MLT Week-1

Sherry Thomas
21f3001449

Contents

Abstract

The week provides an introduction to Machine Learning and subsequently delves into the syllabus with a focus on unsupervised learning. The two primary areas of study covered are representation learning and Principal Component Analysis (PCA).

Introduction to Machine Learning

Machine Learning is a sub-field of artificial intelligence concerned with the design of algorithms and statistical models that allow computers to learn from and make predictions or decisions based on data, without being explicitly programmed. It utilizes mathematical optimization, algorithms, and computational models to analyze and understand patterns in data and make predictions about future outcomes.

It can be further explained as follows:

- Why: Machine Learning is used to automate tasks that would otherwise require human intelligence, to process vast amounts of data, and to make predictions or decisions with greater accuracy than traditional approaches. It also has surged in popularity in recent years.
- Where: Machine Learning is applied in various fields such as computer vision, natural language processing, finance, and healthcare, among others.Where: Machine Learning is applied in various fields such as computer vision, natural language processing, finance, and healthcare, among others.

- What: Machine Learning departs from traditional procedural approaches, instead it is driven by data analysis. Rather than memorizing specific examples, it seeks to generalize patterns in the data. Machine Learning is not based on magic, rather it relies on mathematical principles and algorithms.

Broad Paradigms of Machine Learning

1. **Supervised Learning**:Supervised Machine Learning is a type of machine learning where the algorithm is trained on a labeled dataset, meaning that the data includes both inputs and their corresponding outputs. The goal of supervised learning is to build a model that can accurately predict the output for new, unseen input data. Few examples:

- Linear regression for predicting a continuous output
- Logistic regression for binary classification problems
- Decision trees for non-linear classification and regression problems
- Support Vector Machines for binary and multi-class classification problems
- Neural Networks for complex non-linear problems in various domains such as computer vision, natural language processing, and speech recognition

2. **Unsupervised Learning**: Unsupervised Machine Learning is a type of machine learning where the algorithm is trained on an unlabeled dataset, meaning that only the inputs are provided and no corresponding outputs. The goal of unsupervised learning is to uncover patterns or relationships within the data without any prior knowledge or guidance. Few examples:

- Clustering algorithms such as K-means, hierarchical clustering, and density-based clustering, used to group similar data points together into clusters
- Dimensionality reduction techniques such as Principal Component Analysis (PCA), used to reduce the number of features in a dataset while preserving the maximum amount of information
- Anomaly detection algorithms used to identify unusual data points that deviate from the normal patterns in the data

3. **Sequential learning**: Sequential Machine Learning (also known as time-series prediction) is a type of machine learning that is focused on making predictions based on sequences of data. It involves training the model on a sequence of inputs, such that the predictions for each time step depend on the previous time steps. Few examples:

- Time series forecasting, used to predict future values based on past trends and patterns in data such as stock prices, weather patterns, and energy consumption
- Speech recognition, used to transcribe speech into text by recognizing patterns in audio signals
- Natural language processing, used to analyze and make predictions about sequences of text data

Representation Learning

Representation learning is a sub-field of machine learning that focuses on learning meaningful and compact representations of complex data for tasks such as dimensionality reduction, clustering, and classification.

Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$, we need to find a representation for it with the minimum reconstruction error.

Let $\mathbf{w}$, where $||\mathbf{w}|| = 1$, be the best linear representation of the dataset.

The representation is given by,

$$\frac{(\mathbf{x}_i^T \mathbf{w})}{\mathbf{w}^T \mathbf{w}} \mathbf{w}$$

$$\text{But } ||\mathbf{w}|| = 1$$

$$\therefore \text{ Projection } = (\mathbf{x}_i^T \mathbf{w})\mathbf{w}$$

The reconstruction error is given by,

$$\text{Reconstruction Error}(f(\mathbf{w})) = \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w})\mathbf{w}||^2$$

where $\mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w})\mathbf{w}$ is the residue and can be given by $\mathbf{x}'$.

Our objective is to minimize the reconstruction error. Minimizing it, we get,

$$\begin{aligned}
\min_{\mathbf{w} \in ||\mathbf{w}||=1} f(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^{n} -(\mathbf{x}_i^T \mathbf{w})^2 \\
\therefore \max_{\mathbf{w} \in ||\mathbf{w}||=1} f(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{w})^2 \\
&= \mathbf{w}^T (\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T) \mathbf{w} \\
\max_{\mathbf{w} \in ||\mathbf{w}||=1} f(\mathbf{w}) &= \mathbf{w}^T \mathbf{C} \mathbf{w}
\end{aligned}$$

where $\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T$ is the Covariance Matrix and $\mathbf{C} \in \mathbb{R}^{d \times d}$.

From the above equation we find that $\mathbf{w}$ is the eigenvector corresponding to the largest eigenvalue λ of $\mathbf{C}$.

$\mathbf{w}$ is also called the First Principal Component of the dataset.

Potential Algorithm

With this information, we can give the following algorithm:\ Given a dataset $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$,

- Center the dataset

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

$$\mathbf{x}_i = \mathbf{x}_i - \qquad \forall i$$

- Find the best representation $\mathbf{w} \in \mathbb{R}^d$ and $||\mathbf{w}|| = 1$.
- Replace $\mathbf{x}_i = \mathbf{x}_i - (\mathbf{x}_i^T \mathbf{w})\mathbf{w} \quad \forall i$
- Repeat the first three steps until residues become zero and we obtain $\mathbf{w}_2, \mathbf{w}_3, \dots, \mathbf{w}_d$.

But is this the best way? How many $\mathbf{w}$ do we need for optimal compression?

