Introduction to Machine Learning

Praanesh Balakrishnan Nair

January 28, 2025

Contents

1 Introduction

Quote by Herbert Alexander Simon:

> Learning is the process by which any system improves its performance from experience

2 Well-posed Learning Problem by Tom Mitchell (1998)

A computer program

- Performs Task T

- Has some Performance P
- Learns from Experience E

Sl. No	Task	Performance	Experience
1.	Classifying Emails as Spam/Not Spam	Number of emails correctly classified	Watching you Label Ema
2.	Playing Chess	Percent of games won	Watch enemy play
3.	Handwriting Recognition	Percent of correct recognitions	Sample images

3 Components of a dataset

3.1 Features

- Individual measurable properties, which are going to be used as input to the machine learning model. Eg. Age of people, Dimensions of a house, etc

3.2 Data Points

- Multiple samples of features.
- Eg:

Sl. No	Age	Height	Weight	BP
1	19	175	68	999
2	25	169	69	0
...	...	...	...	...

Each of these rows are data points. In each data point, you have different samples of the same features

3.3 Feature Vector

- Features in one data-point is often mathematically represented as a Vector
- Eg:

Sl. No	Age	Height	Weight	BP	Feature Vector
1	19	175	68	999	[19, 175, 68]
2	25	169	69	0	[25, 169, 0]
...	...	...	...	...	...

3.4 Distance and Similarity Matrix

3.4.1 Distance Matrix

- Basically Adjacency Matrix
- $d(i, j)$ = distance between i^{th} data point and j^{th} data point.
- It's Symmetric
- Diagonal Elements are 0
- The actual distance between i^{th} data point and j^{th} data point can be measured in many ways:

1. **Euclidean Distance** $= \sqrt{\sum (x_i - y_i)^2}$
2. **Manhattan distance** $= \sum |x_i - y_i|$
3. **Cosine Distance** $= 1 - \frac{x.y}{|x||y|}$

4 Classification

Here, you input some data, and the output is a classification of it.

4.1 Types of Classification Learning

4.1.1 Supervised Learning

- Classify features X_i into classes/labels Y_i
- You find the pattern of data that is associated with one label, and use that pattern to classify.

4.1.2 Unsupervised Learning

- You have only features X_i and no labels
- You find patterns, so that similar patterns form one label, and anything different will be given another label.

4.1.3 Semi-supervised Learning

The entire dataset consists of labelled and unlabelled data

1. Perform supervised learning on the labelled data
2. Now you use this to predict the labels of the unlabelled data. The predicted labels are called **psuedo-labels**.
3. Now do supervised learning on the combined data

4.2 Model Accuracy

4.2.1 Error Rate

- In classification, the model accuracy is quantified by the **error rate**.
- **Error Rate** $= \frac{\textbf{number of misclassifications}}{\textbf{total number of data points}}$
- **Error Rate** $= \frac{\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)}{n}$, where $I(y_i \neq \hat{y}_i)$ is 1 if it's a mismatch, and 0 if it's a match

4.2.2 Confusion Matrix

- Matrix where:
  - rows signify Ground truth (Row1: +, Row2: -)
  - columns signify predicted output (Column1: +, Column2: -)
- If row and column have same sign, it means the model has predicted correctly (it's a **true output**).
- If row and column have opposite signs, it means the model has predicted incorrectly (it's a **false output**).

	Predicted +	Predicted -
Actual +	**True Positive**	**False Negative**
Actual -	**False Positive**	**True Negative**

- Here are things you can derive from the confusion matrix:

	Predicted +	Predicted -	
Actual +	**True Positive**	**False Negative**	Sensitivity/Recall
Actual -	**False Positive**	**True Negative**	Specificity
	Precision	Negative Predictive Value	

  - **Sensitivity** $= \frac{\textbf{Diag. Element of Row 0}}{\textbf{Row 0}} = \frac{TP}{TP+FN}$
  - **Specificity** $= \frac{\textbf{Diag. Element of Row 1}}{\textbf{Row 1}} = \frac{TN}{FP+TN}$
  - **Precision** $= \frac{\textbf{Diag. Element of Column 0}}{\textbf{Column 0}} = \frac{TP}{TP+FP}$
  - **Negative Predictive Value** $= \frac{\textbf{Diag. Element of Column 1}}{\textbf{Column 1}} = \frac{TN}{FN+TN}$

4.2.3 Receiver Operating Characteristic Curve (ROC) and Area Under the Curve (AUC)

- ROC is the plot between True Positive and False Positive
- AUC is the area under ROC
- $0 \leq AUC \leq 1$
- $AUC = \int_0^1 (\textbf{ROC Curve})$

4.3 Model Validation Techniques

4.3.1 Internal Validation

- Seperation betwen clusters should be high
- Cohesion (distance between points in a cluster) should be low

4.3.2 External Validation

1. Dice Coefficient
   - $D(A, B) = \frac{2|A \cap B|}{|A|+|B|}$
   - If $D(A, B) = 0$, then there's no overlap. Similarly if $D(A, B) = 1$, they are the same set.
   - A could be the data we have and B could be some external data.
2. Jaccard Similarity Index
   - $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$

5 Regression

Here, you input some data and you get a quantitative response.

5.1 Model Accuracy

- In regression, it's called the **quality of fit** and it's the quantification of the degree of closeness of predicted response and the true response
- The most commonly used measure for this, is the **Mean Square Error (MSE)**.
- $MSE = \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{n}$, where y_i is the true value, $f(x_i)$ is the predicted value
- Accuracy of the Model $\propto \frac{1}{MSE}$