Week 7 Video 1

Clustering
Clustering

- A type of *Structure Discovery* algorithm
Clustering

- You have a large number of data points
- You want to find what structure there is among the data points
- You don’t know anything a priori about the structure
- Clustering tries to find data points that “group together”
Trivial Example

- Let’s say your data has two variables
  - Probability the student knows the skill from BKT (Pknow)
  - Unitized Time

- Note: clustering works for (and is effective in) large feature spaces
k-Means Clustering Algorithm
Not the only clustering algorithm

- Just the simplest
- We’ll discuss fancier ones as the week goes on
How did we get these clusters?

- First we decided how many clusters we wanted, 5
  - How did we do that? More on this in the next lecture

- We picked starting values for the “centroids” of the clusters...
  - Usually chosen randomly
  - Sometimes there are good reasons to start with specific initial values...
Then...

- We classify every point as to which centroid it’s closest to
  - This defines the clusters
  - Typically visualized as a voronoi diagram
Then...

- We re-fit the centroids as the center of the points in each cluster
Then...

- Repeat the process until the centroids stop moving
- “Convergence”
Note that there are some outliers.
What if we start with these points?
Not very good clusters
What happens if your starting points are in strange places?

Not trivial to avoid, considering the full span of possible data distributions
One Solution

- Run several times, involving different starting points
Exercises

- Take the following examples
- (The slides will be available in course materials so you can work through them)
- And execute k-means for them
- Do this by hand...
- Focus on getting the concept rather than the exact right answer...
- (Solutions are by hand rather than actually using code, and are not guaranteed to be perfect)
Exercise 7-1-1
Pause Here with In-Video Quiz

- Do this yourself if you want to

- Only quiz option: go ahead
Solution Step 1
Solution Step 2
Solution Step 4
No points switched -- convergence
K-Means did pretty reasonable here
Exercise 7-1-2
Pause Here with In-Video Quiz

- Do this yourself if you want to

- Only quiz option: go ahead
Solution Step 1
Solution Step 2
Solution Step 3
The three clusters in the same data lump might move around for a little while

But really, what we have here is one cluster and two outliers...

k should be 3 rather than 5

See next lecture to learn more
Exercise 7-1-3
Pause Here with In-Video Quiz

- Do this yourself if you want to

- Only quiz option: go ahead
The bottom-right cluster is actually empty!
There was never a point where that centroid was actually closest to any point
Pause Here with In-Video Quiz

- Do this yourself if you want to

- Only quiz option: go ahead
Solution Step 1

The diagram shows a scatter plot with two clusters of data points. The x-axis represents pknow, and the y-axis represents time.

Two lines are drawn in the plot. The line with a negative slope passes through the data points in the lower left and upper right quadrants, while the line with a positive slope passes through the data points in the upper left and lower right quadrants.

There are two red points that stand out from the clusters.

The graph suggests a correlation between pknow and time, with the positive slope indicating a potential increase in pknow with time and the negative slope indicating a decrease.

The diagram is a visual representation of the correlation between the two variables, potentially highlighting areas of interest or anomalies in the data.
Solution Step 3
Solution Step 4

plot of time vs. pknow
Approximate Solution
Notes

- Kind of a weird outcome

- By unlucky initial positioning
  - One data lump at left became three clusters
  - Two clearly distinct data lumps at right became one cluster
Exercise 7-1-5
Pause Here with In-Video Quiz

- Do this yourself if you want to

- Only quiz option: go ahead
Exercise 7-1-5
That actually kind of came out ok...
As you can see

- A lot depends on initial positioning
- And on the number of clusters

- How do you pick which final position and number of clusters to go with?
Next lecture

- Clustering – Validation and Selection of k