-box models are models that are intuitive, transparent, and its decisions are easy to interpret
- Black-box models are models that are harder to interpret and understand why it made its decisions
- Generally, black-box models perform better

Questions to Answer

1. Indicate whether we would expect a flexible model to be better or worse than an inflexible model
   a. The sample size is large and the number of features is small
   b. The relationship between the predictor and response is highly non-linear
2. Under what circumstances would a more flexible model be more useful than a less flexible model and visa versa?
3. What can we do if our model has high performance on the training data but poor performance on the test data?
4. Why do we want to split the data into a train, validate, and test set?
5. When would a black-box model be more useful than a white-box model?
6. Should you drop duplicate data points when training your models?