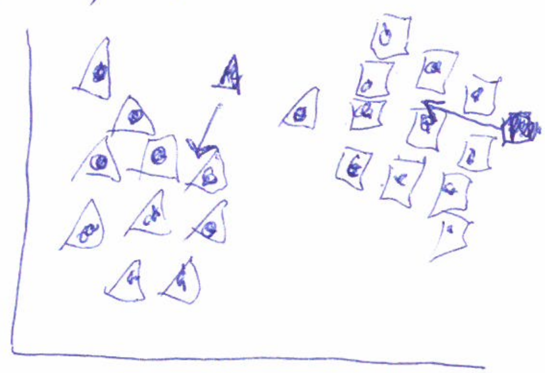
K-Means Clustering

- unsupervised learning
- given features discover clusters ex) Developed / underdeveloped nations
  - species from measurements
  - ~~party~~ partisanship
  - ~~neighborhood~~ poor/wealth neighborhood boundaries

Define $x_1, \dots, x_n$ points

Place $c_1, \dots, c_k$ centroids randomly

Eg Weight



K = 2

Height

① For each x_i:

→ find the nearest centroid $\min d(x_i, c_j)$

→ assign x_i to cluster c_j

② For each c_j:

→ recalculate position of cluster j using the mean

$$\bar{\Delta} = \frac{1}{n_\Delta} \sum_{i \in \Delta} x_{i\Delta} \qquad \bar{\square} = \frac{1}{n_\square} \sum_{i = \square} x_{i\square}$$

Repeat ① & ② until no cluster

3]

Alg	Type	+	−	Best use	Tuning
Neural Networks	supervised	+ very accurate high perf	- slow - uninterpretable - hsn cu psm - complex model structure - need a lot of data	Classification (text & non-text)	- Layers - Backprop iterations - Nodes - Loss functions ↩
K means clusters	~~clust~~ unsupervised	+ fast + simple + interpretation	- specify clusters - clusters should be established	Classification text or non-text	- k
Topic models	unsupervised	+ interpretation + ~~pe~~ useful	- complex/slow - dependent upon structure of data - ~~specify~~ choose topics		- k - sampling method

ʌ

Algor!	Type	+	−	Best Use	Tuning
Random Forests	supervisd	+ high accury ~~+ higher interp~~ + good interpretn	− less interpretn than DT − very compu intensive	classifcn	− number of trees grown − tree depth
Support Vector Machines	supervisd	+ high accur + fast - scales well + allows for regularizn	− ~~Very~~ uninterpretable − sensitive to overfitting − complicated model tuning	regresn classifcns (text & non-text)	~~a~~ - kernel => liner => radial basis fnctn (default) − C - regularizatn param − gamma - radius of influence
Linear Regresn / LASSO	supervisd	+ simple + fast + interpretable	− cannot used for classification − ~~they~~ not flexible − poor performance	regresn (non-text)	− regularizatn penalty λ
Logistic regresn (regularized)	supervisd	+ interpretation	− inflexible for complex data	classification (non-text)	− regularizn penalty λ

Algorithm	Type	+	−	Best Use	Tuning Parameters
K-nn	supervised	+ Very fast + nonparametric + useful for classification or regression	− overfitting − poor accuracy − Low interpret − sensitive to irrelevant features ↓ scale features	classification regression w/ low # of features	K - # of points used in estimation ↑K ↑ prevents overfitting but may reduce accuracy
Naive Bayes	supervised	+ very fast + few assumptions ~~+ high accuracy~~ + highly interpretable + robust to violation of assumptions + Small n	− mostly useful for classification − middling accuracy − TF-IDF pre-processing necessary	document classification	Laplace smoothing
Decision trees	supervised	+ high accuracy + high interpretability for	− Skilled pruning to avoid overfitting − Sensitive to small data perturbations − computationally intensive	classification	↓ pruning / tree depth