CS7641: Machine Learning
Unsupervised Learning Report
Rubric
Fall 2025

Due Dates

- **11/04/2025** Grading Instructions Email sent to Team.
- **11/05/2025** First Two Calibration Reports on your sheet in Teams graded by EOD.
- **11/07/2025** Report on Calibration from Head TAs by EOD.
- **11/11/2025** Roughly 33-50% of Reports Graded.
- **11/14/2025** Roughly 50-80% of Reports Graded.
- **11/18/2025** Full 100% of Reports Graded by EOD.
- **11/19/2025** Stats run. Any outstanding or missed reports. Grades Posted to students.

Clustering

First, we should understand how to validate clusters. Unlike supervised learning, we don't have examples with labels. Therefore, there are many ways to approach this which haven't been described in lecture, so any way is acceptable as long as it's justified. On the other hand, if a student uses a dataset that has labels, then cluster validation is done by finding out the number of mislabeled samples. If they don't have labels, then, apart from the techniques in the overview above, a popular way to evaluate is to find out inter-cluster and intra-cluster distances. Inter is the distance between clusters and intra is the distance within clusters. You want clusters to be as densely packed and as far away from each other as possible. You can express this quantity as a ratio: intra / inter. The ratio has to be near 0 for a good model. Basically, you choose K when this ratio is lowest.

Overview of Clustering Algorithms

- Expectation Maximization (EM): EM is primarily a statistical technique used for finding the maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. It's widely used in scenarios like Gaussian Mixture Models (GMM) for clustering.
- K-Means: Grouping data into 'k' distinct, non-overlapping clusters based on their distances to the center of these clusters.
- Density-Based Clustering: As the name suggests, this method groups together points that are closely packed in a region of a feature space, based on density. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is one of the most popular algorithms in this category.
- Hierarchical Clustering: To build a tree of clusters, which can be visualized as a dendrogram, where each level of the tree represents a partition of the data into clusters at that distance level.

Algorithmic Principle

- EM: The EM algorithm alternates between the following two steps:
  - E-Step: Estimate the expected value of the latent variables given observed data and current estimate of parameters.
  - M-Step: Maximize the expected log-likelihood found on the E step with respect to the parameters.
- K-Means: Iteratively assign points to the nearest cluster centroid and then recalculate the centroids based on the points assigned to them.
- Density-Based Clustering: The fundamental idea here is that a cluster is a dense region of points, which is separated by regions of lower point density. For example, in DBSCAN, if a point has a certain number of neighbors within a specified radius, it becomes a core point, and all points reachable from a core point belong to the same cluster.
- Hierarchical Clustering: Can be either agglomerative (bottom-up) or divisive (top-down). Agglomerative starts with each data point as a separate cluster and then merges them based on distance, while divisive starts with one big cluster and splits it.

Output

- EM: It often provides soft assignments, meaning each data point is assigned a probability of belonging to each cluster, rather than a definitive cluster assignment.
- K-Means: Provides hard assignments where each point is definitively assigned to one cluster.
- Density-Based Clustering: Provides hard assignments where each point either belongs to a cluster or is considered an outlier.
- Hierarchical Clustering: Rather than a simple cluster assignment, it provides a dendrogram, a tree-like diagram showing the arrangement of the clusters.

Dealing with Noise

- EM: EM, especially when used with GMMs, doesn't inherently handle noise well. All points will be assigned to some cluster, even if they are anomalous.
- K-Means: Noise and outliers can affect the positioning of the centroids and, therefore, the quality of the clusters.
- Density-Based Clustering: Algorithms like DBSCAN are explicitly designed to handle noise. Those points that do not fall into the dense region (or are not close to it) are treated as noise or outliers.
- Hierarchical Clustering: Outliers can influence the structure of the dendrogram, especially in the early stages of agglomerative clustering.

Shape of Clusters

- EM: When using GMMs, clusters are often elliptical or spherical, given the Gaussian assumptions.
- K-Means: Assumes spherical clusters due to its reliance on centroid-based distance. Struggles with clusters of arbitrary shapes.
- Density-Based Clustering: Can discover clusters of arbitrary shapes, since it's based on density rather than distance from a centroid.
- Hierarchical Clustering: Does not assume a specific shape for clusters, but the shape can be influenced by the choice of distance metric and linkage criteria.

