Machine Learning for social sciences

Natural Language Processing : Session 1

William Aboucaya

Introduction

As you already know, machine learning can be used for many use cases :

Introduction

As you already know, machine learning can be used for many use cases :

- Value prediction

Introduction

As you already know, machine learning can be used for many use cases :

- Value prediction
- Multi-label classification

Introduction

As you already know, machine learning can be used for many use cases :

- Value prediction
- Multi-label classification
- Similarity measurement

Introduction

As you already know, machine learning can be used for many use cases :

- Value prediction
- Multi-label classification
- Similarity measurement
- and even more...

But what if we want to regroup data into groups without supervising the definition of these groups ?

Supervized vs. unsupervized learning

Supervised Learning

- Learns from labeled data to make predictions or classify data.
- **Key Characteristics :**
  - Data includes input-output pairs.
  - Goal : Minimize error by matching predictions to known outputs.

Supervized vs. unsupervized learning

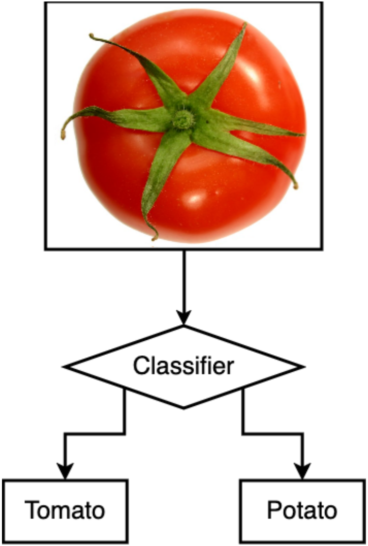
Supervised Learning

- Learns from labeled data to make predictions or classify data.
- **Key Characteristics :**
  - Data includes input-output pairs.
  - Goal : Minimize error by matching predictions to known outputs.
- **Examples :**
  - Image classification
  - Spam email detection
  - Stock price prediction

Supervized vs. unsupervized learning

Supervised Learning

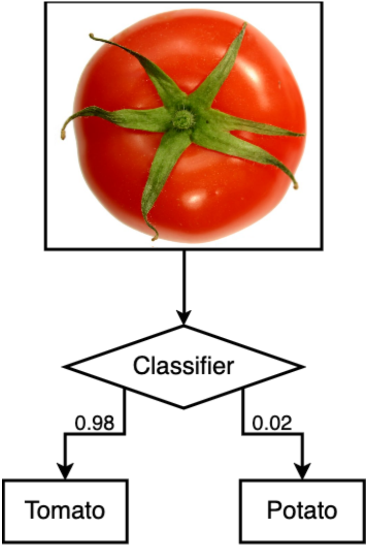
- Learns from labeled data to make predictions or classify data.
- **Key Characteristics :**
  - Data includes input-output pairs.
  - Goal : Minimize error by matching predictions to known outputs.
- **Examples :**
  - Image classification
  - Spam email detection
  - Stock price prediction



Supervized vs. unsupervized learning

Supervised Learning

- Learns from labeled data to make predictions or classify data.
- **Key Characteristics :**
  - Data includes input-output pairs.
  - Goal : Minimize error by matching predictions to known outputs.
- **Examples :**
  - Image classification
  - Spam email detection
  - Stock price prediction



Supervized vs. unsupervized learning

Unsupervised Learning

- Learns patterns and structures from unlabeled data.
- **Key Characteristics :**
  - Data has no explicit labels or outputs.
  - Goal : Discover hidden patterns or groupings.

Supervized vs. unsupervized learning

Unsupervised Learning

- Learns patterns and structures from unlabeled data.
- **Key Characteristics :**
  - Data has no explicit labels or outputs.
  - Goal : Discover hidden patterns or groupings.
- **Examples :**
  - Customer segmentation
  - Dimensionality reduction
  - Anomaly detection

Supervized vs. unsupervized learning

Unsupervised Learning : the example of semantic textual similarity

Objective : Quantifying the degree of similarity between two texts based on their meaning rather than their vocabulary.

Supervized vs. unsupervized learning

Unsupervised Learning : the example of semantic textual similarity

Objective : Quantifying the degree of similarity between two texts based on their meaning rather than their vocabulary.

