Thursday, 11/19/2020

EE 660

MACHINE LEARNING FROM SIGNALS: FOUNDATIONS AND METHODS

Prof. B. Keith Jenkins

Lecture 26

Announcements

- Course (learning experience) evaluations
  - => Please fill this out online
- Today is our last regular lecture; no lecture next Tuesday because of our Quiz
- End-of-Semester Quiz Information (v1) has been posted (format, ground rules, tips)
- There will be an (optional) review session on Saturday (11/11), 12:30 PM - 1:30 PM
  - Using Zoom or Webex (TBA by email)
  - H9W12 is due tomorrow

Today's Topics

- End-of-Semester Quiz Information
- Unsupervised learning (part 3)
  - Agglomerative hierarchical clustering (part 2)
    - Farthest neighbor algorithm
  - Clustering metrics: how to choose K

EE 660 | **End-of-Semester Quiz Information** | Posted: Thur., 11/19/2020
Jenkins

v1.0

Introduction

1. The quiz will be mostly short-answer problems with perhaps some medium-length-answer problems also.
   (a) Short-answer problems will be primarily multiple choice, select all that apply, true/false, etc., in some cases asking for justification or explanation that you will type in. Our current plan is for you to enter your answers to these problems online using your computer (e.g., clicking on correct answers, writing in short text justifications, etc.).
   (b) Other problems (short- and medium-length answer problems) will require you to write out your response. Our current plan is for you to work out the answers to these problems, and provide your final answers, on paper. This will proceed similarly to how an in-class exam or quiz would. At the end of the quiz, you will take pictures of your solutions, convert to pdf, and upload to D2L like you would a homework assignment. You will be allowed some time after the quiz ends to do this.

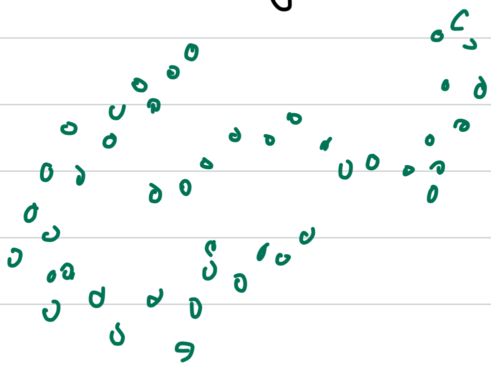
Ground Rules and Requirements

1. The quiz will be open book and open class materials. You may use paper notes/books and/or electronic notes/books, that were part of our EE 660 class. You may use a calculator (software or hardware based).
2. You may not use materials that were not part of our EE 660 class. Use of the internet searching, or communicating with anyone besides the instructor and TAs, is not allowed. To maintain fairness to all, we expect everyone to take the quiz in good faith and with respect of your fellow classmates. Of course, anyone found to be violating these ground rules, will be penalized; note that for significant violations, the penalty can be worse than receiving a 0 on the quiz (e.g., a substantial lowering of the course letter grade).
3. The quiz will be structured as a Zoom class/meeting. **Everyone is required to have a live video feed showing themselves, and ability to see chat messages and hear audio from the instructor/TA, for the duration of the quiz.** You may use your laptop, a webcam, or smartphone camera, etc., for you video feed. The zoom meeting may be recorded.

Tip

1. You are strongly encouraged to write out a some formula sheet(s) in advance. In a short-answer quiz, browsing through course materials to find a formula or other information, can eat up a lot of your time. It might even be helpful to include a reference location with each formula or item you write (e.g., page number in the text, or lecture/page number in class notes), for quick access to more information.

N.N. agl. clustering — tends to be good at linking long strings of points.

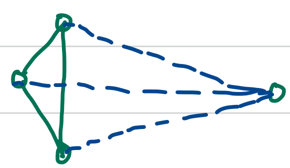


Resulting graph is a tree (no closed loops).
Continue to $\hat{k}_{final} = 1 \Rightarrow$ spanning tree

Use $\delta_{min} \Rightarrow$ minimal spanning tree. (a spanning tree that has min. total length of all edges.)

Farthest Neighbor Algorithm

→ Uses δ_{max}.



Merge rule: connect all nodes in one cluster to all nodes in other cluster.
⇒ Each cluster is represented by a fully connected subgraph.

Comments:

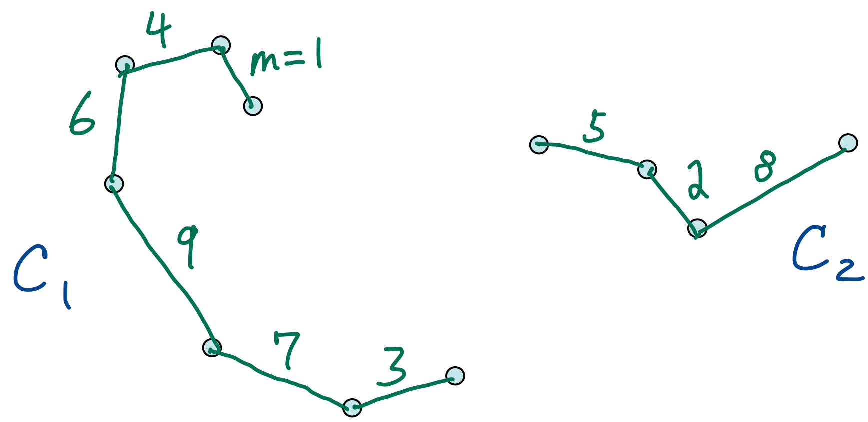
F.N. agl. clust. : prefers compact, dense clusters.

