How is the current course lecture speed?

- Too fast
- Good
- Too slow

Recitation:
K-means, GMMs, EM, PCA

Lyle Ungar

University of Pennsylvania

How are k-means and k-NN related?

- What do they each require you to specify?
- How do the two k's relate?
- When is one likely to be better or worse than the other?

How many clusters?

How would you pick K?

What is the optimal K for PCA?

- **Approximate *x* in terms of *K* eigenvectors, V**
  - $\hat{x}_i = \Sigma_k \; z_{ik} \; v_k$ or $\hat{X}_K = Z_K V_K^T$
- **Distortion**
  - $|| X - \hat{X}_K ||_F^2$

What is the optimal K for K-means?

$$J(\mu, r) = \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} ||\mu_k - \mathbf{x}_i||_2^2$$

r_{ik} 1 if point *i* in cluster *k*

μ_k centroid of cluster *k*

Radial Basis Functions (RBF)

- What is the algorithm?
- What are the hyperparameters?
- Does it use hard or soft clustering?

How to model HW grade distribution?

- Or donation amounts
- Zero-inflated models
  - $p(x) = \pi_1 N(0,0) + \pi_2 N(\mu, \Sigma)$

EM

- In a GMM, the E-step finds the expected value of what?
- In data imputation via feature averaging, the E-step finds the expected value of what?
- In a GMM, the M-step finds the MLE estimate of what?
- In data imputation via feature averaging, the M-step finds the MLE value of what?

EM for missing data

- **Does EM for imputation work when the data are not missing at random?**
  - Why or why not?

How to handle categorical data?

x_1	x_{1R}	x_{1G}	x_{1B}	x_{1NA}
R	1	0	0	0
G	0	1	0	0
B	0	0	1	0
R	1	0	0	0
NA	0	0	0	1

“one hot coding”

What if there are *lots* of categories?

- ZIP codes (42,000)
- FIPS codes
- SIC Codes

1623	Water, Sewer, Pipeline, Comm & Power Line Construction
1629	Heavy Construction, Not Elsewhere Classified[6]
1700	Construction - Special Trade Contractors
1731	Electrical Work
2000	Food and Kindred Products
2011	Meat Packing Plants
2013	Sausages & Other Prepared Meat Products
2015	Poultry Slaughtering and Processing
2020	Dairy Products
2024	Ice Cream & Frozen Desserts
2030	Canned, Frozen & Preserved Fruit, Veg & Food Specialties
2033	Canned, Fruits, Veg, Preserves, Jams & Jellies

What if there are *lots* of categories?

- **Dimensionality reduce: cluster, PCA, …**
- **Possible features**
  - Geolocation
  - Demographics
  - Co-occurrence
  - Product sales, twitter language, …
- **Often someone has already done the clustering**

Eigenwords: SVD practice

I ate ham
You ate cheese
You ate

context

		Word Before					Word After				
		ate	cheese	ham	I	You	ate	cheese	ham	I	You
word	ate	0	0	0	1	2	0	1	1	0	0
	cheese	1	0	0	0	0	0	0	0	0	0
	ham	1	0	0	0	0	0	0	0	0	0
	I	0	0	0	0	0	1	0	0	0	0
	You	0	0	0	0	0	2	0	0	0	0

https://colab.research.google.com/drive/1qDvkPO-t0bQEkG30lDhboTcaiTc_gtCT#scrollTo=Mgy0mHZppM_1&uniqifier=1