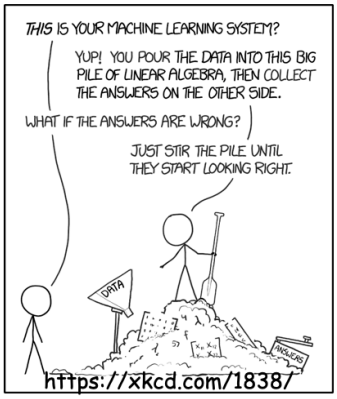
Important links

Contents



DSCI 563 Unsupervised Learning

- Course Jupyter book
- Course GitHub page
- Slack Channel
- Canvas
- Gradescope
- iClicker Cloud (Section 001)
- iClicker Cloud (Section 002)

- YouTube videos
- Class + office hours calendar

This course is about identifying underlying structure in data. We will talk about clustering, data representation (e.g., dimensionality reduction and word embeddings), and recommendation systems.

▶ Click to expand!
▶ Click to expand!
▶ Click to expand!

Format

▶ Click to expand!

Tentative Lecture Schedule

This course occurs during **Block 5** in the 2021/22 school year.

Lecture	Topic	Assigned videos	Resources and optional readings
0	Course Information		
1	K-Means and intro to GMMs	 • Videos: 14.1, 14.2,14.3	• `sklearn` clustering documentation • "Spaghetti Sauce" talk by Malcom Gladwell • Visualizing-k-means-clustering • Visualizing K-Means algorithm with D3.js • Clustering with Scikit with GIFs
2	DBSCAN and Hierarchical Clustering	 • Videos: 15.1, 15.2, 15.3	• Comparison of sklearn clustering algorithms • DBSCAN Visualization • Clustering with Scikit with GIFs
3	Dimensionality Reduction Intro	 • Videos: 17.1, 17.2, 17.3	• PCA visualization • Introduction to Machine Learning with Python book Chapter 3 • Mike's PCA video from CPSC 340 • StatQuest PCA video
4	More PCA, LSA, NMF, Autoencoders	No videos	
5	Word Vectors, Word Embeddings	 • Videos: 18.1, 18.2, 18.3	Word2Vec papers: • Distributed representations of words and phrases and their compositionality • Efficient estimation of word representations in vector space • word2vec Explained • Debiasing Word Embeddings

Lecture	Topic	Assigned videos	Resources and optional readings
6	Topic modeling	No videos	• Dave Blei video lecture, paper
7	Recommender Systems I	No videos	• Collaborative filtering for recommendation systems in Python, by N. Hug • How Netflix's Recommendations System Works
8	Recommender Systems II	No videos	• SVDfeature

Datasets

Here is the list of Kaggle datasets we'll use in the lectures.

- A small subset of 200 Bird Species with 11,788 Images (available here)
- A tiny subset of Food-101 (available here)
- Credit Card Dataset for Clustering
- Countries of the World
- Airline Sentiment
- Jester 1.7M jokes ratings dataset
- Amazon ratings data

If you want to be extra prepared, you may want to download these datasets in advance and save them under the `lectures/data` directory in your local copy of the repository. During labs, you will be given time to work on your own or in groups. There will be a lot of opportunity for discussion and getting help during lab sessions.

We are providing you with a `conda` environment file which is available here. You can download this file and create a conda environment for the course and activate it as follows.

```
conda env create -f env-dsci-563.yml
conda activate 563
```

We've only attempted to install this environment file on a few machines, and you may encounter issues with certain packages from the `yml` file when executing the commands above. This is not uncommon and may suggest that the specified package version is not yet available for your operating system via `conda`. When this occurs, you have a couple of options:

1. Modify the local version of the `yml` file to remove the line containing that package.
2. Create the environment without that package.
3. Activate the environment and install the package manually either with `conda install` or `pip install` in the environment.

Note that this is not a complete list of the packages we'll be using in the course and there might be a few packages you will be installing using `conda install` later in the course. But this is a good enough list to get you started.

▶ Click to expand!

▶ Click to expand!

Please see the general MDS policies.