LING3401 Linguistics and Information Technology

Tutorial: Machine learning basics II

Yige Chen

The Chinese University of Hong Kong

February 12, 2025

Acknowledgment

- Some of the tutorial materials are based on: Dan Jurafsky and James H. Martin. *Speech and Language Processing* (3rd ed. draft). 2024.
- GPT-4o and DeepSeek-R1 helped me write more than half of today's codes. Thanks GPT and DeepSeek!

Machine learning basics

- Supervised Learning: learning a mapping from inputs (e.g., text) to outputs (e.g., labels) using labeled data
- Unsupervised Learning: finding patterns or structures in data without labeled outputs

Machine learning basics

- Pop quiz! Are these examples of supervised or unsupervised learning?
    - You are given a dataset of movie reviews. Each review is labeled as either “positive” or “negative”. Your task is to train a model to predict whether a new review is positive or negative.
    - You have a collection of news articles, but they are not labeled. Your task is to group the articles into different topics based on patterns in the text.

Training, validation, and testing

- Training, validation, and testing
  - Training set: used to train the model by adjusting weights based on the data.
  - Validation set: used to tune hyperparameters and prevent overfitting.
  - Test set: assesses the final performance of the trained model on unseen data.
- Overfitting and generalization
  - Overfitting: the model learns noise in the training data and performs poorly on new data.
  - Generalization: the model performs well on unseen data by capturing the underlying patterns.
- Evaluation metrics
  - Examples: accuracy, precision, recall, F1-score, perplexity, etc.
  - Metrics are chosen based on the task (e.g., classification vs. generation).

Training, validation, and testing

- Pop quiz! Why do we need a training set?
  1. Because the model learns best through telepathy
  2. To fine-tune hyperparameters and prevent overfitting
  3. To update the model's weights and learn patterns from data
  4. To evaluate the model's performance after training

Training, validation, and testing

- Pop quiz! Why do we need a validation set?
  1. To evaluate how well the model generalizes to unseen data
  2. To tune hyperparameters and assess model performance before final testing
  3. To train the model by updating its weights
  4. To give the model a motivational speech before the test

Training, validation, and testing

- Pop quiz! Why do we need a test set?
  1. To evaluate how well the model generalizes to unseen data
  2. To check if the model can finally become sentient and take over the world
  3. To adjust hyperparameters for better performance
  4. To help the model learn patterns from labeled data

Have a try!

- Take a look at Part 1 of today's Colab notebook, and see the examples of supervised learning!
    - It is a text classification task (sentiment analysis)
    - We are using four machine learning algorithms: support vector machines, logistic regression, naïve Bayes, and random forest
- Pay attention to
    - What ML algorithms do they use
    - The training/test split. Does it use a validation set? Does it employ cross-validation?
    - How the models are evaluated?
    - How are the results different wrt. ML algorithms?

Have a try!

- Take a look at Part 2 of today's Colab notebook, and see the examples of unsupervised learning!
    - This is a clustering task that is performed using word embeddings
    - We employ k-means to cluster (i.e., to group) the words into several groups based on their embeddings
    - We are using 100-dimensional GloVe embeddings
    - Since visualizing 100 dimensions is impractical, we will reduce the dimensionality to 2
    - Note: seeing 2 graphs does not mean clustering was done 2 times. It was performed using the 100-dimensional embeddings but was represented in a 2D space with 2 different dimensionality reduction methods

Word embeddings and dimensionality reduction

- We will come back to this next week during the lecture!
- Word embeddings: dense vector representations of words, capturing their meaning based on context
- So in fact as long as we visualize the clustering results, we are essentially visualizing 100-dimensional GloVe embeddings in a 2D space as well!
  - So reducing the 100D space to 2D is dimensionality reduction
- You will also see how the 100-dimensional GloVe embeddings can be represented in a 3D space next week during the tutorial

Miscellaneous

- Please do not hesitate to ask questions
- We enjoy feedback from you, so please let us know if you feel there's anything we could have done better
- It would be great if you'd bring your laptop to the class every week