Example :

- **Sentence 1 :** This NLP class is fantastic 😍
- **Sentence 2 :** I really like this natural language processing course
- **Similarity :** 0.703

. Results obtained using `sentence-transformers/all-mpnet-base-v2`

Supervized vs. unsupervized learning

Unsupervised Learning : the example of semantic textual similarity

Objective : Quantifying the degree of similarity between two texts based on their meaning rather than their vocabulary.

Example :

- **Sentence 1 :** This NLP class is fantastic 😍
- **Sentence 2 :** This class is extremely boring 😴
- **Similarity :** 0.491

. Results obtained using `sentence-transformers/all-mpnet-base-v2`

Supervized vs. unsupervized learning

Unsupervised Learning : the example of semantic textual similarity

Objective : Quantifying the degree of similarity between two texts based on their meaning rather than their vocabulary.

Example :

- **Sentence 1 :** This NLP class is fantastic 😍
- **Sentence 2 :** I would like a falafel sandwich with Algerian sauce and a drink please.
- **Similarity :** 0.081

. Results obtained using `sentence-transformers/all-mpnet-base-v2`

Clustering

Clustering is an unsupervised learning technique used to group similar data points. It identifies patterns or structures in the data without requiring labels.

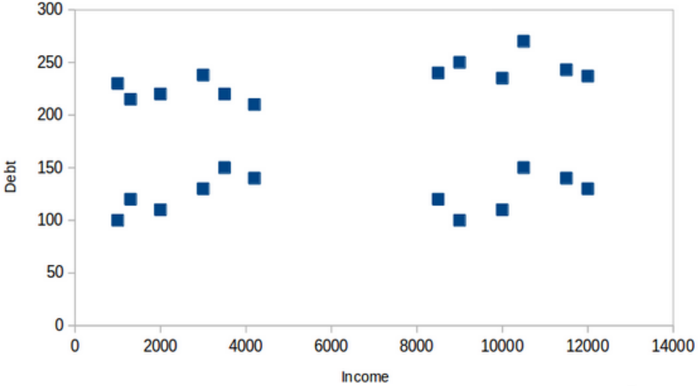
Clustering

Clustering is an unsupervised learning technique used to group similar data points. It identifies patterns or structures in the data without requiring labels. Popular algorithms include :

- K-Means Clustering
- Louvain Clustering
- DBSCAN (Density-Based Spatial Clustering)

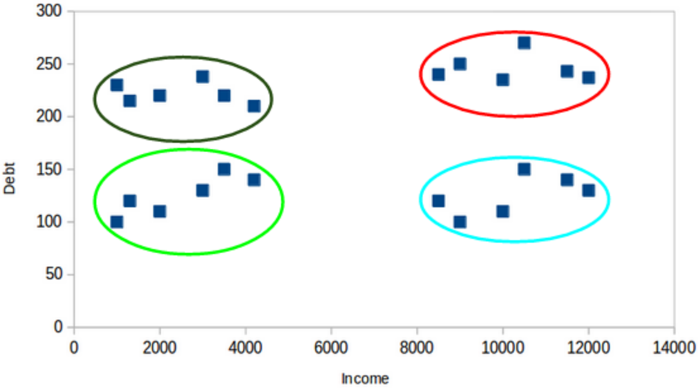
Clustering

Example : K-Means clustering of people based on their debt and annual income (with $K = 4$)



Clustering

Example : K-Means clustering of people based on their debt and annual income (with $K = 4$)



Problem

So you can cluster numeric data, graphs, etc.

Problem

So you can cluster numeric data, graphs, etc.

But we are in an NLP class, can we cluster texts ? Like, based on topics ? 🤨

Topic Modeling

Definition : Topic modeling is an unsupervised machine learning technique that discovers abstract topics within a collection of documents.

Topic Modeling

Definition : Topic modeling is an unsupervised machine learning technique that discovers abstract topics within a collection of documents.

Goal : Automatically identify clusters of words (topics) that frequently occur together in documents.

Topic Modeling

Definition : Topic modeling is an unsupervised machine learning technique that discovers abstract topics within a collection of documents.

Goal : Automatically identify clusters of words (topics) that frequently occur together in documents.

Applications :

- Document categorization
- Information retrieval
- Recommender systems
- Social media analysis

LDA

Latent Dirichlet Allocation is a simple topic modeling method. Based on the co-occurrences of words in a series of documents, the LDA algorithm extracts the topics and their associated representative words (or n-grams).