Def.: <u>Diameter of a cluster</u> is the length of the longest edge.

<u>Diameter of a clustering</u> is the length of the longest edge over all clusters.

F.N. alg — at each step, merges the 2 clusters that give the smallest increase in diameter of clustering.

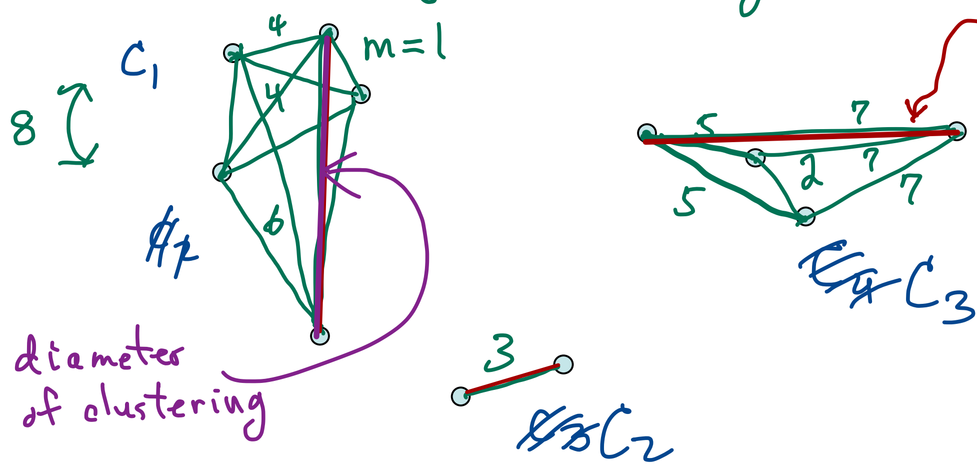
Example: (Nearest neighbor clustering)



H.C.:

$\delta' \geq \delta_{halt}$

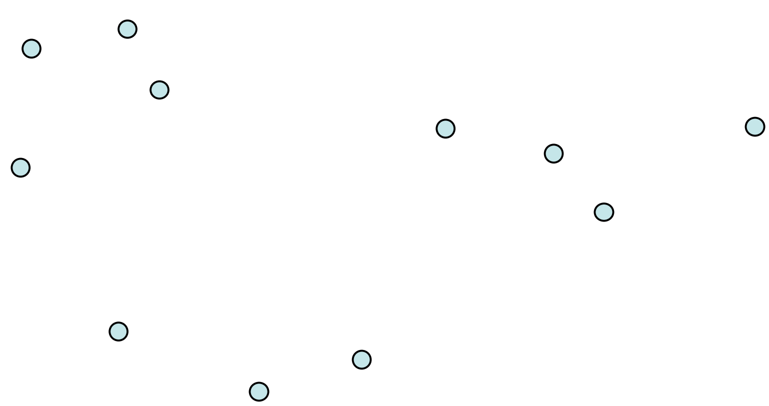
Farthest neighbor clustering



H.C.:

$\hat{K} \leq 3.$

diameter of C_3

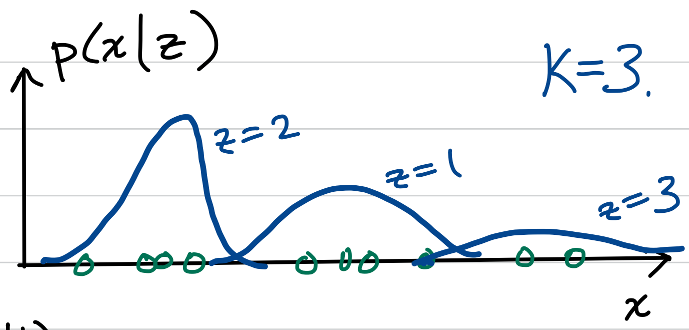


Clustering Metrics: How to Choose K [Xu paper, Sec. II M]

→ Need some criterion or measure for quality of each clustering

For probabilistic mixture models

$\underline{\theta}$ = unknown parameters in $p(x \mid z, \underline{\theta})$



Consider $$p(\underline{x}_i \mid \mathcal{M}_k, \underline{\theta}) = \sum_{k=1}^{K} \pi_k^{(K)} \, p(\underline{x}_i \mid z_i = k, \mathcal{M}_K, \underline{\theta})$$

