Week 10 Lecture Key Points

Important Points for the Final Exam

I highly recommend watching the Week 12 lecture recording for the key topics that will be important for the final exam. This document will summarize the key points covered in the Week 10 lecture.

What is Cluster Analysis?

Refer to slide 5 for details.

Can we manually cluster data points?

Yes, for 2D data, you can use a scatter plot. For higher-dimensional data, a scatter plot matrix or a tour can be used. However, manually clustering high-dimensional data is time-consuming, which is why we rely on clustering algorithms.

What is a Distance Measure?

A distance measure tells you how dissimilar data points are. It indicates whether two data points are "similar" or "close."

What distance measures are available?

The common ones include:

- Euclidean distance
- Mahalanobis distance (statistical)
- Manhattan distance

Other distance measures not mentioned in the lecture are also available.

Can any function be a distance metric?

No. It must follow the rules outlined on slide 14.

Can we apply Euclidean distance to any variables?

No. It should only be used for numeric variables. For variables like dates or categorical data, other distance metrics or clustering algorithms are more appropriate, although they were not covered in the lecture.

What is K-Means trying to achieve?

K-Means aims to find a grouping that minimizes the variance within clusters.

What are key characteristics of K-Means?

You need to specify the number of clusters in advance. K-Means typically works well with spherical clusters.

Why are there different implementations of K-Means?

K-Means is an NP-hard problem, meaning it cannot be solved perfectly. As a result, different approximation algorithms are used.

Briefly, how does K-Means work?

K-Means initially divides the data into k random clusters. Then it reassigns data points to the nearest cluster mean, updates the means, and repeats this process until no further reassignments are needed.

How many clusters should I choose for K-Means?

You can try different values of k and assess which provides the best results. There is no correct number of clusters, since we do not know the natural grouping of the data. It all depends on the context.

Why is it important to set a seed for K-Means?

Since K-Means begins with a random division of data into clusters, setting a seed allows you to control this randomness.

Why did you scale the dataset in the lecture?

When a dataset contains variables with different scales, scaling ensures that each variable contributes equally to the distance measure.

Why do we use a `table`?

We use a table to compare groupings, either between the true grouping variable and the clustering method, or between two clustering methods.

What is label switching?

Cluster labels themselves don't carry meaning. You can rename them, which is particularly useful when creating a two-way table to ensure large values appear on the diagonal.

Why do we check group means?

Checking group means helps us understand whether the cluster centers are far apart. This is an important criterion for determining cluster quality.

Why do we check clusters using a tour?

It's important to visualize clusters in the data space to ensure they are meaningful. Refer to slide 30 for more details.

Why do we need cluster statistics?

Although there is no definitive "correct" answer in clustering, we need to determine the final number of clusters. Cluster statistics help guide this decision.

Why didn't we cover the second type of hierarchical clustering?

There are two types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative is faster, and divisive is slower. We only covered agglomerative because it's the default method in most statistical software.

In brief, how does hierarchical clustering work?

First, calculate the distance matrix and choose a linkage method. Then, the two closest data points are merged. This process is repeated until all points are grouped into one cluster.

What's the difference between a distance measure and a linkage method?

A linkage method is a specialized distance measure. While a standard distance measure calculates dissimilarity between two points, a linkage method measures dissimilarity between two groups of points.

Why can we obtain multiple clustering solutions from a dendrogram?

Different cluster solutions are possible depending on where the dendrogram is cut, yielding clusters of varying sizes.

How should we appropriately cut the dendrogram?

To cut the tree appropriately, consider the tolerance of each solution. Tolerance is the vertical height between successive horizontal bars. For example, the height difference between the top two horizontal bars represents the tolerance for a 1-cluster solution, while the difference between the second and third bars gives the tolerance for a 2-cluster solution. The greater the tolerance, the more stable the solution. If two solutions have similar tolerances, both are valid. A solution with very small tolerance is likely not a reasonable choice.

What are the general steps for hierarchical clustering?

Refer to the last slide for a summary of the steps.

What’s the relationship between EDA and clustering?

While some forms of cluster analysis may not fall under EDA, clustering is typically considered a part of exploratory data analysis. It is used to uncover patterns or groupings in the data without making inferences about the population, a characteristic more aligned with inferential analysis.

When should we use K-Means versus hierarchical clustering?

Use K-Means if you know the number of groups in your data. Hierarchical clustering is better suited if you suspect there’s a hierarchical structure in the data. K-Means tends to be faster since you can limit the number of iterations, but it depends on the random seed, so the solution isn’t unique. In general, neither K-Means nor hierarchical clustering explains the meaning of the clusters—you’ll need to interpret the results to find the solution that makes the most sense and has practical relevance.