Principal Component Analysis

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset by identifying the most important features, or principal components, that explain the maximum amount of variance in the data. PCA accomplishes this by transforming the original dataset into a new set of variables that are uncorrelated and ordered by their importance in explaining the variance in the data. This can be helpful for visualizing high-dimensional data or for preprocessing data before performing machine learning tasks.

Following the potential algorithm from above, using the $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$, we get,

$$\forall i \quad \mathbf{x}_i - ((\mathbf{x}_i^T \mathbf{w}_1)\mathbf{w}_1 + (\mathbf{x}_i^T \mathbf{w}_2)\mathbf{w}_2 + \dots + (\mathbf{x}_i^T \mathbf{w}_d)\mathbf{w}_d) = 0$$

$$\therefore \mathbf{x}_i = (\mathbf{x}_i^T \mathbf{w}_1)\mathbf{w}_1 + (\mathbf{x}_i^T \mathbf{w}_2)\mathbf{w}_2 + \dots + (\mathbf{x}_i^T \mathbf{w}_d)\mathbf{w}_d$$

From the above equation, we can say that we can represent the data using $\{\mathbf{x}_i^T \mathbf{w}_1, \mathbf{x}_i^T \mathbf{w}_2, \dots, \mathbf{x}_i^T \mathbf{w}_d\}$, which are constants, and $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$ which are vectors.

Therefore, data which was initially $d \times n$ can now be represented with $d(d+n)$ elements which doesn't seem great at first.

But if the data lies in a lower sub-space, the residues can be zero without the need for d principal components.

If the data can be represented using only k principal components, where $k << d$, the data now can be compressed from $d \times n$ to $k(d+n)$.

Approximate Representation

If the data can be approximately represented by a lower sub-space, is it enough that we use only those k projections? How much variance should be covered?

Given a centered dataset $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$, let $\mathbf{C}$ be its covariance matrix, and let $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$ be its eigenvalues, which are non-negative because the covariance matrix is a positive semi-definite matrix, placed in descending order, and let $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$ be its corresponding eigenvectors of unit length.

The eigen equation for the covariance matrix can be given by,

$$\begin{aligned}
\mathbf{C}\mathbf{w} &= \lambda\mathbf{w} \\
\mathbf{w}^T\mathbf{C}\mathbf{w} &= \mathbf{w}^T\lambda\mathbf{w} \\
\therefore\lambda &= \mathbf{w}^T\mathbf{C}\mathbf{w} \qquad \{\mathbf{w}^T\mathbf{w} = 1\} \\
\lambda &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w})^2
\end{aligned}$$

Therefore, as the mean is zero, λ gives the variance captured by the eigenvector $\mathbf{w}$.

A good rule of thumb is that the variance captured by P.C.A. should be at least 95%. If the first k eigenvectors capture the required variance, this is given by,

$$\frac{\displaystyle\sum_{j=1}^{k}\lambda_j}{\displaystyle\sum_{i=1}^{d}\lambda_i} \geq 0.95$$

Hence, the higher the variance captured, the lower is the error obtained.

P.C.A. Algorithm

Given a centered dataset $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$, let $\mathbf{C}$ be its covariance matrix.\ The algorithm is as follows:

- Step 1: Find the eigenvalues and eigenvectors of $\mathbf{C}$. Let $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$ be its eigenvalues placed in the descending order, and let $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\}$ be its corresponding eigenvectors of unit length.
- Step 2: Calculate k, where k is the number of required top eigenvalues and eigenvectors, according to the required variance that needs to be covered.
- Step 3: Project the data onto the eigenvectors, and obtain the required representation as a linear combination of those projections.

In short, we can say that P.C.A. is a dimensionality reduction technique that finds combination of features that are de-correlated (independent of each other).

Credits

- Professor Arun Rajkumar: The content as well as the notations are from his slides and lecture.

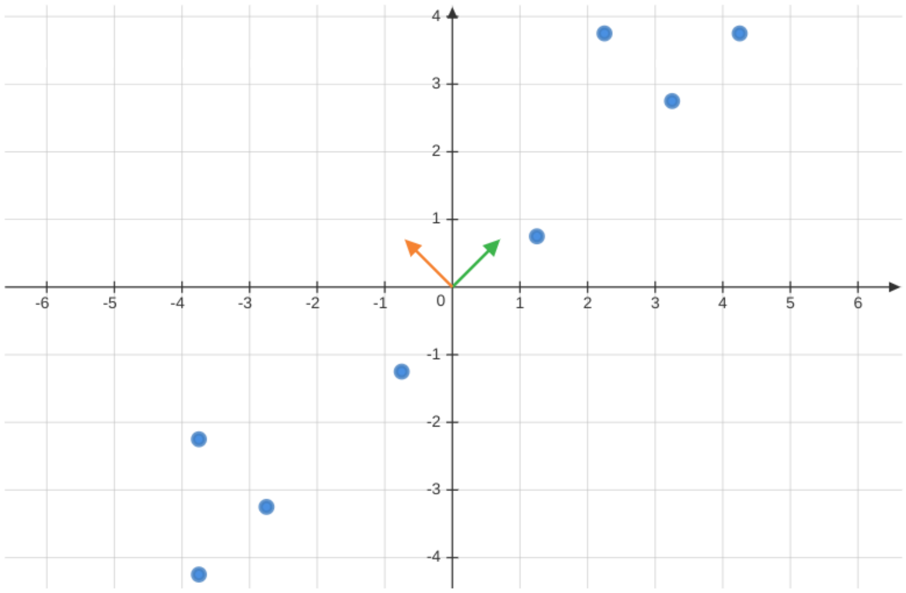

Figure 1: The dataset depicted in the diagram has two principal components: the green vector represents the first PC, whereas the red vector corresponds to the second PC.