LDA

Latent Dirichlet Allocation is a simple topic modeling method. Based on the co-occurrences of words in a series of documents, the LDA algorithm extracts the topics and their associated representative words (or n-grams).

Assume a collection of documents about topics like sports, politics, and technology.

LDA

Latent Dirichlet Allocation is a simple topic modeling method. Based on the co-occurrences of words in a series of documents, the LDA algorithm extracts the topics and their associated representative words (or n-grams).

Assume a collection of documents about topics like sports, politics, and technology.

LDA uncovers :

- Topic 1 (Sports) : *game, team, player, score, match*.
- Topic 2 (Politics) : *election, policy, government, vote*.
- Topic 3 (Technology) : *software, AI, computer, NLP*.

LDA

Lab 1 : Exploration of an online participatory process using LDA

Let's dive into it 🤓

LDA

Limitations :

- Setting the number of topics to extract is difficult.

LDA

Limitations :

- Setting the number of topics to extract is difficult.
- Interpreting the extracted topics can be subjective.

Limitations :

- Setting the number of topics to extract is difficult.
- Interpreting the extracted topics can be subjective.
- Requires careful pre-processing of text data.

Limitations :

- Setting the number of topics to extract is difficult.
- Interpreting the extracted topics can be subjective.
- Requires careful pre-processing of text data.
- May struggle with very large or very small datasets.

LDA

Limitations :

- Setting the number of topics to extract is difficult.
- Interpreting the extracted topics can be subjective.
- Requires careful pre-processing of text data.
- May struggle with very large or very small datasets.

So how can we solve (some of) these issues ?

BERTopic

BERTopic is a topic modeling technique that leverages pre-trained language models and clustering algorithms.

BERTopic

BERTopic is a topic modeling technique that leverages pre-trained language models and clustering algorithms.

It identifies coherent topics from textual data, allowing for **dynamic topic reduction**.

BERTopic

BERTopic is a topic modeling technique that leverages pre-trained language models and clustering algorithms.

It identifies coherent topics from textual data, allowing for **dynamic topic reduction**.

Combines the strength of **transformer-based embeddings** (e.g., BERT) with techniques like UMAP for dimensionality reduction.

BERTopic

How does it work ?

- **Text Embedding** : Text is converted into embeddings (i.e., vectors) using models like BERT or SBERT.

BERTopic

How does it work ?

- **Text Embedding** : Text is converted into embeddings (i.e., vectors) using models like BERT or SBERT.
- **Dimensionality Reduction** : UMAP reduces embedding dimensions for efficient clustering.

BERTopic

How does it work ?

- **Text Embedding** : Text is converted into embeddings (i.e., vectors) using models like BERT or SBERT.
- **Dimensionality Reduction** : UMAP reduces embedding dimensions for efficient clustering.
- **Clustering** : HDBSCAN groups similar data points into clusters.

BERTopic

How does it work ?

- **Text Embedding** : Text is converted into embeddings (i.e., vectors) using models like BERT or SBERT.
- **Dimensionality Reduction** : UMAP reduces embedding dimensions for efficient clustering.
- **Clustering** : HDBSCAN groups similar data points into clusters.
- **Topic Representation** : Key terms are extracted from clusters to represent topics.

BERTopic

Main advantages :

- **Dynamic Topic Reduction** : Easily merge or split topics based on coherence and size.

BERTopic

Main advantages :

- **Dynamic Topic Reduction** : Easily merge or split topics based on coherence and size.
- **Light pre-processing** : No need to remove stop-words, transformer-based models use them as context.

BERTopic

Main advantages :

- **Dynamic Topic Reduction** : Easily merge or split topics based on coherence and size.
- **Light pre-processing** : No need to remove stop-words, transformer-based models use them as context.
- **Flexible Backends** : Compatible with a variety of embedding models (e.g., BERT, DistilBERT, GPT).

BERTopic

Main advantages :

- **Dynamic Topic Reduction** : Easily merge or split topics based on coherence and size.
- **Light pre-processing** : No need to remove stop-words, transformer-based models use them as context.
- **Flexible Backends** : Compatible with a variety of embedding models (e.g., BERT, DistilBERT, GPT).

Let's see what we can do with it 🤓