Parameter Sensitivity

- EM: Typically requires the number of clusters to be specified in advance (e.g., the number of Gaussian components in GMMs).
- K-Means: Requires the number of clusters 'k' to be specified in advance. Initialization of the centroids (starting position) can also influence the outcome.
- Density-Based Clustering: Doesn't need the number of clusters as a parameter, but parameters like density threshold and distance can be crucial for the results and might need fine-tuning.
- Hierarchical Clustering: Requires choice of linkage method (like single, complete, average) and a distance metric. The level at which to "cut" the dendrogram to define clusters also needs to be chosen.

Scalability

- EM: For large datasets, EM can be computationally intensive, especially for models with many parameters.
- K-Means: Relatively scalable and can be used with large datasets, especially with variations like MiniBatch K-means.
- Density-Based Clustering: Algorithms like DBSCAN can be more efficient than EM on large datasets, but their performance can degrade with increasing dimensionality due to the curse of dimensionality.
- Hierarchical Clustering: Not as scalable as k-means or DBSCAN for large datasets because of its $O(n^2)$ or worse complexity for many algorithms.

Non-Linear Manifold Learning Algorithms

Sammon Mapping

- Purpose: Reduce dimensionality while preserving the pairwise distances between data points.
- Principle: It uses a non-linear projection method where the cost function (Sammon's stress) emphasizes the preservation of small pairwise distances. The aim is to minimize the difference between the original distances and the distances in the reduced-dimensional space, especially for close points.
- Applications: Mostly used for visualization purposes.

Isomap (Isometric Mapping)

- Purpose: Preserve the geodesic (or intrinsic) distances between data points.
- Principle: It's a combination of multi-dimensional scaling (MDS) and a neighborhood graph. First, a neighborhood graph is constructed, and the shortest path between points (approximating geodesic distances) is computed using the graph. Then, MDS is applied to these geodesic distances.
- Applications: Used in several applications where the intrinsic manifold structure is more important than the Euclidean structure, like in certain medical images.

Laplacian Eigenmaps

- Purpose: Map nearby inputs into nearby outputs, designed specifically for clustering and visualization.
- Principle: Constructs a neighborhood graph and then finds a low-dimensional representation that preserves the local properties of this graph. The technique uses the spectrum (eigenvalues) of the Laplacian of this graph to find the embedding.
- Applications: Used in image segmentation, clustering, and visualization of manifold structures.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Purpose: Preserve the pairwise similarities between data points for visualization (typically in 2D or 3D).
- Principle: It converts high-dimensional Euclidean distances between data points into conditional probabilities representing similarities. Then, it minimizes the divergence between these probabilities and a similar set in the low-dimensional space, using a t-distribution to compute similarities in the reduced space.
- Applications: Predominantly used for visualization, especially in high-dimensional data visualization like gene expression data or deep learning activations.

UMAP (Uniform Manifold Approximation and Projection)

- Purpose: Dimensionality reduction aimed at preserving both local and global structures in the data.
- Principle: UMAP starts by approximating the data manifold in the high-dimensional space using a fuzzy topological representation. This means, instead of creating a strict k-nearest neighbors graph, UMAP allows for a more nuanced connection between data points. The local structure is captured similarly to t-SNE, by comparing distances in the original space to a distribution (in UMAP's case, usually the hyperbolic space, which is a space of constant negative curvature). For the low-dimensional representation, UMAP also employs a distribution comparison, but here it uses a different, flatter distribution. The aim is to find a low-dimensional representation where the two distributions are as close as possible.
- Advantages over other methods:
  - Speed: UMAP tends to be faster than t-SNE, especially for larger datasets.
  - Consistency: While t-SNE can sometimes produce different visualizations with different runs or perplexity settings, UMAP is more consistent.
  - Global Structure: UMAP retains more of the global structure than t-SNE, which is predominantly focused on preserving local clusters.
  - Versatility: While primarily used for visualization, UMAP can be used as a general non-linear dimensionality reduction method for other tasks as well.
- Applications: Data visualization, especially for high-dimensional datasets like gene expression data. Preprocessing step for machine learning tasks to reduce dimensionality. Exploratory data analysis to understand inherent clustering or structure in the data.

