**AIML unit 2 – lessons up to LA recap**

***What is Machine Learning?***

Machine Learning (ML) is the study of algorithms that improve automatically through experience over a **training set** of data

Problems are classified in two categories:

* Problems with algorithmic solution

Encoding **a set of rules** that solve the problem

* Problems **without** an algorithmic solution

*ML is about writing the* ***optimization*** *for the* ***hypothesis***

***Terminologies***

The data set for learning comes from past experience, the "algo" adapts to unseen data, conforming it to the hypothesis and making a prediction (e.g., pattern finding in data)

An ML algorithm is called a **model**, the pattern finding is called the **learning** of a model

Generally, the model improves after learning from the data, learning things that cannot be learnt directly from data

**Limits of ML – Causality and Correlation**

Simply put, Correlation is when two things happen together, while Causation is when one thing causes another thing to happen.

Important concepts:

* Prob/Stat: correlation coefficient, covariance, standard deviation, mean

***The inductive bias***

The inductive bias (also known as learning bias) of a learning algorithm is the set of assumptions that the learner uses to predict outputs of given inputs that it has not encountered.

***The Correlation coefficient***

The correlation coefficient exists to numerically measure the relationship between two variables, i.e. how two variables are *linearly* related

Informally, two variables are *dependent* if they are *correlated* and viceversa

***A picture containing text

Description automatically generated***One of the most important correlation coefficients is the ***Pearson correlation coefficient:***

Where:

* Cov(X,Y) is the **covariance**
* 𝜎𝑋 is the **standard deviation** of X
* 𝜎Y is the **standard deviation** of Y

Chart, scatter chart

Description automatically generated

**Remark from Probability:**

* Cov(X,Y) = E[(X – E[X])(Y – E[Y])]

Cov(X,Y) = E[XY] – E[X]E[Y]

* 𝜎X = *√* E[X²] – (E[X]²)

PCC will be between 1 and -1, so a *linear equation* describes the relationship between X and Y, with all data points on a line

PCC = 0 means that there’s **no correlation** between the variables

Text

Description automatically generated

Where:

* i is the index of which I take x, y
* x̄, ȳ are the (standard) means, respectively, of x and y

Finally, on an important note **correlation does not imply causation** (beware of spurious correlation)

**Learning Paradigms**

We will focus on two learning paradigms:

* Supervised learning
* Unsupervised learning

***Supervised learning – Hypothesis over an Hypotheses set***

SL is a ML paradigm for problems where data consists of labeled examples, with each datum containing **features** and an **associated label**

* The labelled data means some input data is already tagged with the correct output
* The training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly

**The goal is to learn a function that maps features to labels over a set of possible functions,** aka predict the label when given as input unseen data

Diagram

Description automatically generated

The two types of SL algorithms are:

* **Regression:**

used for the prediction of **continuous variables** (∃ relationship between input and output)

* **Classification:**

used when the output variable is **categorical** (has classes: yes/no, male/female)

***The Bayes optimal classifier***

***What is the most probable classification of the new instance given the training data?***

**The Bayes optimal classifier is a probabilistic model that makes the most probable prediction for a new example, given the training dataset**

In general, the most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities

**Text, letter

Description automatically generated**

And D is an **unknown generator process** (we will refer to this as an unknown probability distribution D over input pairs (x, y))

***Unsupervised learning – Act on data without supervision (like humans)***

***Diagram

Description automatically generated***

UL algorithms learn from **untagged** data instead, exhibiting self-organization that captures patterns as **probability densities**

We have input data but no corresponding output data

**The goal is that through *mimicry,* the machine is forced to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**

Therefore, **there isn’t any training upon a given dataset**

The types of UL algorithms are:

* **Clustering**

Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities

* **Association**

An association rule is an unsupervised learning method which is used for finding the relationships between variables, determining the set of items that occurs together in the dataset

**The concept of overfitting and underfitting**

* **Underfitting – high bias**

A machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data, i.e., it only performs well on training data but performs poorly on testing data

* **Overfitting – high variance**

A machine learning algorithm is said to have overfitting when the model does not make accurate predictions on testing data (e.g., too much noisy training data, the model **learns** the noise as a concept, reducing generalization)

Diagram

Description automatically generated

**Inductive, Deductive and Transductive learning**

**Induction** is a concept commonly present and used in traditional Supervised Learning, the reasoning from observed training cases to general rules, which are then applied to test cases

**Induction is not truth-preserving** as new knowledge can compromise old knowledge

However, **transduction** is reasoning from observed specific training cases to specific test cases, meaning that transductive learning techniques have observed all the data beforehand (training and testing)

The model already encounters both the training and testing, while, under the inductive learning policy, **only** the training data is encountered when training the model

**Deduction** is applied to obtain generalizations from a solved example and its explanation, obtaining general knowledge from general knowledge

**Deduction is truth-preserving** as new knowledge cannot compromise old knowledge