Lecture 28 announcements

- Please fill out course evaluations online; look for email from "USC Learning Experience Evaluations" or c-evals@usc.edu. Deadline is next Thursday (12/6).
- Final exam is Tuesday, 12/11, 2:00 PM - 4:00 PM PST; on-campus exam location is SGM 124.
- DEN students have been notified by denexam@usc.edu of their exam locations.
- Sample final exam and its solution will be posted.
- Project reports are due Tuesday, 12/4, 2:00 PM.
  - Team projects - turn in one written report (for the team) plus code to group-project folder
  - Individual projects - turn in one written report plus one code file to individual-project folder
  - Everyone - follow the provided template (D2L, Week 15) for your report organization

Lecture 28 outline

- Finish unsupervised Learning
  - Clustering metrics: how to choose K (part 2)
- Final exam: coverage, ground rules, timing → ONLINE POLL COMING
- Review

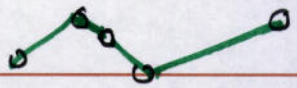
How to choose K in U.L. – Nonstatistical methods

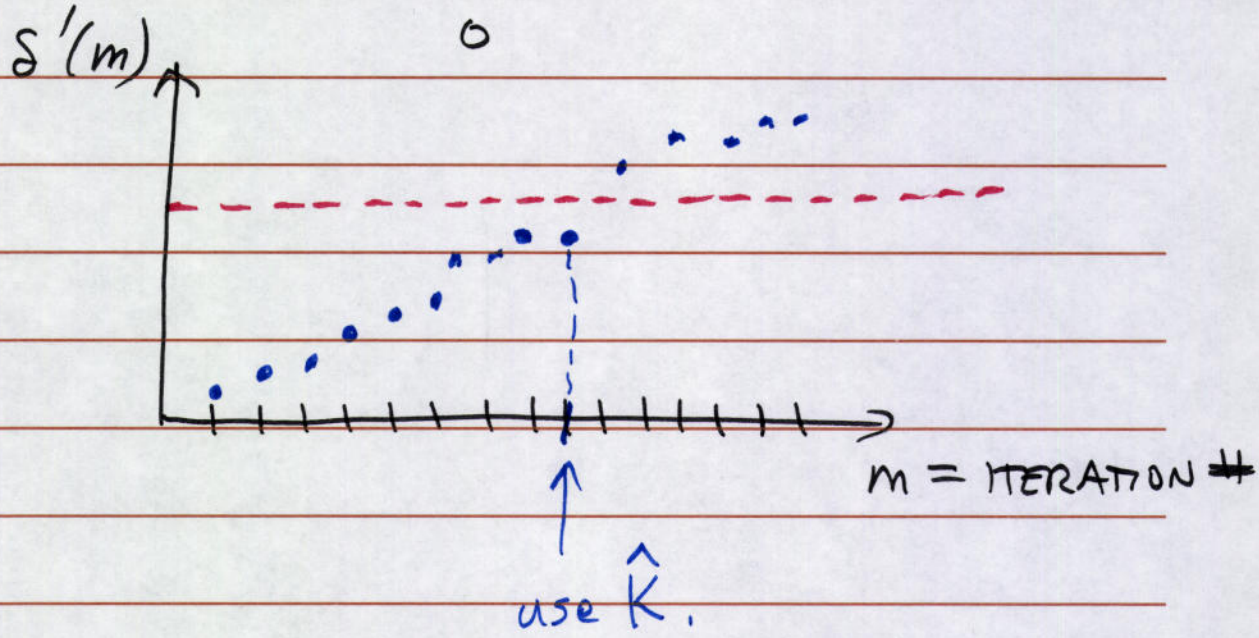
1. For agglomerative hierarchical clustering, can use the stepwise criterion:

Ex: $\delta'(m) = \min_{j,k} \delta_{jk}$

in which for NN/single linkage:

$\delta_{jk} = \delta_{min}(C_j, C_k)$





2. FOR GRAPHICAL METHODS GENERALLY (OR ANY CLUSTERING METHOD), CAN USE OTHER AD HOC MEASURES

CALINSKI AND HARABASZ INDEX:

$$CH(K) \triangleq \frac{\left(\sum_{k=1}^{K} N_k \left\| \underline{m}_k - \underline{m} \right\|_2^2\right)/(K-1)}{\left(\sum_{k=1}^{K} \sum_{\underline{x}_i \in C_k} \left\| \underline{x}_i - \underline{m}_k \right\|_2^2\right)/(N-K)}$$

~ SAMPLE VARIANCE OF CLUSTER MEANS

SAMPLE WITHIN-CLUSTER VARIANCE.

UN-NORM'D SAMPLE VARIANCE OF CLUSTER C_k.

IN WHICH: $N_k = \#$ DATA PTS. IN C_k

$$\underline{m}_k \triangleq \frac{1}{N_k} \sum_{\underline{x}_i \in C_k} \underline{x}_i$$

$$\underline{m} \triangleq \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i .$$

LET $\hat{K} = \underset{K}{\operatorname{argmax}}\ CH(K)$

CH(K) WAS FOUND THE BEST IN A SYSTEMATIC COMPARISON OF 30 AD HOC CLUSTER QUALITY MEASURES.

IN PRACTICE IT'S OFTEN BEST TO USE A FEW CLUSTER QUALITY MEASURES, INCLUDING CH(k).

FINAL EXAM

1. COVERAGE

WILL COVER:

- LECTURES 19 - 28, AND RELATED DISCUSSION, HWs, & HANDOUTS, READING
- ALSO:
  - LOGISTIC REGRESSION: LECT. 6 + MURPHY Ch. 8
  - APPROX. / GENERALIZ. TRADE-OFF + OTHER TARGET TYPES: LECT. 12 + AML 2.2.4, 2.3
  - OVERFITTING: LECT. 13 + AML 4.1

& RELATED HWs, DISCUSSION.

2. TIME PERIOD.

~~2h~~	2h	8	ONLINE POLL WILL BE SENT.
~~1h 40m~~	1h 30m	14	
~~1h 20m~~			

3. MATERIALS ALLOWED.

LIKE MIDTERM (2 SHEETS)	19	
OTHER	0	
4 CALCULATORS		
SIMPLE OR SCI. CALC.	5	
NO CALC.	10	