Part 1: Review of Selected Topics

Part 1, session 6a of Data Mining Intro

2025-01-21 17:30 GMT

Abstract

Part 1 review: discussion of selected topics.

Table of Contents

Introduction

The aim of this session is to review sessions of the previous week, in order to prepare to present this material going forward. The primary reference is `eda_v3`:

- Notes on Exploratory Data Analysis (EDA)
  - MATH4350 version 3
  - by Karen Trageser

The subsequent material for the course is based on the following `ml` reference:

- Machine Learning: a Concise Introduction
  - by Steven W. Knox
  - next edition pending publication by Wiley
  - (Course notes: ML Topics & Techniques, version 2.93)

We first compare the `eda_v3` content to the selection of last week's topics. We'll then form small teams to select topics for discussion in the remainder of the session.

Overview of Topics

Here are the chapter titles of `eda_v3`.

EDA v3 chapters

chpt	title
1	Introduction
2	Exploratory Data Analysis (EDA)
3	Unsupervised Learning
4	Some Linear Algebra
5	Dimension Reduction
6	Topic Modeling
7	Sampling
8	Time Series
9	App-A: Probability Review
10	App-B: SVD Notes by Carla Martin

And here are the topics we discussed last week.

Topics discussed Jan 13-17

date	time	topic
2025-01-13	AM	Exploratory Data Analysis
2025-01-13	PM	Conditional Distributions
2025-01-14	AM	Clustering: EDA in Higher Dimensions
2025-01-14	PM	Text Analysis
2025-01-15	AM	Sampling and Study Design
2025-01-15	PM	Linear Algebra
2025-01-16	AM	Dimension Reduction
2025-01-17	AM	Time Series
2025-01-17	PM	Time Series & Point Processes: Frequency Analysis

Class Exercise: Select Topics for Further Discussion

Which topics presented last week merit further discussion? Which topics from `eda_v3` would you like to know more about? Form a team of 2-4 classmates, and take 10 minutes to record the 1-2 topics of greatest interest to you. We'll then use the remainder of the session to discuss them and follow up with write-ups.

Topics Nominated for Further Discussion

Here are the topics nominated for further discussion by one or more teams.

Topics Nominated by Teams

topic	n_teams
Time Series Forecasting	2
LDA: Latent Dirichlet Allocation	2
SVD:Singular Value Decomposition	1
KDE: Kernel Density Estimation	1
Clustering	1
PCA v MDS v SNE: when to use?	1