UMC 301: Applied Data Science and Artificial Intelligence

Naman Mishra

August 2024

Contents

I Data science / AI **3**

II Machine learning **5**
II.1 Continuous & categorical data 5
II.2 Types of models 5
II.2.1 Predictive & generative models 6
II.3 Foundational model 6

The course

Lecture 1.
Friday
August 02

Chapter I

Data science / AI

How does Siri process the query "Show me a good breakfast restaurant near me"?

(i) Convert speech to text.

(ii) Understand the semantics of the query (for example, understand keywords like "breakfast", "restaurant") and formulate a structured query (place type = restaurant, meal type = breakfast, rating 3–5, distance 0–3 km).

- Also needs your current location!

(iii) Search for restaurants filtered by the structure above and rank based on some metric

- star rating,
- go through all reviews and understand the sentiments
- check your own past history, and try to give you a personalized recommendation
  - either recommend similar restaurants,
  - or recommend something completely different with a message like "would you like to try something new?"

or some combination of these (by another model).

An AI system for predicting cricket

The data comes from

- past performance of the player;
- weather, pitch conditions;

- other *factors*.

Data is usually stored in a tabular format representing

- each factor as a column,
- each data point as a row.

Another piece of data one could analyze is the net practice *videos*. An easier version of this is, read the *text* commentary from some source. Alternatively, record the *audio* commentary and analyze that. Could also look at the final pose of a batter once they have completed a shot as a single *image*.

Thus we can categorize data as

- tabular,
- timeseries (collection of tabular data),
- image (table of pixel values),
- video (time series of images),
- text (list of tokens),
- audio (time series of sound waves, or a spectrogram).

Chapter II

Machine learning

Definition II.1 (Training)**.** Given a family of functions $\mathcal{H} = \{h_\theta\}_{\theta\in\Theta}$ from X to Y and a loss function $L\colon \mathcal{H} \to \mathbb{R}$, *training* is the process of finding a $\theta \in \Theta$ that minimizes the loss $L(h_\theta)$.

Lecture 2. Monday August 05

II.1 Continuous & categorical data

The amount of rainfall tomorrow in millilitres is a continuous variable. Whether or not it will rain is considered a categorical variable.

Catergorical data can further be divided into ordinal and nominal data. If X takes values in $\{1, 2, 3\}$, it is an ordinal variable. If it takes values in $\{\text{red}, \text{green}, \text{blue}\}$, it is a nominal variable.

Predicting continuous data is called *regression*. Predicting categorical data is called *classification*.

II.2 Types of models

- Linear regression
- Gradient boosted tree model, best for tabular and time series data
  - Extreme gradient boosting (XGBoost)
  - Light gradient boosting (LightGBM)
  - Catetgorical gradient boosting (CatBoost)
- Neural networks, best for image, video, text and audio.
  - Convolutional neural networks (CNN) for image and video
  - Recurrent neural networks (RNN)
  - Transformers

II.2.1 Predictive & generative models

Predictive AI

- Input: Any data modality
- Output: Continuous or categorical data

Generative AI

- Input: Any data modality
- Output: Text, image, video, audio

II.3 Foundational model

A foundational model is one that has a broad knowledge and capability. We interact with it using prompts. Can be fine-tuned to specialize in a particular domain.

Examples: GPT, BERT, T5