$\mathcal{M}_k = \mathcal{H} = \{\text{hidden cluster assignments (labels)}\}$

$\rightarrow p(\mathcal{D} \mid \mathcal{M}_k, \underline{\theta})$

$\Rightarrow$ Choose $K^* = $ optimal $K = \underset{K}{\operatorname{argmax}} \; p(\mathcal{D} \mid \mathcal{M}_k, \underline{\hat{\theta}}^{(K)}_{MLE})$

(Choose the K that has largest likelihood)

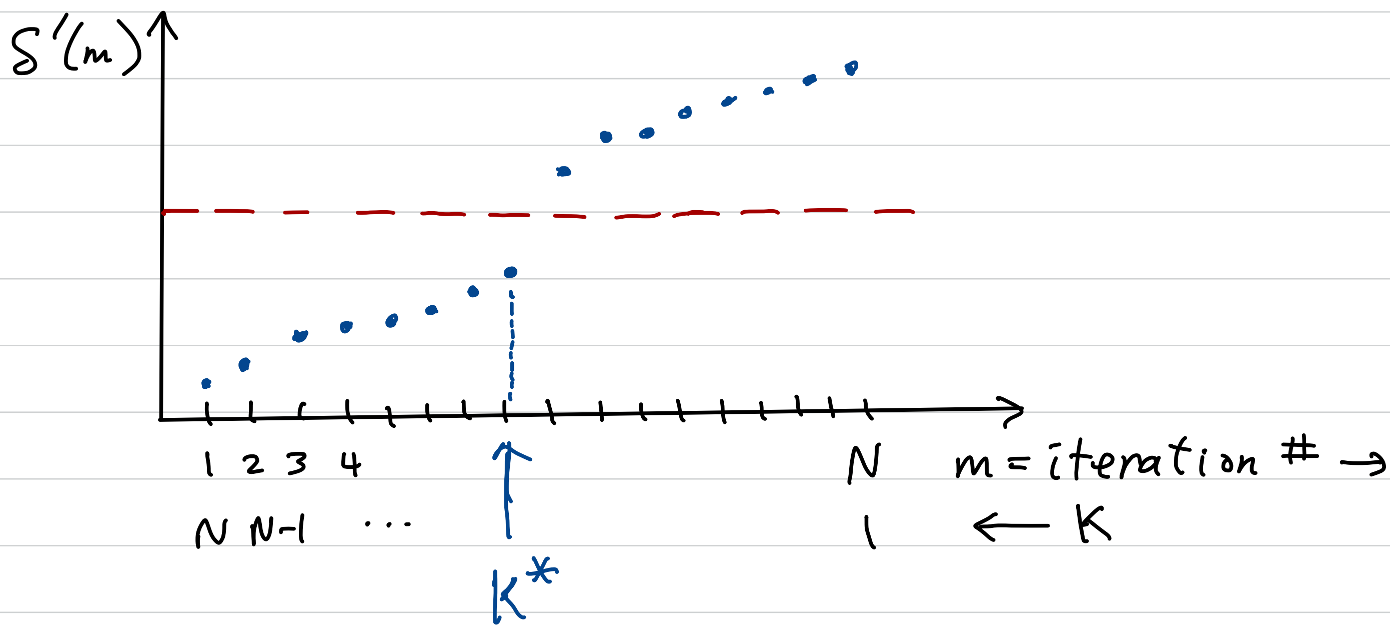
For graphical methods

→ Typically use ad hoc measures

- For agglomerative hierarchical clustering, can use the stepwise criterion:

$$\delta'(m) = \min_{j,k} \delta_{jk} \quad \text{vs. } m = \text{iteration } \#.$$

Ex: NN agl. clust. $\delta_{jk} = \delta_{min}(C_j, C_k)$



- For graphical methods more generally (or any clustering method), can use ad hoc measures

Example of a good measure: Calinski and Harabasz index

$$CH(K) \triangleq \frac{\overbrace{\frac{1}{K-1} \sum_{k=1}^{K} N_k \|\underline{m}_k - \underline{m}\|_2^2}^{\text{Sample variance of cluster means}}}{\underbrace{\frac{1}{N-K} \sum_{k=1}^{K} \underbrace{\sum_{\underline{x}_i \in C_k} \|\underline{x}_i - \underline{m}_k\|_2^2}_{\propto \text{ Sample variance of cluster } C_k}}_{\propto \text{ within-cluster sample variance}}}$$

in which:

N_k = # of data pts. in C_k, $\quad \underline{m}_k \triangleq \frac{1}{N_k} \sum_{\underline{x}_i \in C_k} \underline{x}_i$, $\quad \underline{m} \triangleq \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i$

Let $K^* = \underset{K}{\operatorname{argmax}}\ CH(K)$ ← Was found the best in a systematic comparison of 30 ad hoc cluster quality measures, on a variety of artificial data sets.

If no prior information, then typically use a few quality measures, including CH(K). (cf. Xu paper).