|  |  |
| --- | --- |
| from: | Thrall, Anthony D <adthral@nsa.gov> |
| to: | Tony Thrall <tthrall@alumni.stanford.edu> |
| cc: | Tony Thrall <adthral@uwe.nsa.gov> |
| date: | Dec 23, 2024, 5:39 PM |
| subject: | RE: [Non-DoD Source] eda\_4350 as of 22-Dec |

Karen’s slides 2a: Unsupervised learning.

·         Example: kernel-based density estimation

·         Focus on clustering

Cluster analysis:

·         Goal: partition data into n clusters that account for variability of data

·         Minkowski metric (L\_p): Manhattan (p = 1), Euclidean (p = 2)

·         Cosine similarity

·         Hamming, Levenshtein, Damerau-Levenshtein

·         Distance between sets

o   Jaccard similarity

o   Linkage: single, median, average

o   Ward’s min variance criterion

·         Distance between sets

o   Nearest neighbor

o   Agglomerative clustering

o   Minimal spanning tree

·         Text clustering

o   Term-document matrix: m terms in n docs

o   Remove stop-words; remove suffixes (“stemming”)

o   Term-frequency inverse document frequency (tfidf)

o   Sam-I-Am exercise

·         Association rule-mining