Intuition

When thinking about experiments like those in Step 3, it helps to think in terms of information. The information is already present in your data, if you have more data, you have more information, which means you can potentially find a better model. In Step 3, you perform dimensionality reduction (DR) and then clustering. In each case, you aren't adding new information to your data; instead, you're transforming or reorganizing what's already there. So in theory, these steps alone shouldn't magically create a better model. However, DR and clustering can sometimes act as filters, processes that remove noise or unwanted variation from the data. If your data is noisy, these transformations can help highlight the underlying patterns more clearly and lead to improved models. Make sure you understand how DR or clustering can serve as useful filters!

Datasets for Fall 2025

We require the students to use these two specific datasets. If they are not explicitly using these datasets, they will receive a zero. This should hopefully not be the case but please make a note on the sheet in the comments. For further information on these datasets, please see the SL Report description or the Kaggle repos.

- **Hotel Booking Demand**: Kaggle Repository: *Hotel Booking Demand*
- **US Accidents**: Kaggle Repository: *US Accidents*

Rubric

Total Points: 100 + 5 Extra Credit

As a reminder, please give all or nothing for points. If a section has 3 points for a section, you need to either give 3 or 0 points. This will help us with ambiguity between graders and nuance.

Structured Feedback – CS7641 Fall 2025 Report

Continuing this semester, I want to put the point values by the section feedback. I have an example document I would like everyone to use as a template. You will still need to add individualized feedback for each student. Please reach out if you have any questions!

Requirements (10 points)

- **7 points total:** This is the DOCSTRING-GTusername file on Canvas. All requirements met, full reproducibility ensured.
  - **2 points:** Overleaf and LaTeX used for report. Please double check the report exists.
  - **2 points:** Github Hash. Please double check that this hash works.
  - **2 points:** Instructions to run code. The can be anything but needs to be instructions to help with packages or init to run the code. If the instructions are on Github and not the DOCSTRING, please remove these 2 points.
  - **1 point:** The document is a PDF.
- **1 point:** Report is 8 pages or less. If there is anything more, please make a note to the student for future reports.
- **1 points:** Graphs and text are legible. This means you do not need to do any zooming while on Canvas at 100%.
  - This point can be hard. Many times the student will have some cropped graphs that are hard to read at 100% zoom on a computer. If there are 3 or more graphs that are hard to read the axis or key/legend, please deduct this point.

- **1 points:** Proper References. This is a check to make sure their reference are consistent in formatting. APA, MLA, or IEEE are fine, but need to stay consistent. IEEE does have a website url in the description. This point should be taken off if the student includes something outside the format chosen or text that should be included in the main body.
  - Additionally, if the student does not have a citation outside of the course material or citing the environment, you should take off a point. The rationale here is that the student should be looking for connections to the literature to support their results.

Hypothesis (8 pts)

- **4 points:** Initial Hypothesis
  1. (+2 pts) For stating an explicit hypothesis for what the student expects in the report. This can be a wide range of potential hypotheses, e.g. the student might comment on how the labels in their data will perform better than clustering labels for classification with the NN. The report does not need to have a hypothesis for each section, rather these sections should help develop an overall narrative to keep the paper focused.
  2. (+2 pts) Explicitly stating evidence to support the claim of the hypothesis from a lecture, paper, or generally known theory. The hypothesis must be grounded in evidence before exploration.
- **4 points:** Follow Through with Hypothesis
  1. (+2 pts) Following up with the hypothesis in their discussion or conclusion. It is not enough to just state a hypothesis but follow through from a narrative perspective needs to be met at the end of the report.
  2. (+2 pts) Providing direct evidence from their report to support if the hypothesis has been met or not. This can be short, however there needs to be something explicit to help better summarize.

Step 1: Clustering for Two Datasets (20 pts)

- **10 points:** For each of EM and K-Means Clustering:
  - (+4 pts) How did you choose hyperparameters for each dataset (i.e. the number of clusters)? Two points per dataset.
  - (+2 pts) Visualization or demonstration of the results.
  - (+4 pts) What are the results? Are they meaningful? Descriptive shape of the clusters. This is where a limitation would be addressed. If they student does not mention any connotations, they should be not awarded these three points. There should be different descriptions and rationale for each dataset. If there are the same number of clusters for both datasets, only award 2/4 points.

