Fall 2023

CS 482: Data Mining

Pantelis Monogioudis

t to PDF

Syllabus

Contents

Books

There are three axes that data mining intersects: data, methods and systems.

Data & Methods-oriented books:

- Main textbook on ML methods: *"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"*, 3rd Edition, by Aurélien Géron, 2022. O'Reilly This book will be referred to as `GERON` in the syllabus.
- Supplemental book on data methods that are not covered by Geron's book: "Mining of Massive Datasets" by Jure Leskovec, Anand Rajaraman, Jeff Ullman. Free download. This book will be referred to as `ULLMAN` in this syllabus

Systems-oriented books:

- Supplemental free book to provide additional input for the engineering / applied aspects of data mining and machine learning today: *ML Engineering*, by Andriy Burkov, 2020. Note that the free draft pdf files are at the end of the page.
- Optional book on data-driven application architectures: *"Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems"*, by Martin Kleppmann, Amazon

Skip to main content

Schedule

Lecture	Description	Reading List
1	We start with an *introduction to data mining* and through an end to end data mining application we understand and experience in our Colab notebooks all the stages of a data mining pipeline. At the same time we review elements of probability and statistics necessary for the remainder of the course and discuss your assignments and projects.	GERON Chapter 1
2	Almost all the tasks you will be called to perform as data scientists will have a flavor of *supervised learning*. Here we start with this learning problem and understand the two main tasks - regression and classification.	GERON Chapter2
3.1	We start treating *Decision Tree* as a first example for performing classification tasks. We start with a review of entropy and deep dive on how they work. We then expand into regression trees.	GERON Chapter 6
3.2	*Ensemble Methods* are a natural extension of decision & regression trees. They can handle massive datasets *and* offer partially *explainable* decisions. We cover *Random Forests* , *AdaBoost* and *Gradient Boosting* and pay particular interest to the later as it is routinely used to provide State Of The Art (SOTA) performance in structured data problems.	GERON Chapter 7
4	*Dimensionality reduction* and embeddings are key data representation tools that we need to know. Did you know that your face can be represented as vectors in some space? We review PCA here but we revisit the topic after we present neural networks	GERON Chapter 8
5.1	What happens if training data are not available or very expensive to acquire ? This week we review unsupervised learning and	GERON Chapter 9 and

Skip to main content

Lecture	Description	Reading List
	clustering approaches such as K-means. We review use cases in anomaly and novelty detection.	ULLMAN Chapter 7
5.2	*Similarity Search* naturally follows embedding generation where our latent-space vector representations are used to support search tasks. We will discuss *approximate nearest neighbors* techniques such as Locally Sensitive Hashing (LSH) and their implementation in Facebook AI Similarity Search (FAISS) or other similar systems.	ULLMAN Chapter 3
6	**Good luck in your Midterm**	
7	We are now ready to introduce the more formal version of supervised learning - where everything is treated probabilistically. We will return to linear regression and learn how to solve it via an algorithm that is the workhorse of modern machine learning: the *stochastic gradient descent*.	GERON Chapter 4
8	We continue with our probabilistic treatment of classification and learn how to interpret key performance metrics that arise in such problems. We meet our first neural-like classifier called logistic regressor.	GERON Chapter 3
9	We now expand into unstructured data and *neural networks* that are developed bottom up . We use the Tensorflow playground to understand the tradeoffs between classical and deep architectures and why deep neural networks dominate in unstructured big data applications today.	GERON Chapter 10
10	We then cover *Convolutional Neural Networks*, architectures specifically made to support visual data. They are instrumental to tasks involving feature extraction from images / video.	GERON Chapter 14
11	We revisit *Dimensionality Reduction* and *Embeddings* and present the *Autoencoder* architecture that is used to learn embeddings from data.	GERON Chapter 17

Skip to main content

Lecture	Description	Reading List
12	We now expand to methods that predict on data sequences. We present Recurrent Neural Networks (RNNs) that are heavily used to process sequantial data and perform tasks such as language modeling.	GERON Chapter 15
13	We conclude by presenting *Graph Neural Networks (GNNs)* that are heavily used to process graph data and perform tasks such as node classification.	Own notes
14	This is a review lecture before the final. It also acts as a buffer in case we run late.	
15	**Good luck in your final**	