Step 2: Dimensionality Reduction for Two Datasets (24 pts)

This might be the most involved section for reach report. You might want to double check each of the algorithms.

- **8 points:** For each of RP, PCA, and ICA:
  - (+4 pts) How did they choose the number of components? Supporting results required: Eigenvalues or explained variance for PCA, kurtosis for ICA, reconstruction error for RP.
  - (+2 pts) Visualization or demonstration of the results. This could be in a pair-plot, mapping variance, a table of descriptor values. There needs to be evidence from their experiemntation.
  - (+2 pts) What are the results? Are they meaningful? Transformed features are linear combinations of the original features. Students should analyze and interpret them. Examples include discussing loading scores, assessing the structure of the data in the new feature space, projecting data onto components to visualize patterns, etc. This is where a limitation would be addressed. If they student does not mention any connotations, they should be not awarded these two points.

Step 3: Dimensionality Reduction, then Clustering for Two Datasets (14 pts)

- There are 12 possible combinations of experiments/results. The goal of this section is for the student to demonstrate results from their experiments as well as reason on the outcome. This should be a combination of graphs and, hopefully, tables to better allocate space.
- **7 points:** For each dataset:
  - (+4 pts) Visual demonstration of performance of combination of methods. They should reselect the optimal number of clusters after DR to receive full credit. A table works best here. If there are 12 distinct plots, don't take off points but suggest the student combine graphs to save space.
  - (+3 pts) Comparative analysis between methods. Did PCA or ICA assumptions align better with the data? Did dimensionality reduction help filter noise or improve clustering results? Why or why not?

Step 4: Dimensionality Reduction on Neural Network for One Dataset (12 pts)

- **12 points (4 each):** For each linear method (RP, PCA, ICA) consider the following.
  - (+2 pts) Was there a change in performance? You should be looking for the supporting results to help with their claims.
  - (+2 pts) Was there a change in wall clock time / speed? If the student uses a different metric to judge difference, that will suffice for these points. There needs to be an alternative way to judge, measure, evaluate the results.

Step 5: Clustering on Neural Network for One Dataset (12 pts)

- **12 points (6 each):** For each cluster method (EM and K-Means):
  - (+3 pts) Was there a change in performance? Supporting results should be details for analysis. This is where the student might mention that there is a filtering effect occuring.
  - (+3 pts) Was there a change in wall clock time / speed? If the student uses a different metric to judge difference, that will suffice for these points.

Extra Credit: Non-linear Manifold Learning Algorithm (Up to 5 points)

- **5 points:** Non-linear Manifold Learning Algorithm explained for use and adds significant insights into the datasets and theoretical implications. This will include both a visualization and commentary in their analysis that might highlight clusters or grouping from the DR/clustering results.
- **3 points:** Non-linear Manifold Learning Algorithm explained for use with a visualization but only superficial insights. This might include a comment about visual inspection instead of investigating a class imbalance or how they compare to results found in main report. There must be a visualization to earn points.
- **0 points:** If the student mentions there are Non-Linear Methods and the methods might perform better than current experiments in theory. This is potentially good for their discussion but this is not enough to earn EC. An explicit visualization is needed to earn any EC.

Deductions

Formal writing requirement. Bullet lists (`itemize`, `enumerate`, or `description`) signal draft-style notes rather than polished prose. If *any* section contains a list with more than four or more items, apply a **50 point deduction to the overall report score** (single global penalty, not per occurrence).

Version Control

- 11/03/2025 - TJL changed up the rubric to match Fall 2025 assignment. The different steps were modified to be less ganular and more straight-forward.
- 07/01/2025 - TJL changed up the rubric to match Summer 2025 assignment. Removed Step 5 and redistributed points.
- 03/24/2025 - TJL added more to the rubric to help with consistency of feedback.
- 03/20/2025 - JM refactored rubric to modification on implementation for Spring 2025.
- 03/03/2025 - TJL recalibrated the rubric to modification on implementation for Spring 2025.
- Before Spring 2025 - This rubric and assignment description have been adapted from the original versions written by Pushkar Kolhe and Charles Isbell. For detailed document history, refer to the "History" tab in Overleaf.