Clustering and K-Means

DATA 3301 Introduction to Data Science Lecture Notes 19

Name: _____ Date: _____

Key terms

term	definition (in your own words)
unsupervised learning	
cluster	
centroid	
k-means	
inertia	
convergence	

1. No answer column: unsupervised learning

- **Supervised:** every training row has a _____. **Unsupervised:** we only have the _____; there is nothing to predict and no accuracy to compute.
- Clustering groups observations so those in the same cluster are more _____ to each other than to those in other groups. Clustering ALWAYS needs a _____ measure.
- Euclidean distance from (1, 1) to (4, 5) = √(_____) = ____. Recall from KNN: _____ the features first.
- **Checkpoint.** Write CL (clustering) or CF (classification): spam filter from 10,000 marked emails ____ customer groups ____ Bob Ross "types" ____ penguin species ____

2. Clustering in the wild

- **Bob Ross:** FiveThirtyEight tagged _____ episodes with ____ yes/no elements. Tyner's k-means on 222 paintings found ____ groups, e.g. "shadowy forests." Who named the groups? _____
- **Photo colors:** each pixel is a row with 3 features: _____. Each pixel is repainted with its _____ color: 96,615 colors become 16 (compression).

Big idea: clusters are only as good as your ability to _____ them. 2 or 4 lawmaker clusters? It depends on your _____.

3. The k-means algorithm

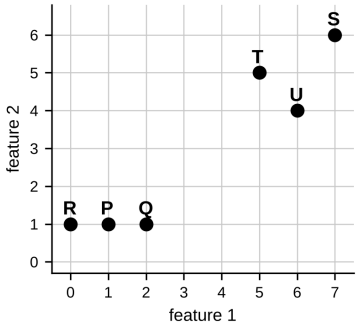
1. Pick ___, the number of clusters (you choose it). **2.** Pick K _____ rows as the starting centroids.

A (assign): each observation joins its _____ centroid. **B (update):** each centroid moves to the _____ of its observations.

Repeat A and B until _____: that is convergence.

- Iris (K = 3): after round 1 the sizes were 56, 87 and ___. Switches in rounds 2 to 8: 19, 12, 5, 4, 3, 1, 1. Converged in round ___ with sizes 50, 54, 46.
- The Smiley data in the live demo: k-means fails on the mouth because it makes _____ clusters.

4. Your turn: one round by hand (K = 2)



Start: C1 at P, C2 at Q. Mark them with X.

Round 1 **squared** distances (no square roots needed to find the closest):

pt	to C1	to C2	joins
P	0	1	C1
Q	1	0	C2
R	____	____	___
S	____	____	___
T	____	____	___
U	____	____	___

- P (1,1) Q (2,1) R (0,1) S (7,6) T (5,5) U (6,4). New C1 = mean of P, R = (____, ___). New C2 = mean of Q, S, T, U = (___, ___).
- Round 2: point ___ switches to C1. Final C1 = (1, 1), C2 = (___, ___). Round 3: nothing changes.

5. Inertia: what k-means makes small

Step 1: distance from a point to its centroid. **Step 2:** _____ it. **Step 3:** _____ over every point.

`inertia = Σ d(x, its centroid)²`

- S = (7, 6), C2 = (6, 5): $(7 - 6)^2 + (6 - 5)^2$ = ___. All six points (add their squared distances): inertia = ___.
- Lower inertia = _____ clusters. Each assign or update step can only make inertia go _____.

6. Starts and scaling

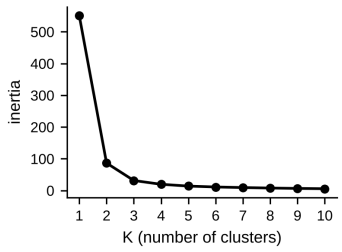
- Same blobs, K = 4: a good start gave inertia 304.4, a bad start _____ (two centroids stuck in one blob). Fix: `n_init` tries several starts and keeps the lowest _____.
- `init='k-means++'` (the default) picks starting centroids that are _____.
- Penguins unscaled: k-means slices only by _____ because grams vary by about ____ and bill length by about 5.5 mm. Fix: _____ first.

7. scikit-learn

```
from sklearn._______ import KMeans
model = KMeans(n_clusters=3, init='random', n_init=10,
random_state=0)
model.fit(X)       # no ___!
model._____________   # one row per centroid
model._______   # cluster number of every row
model._________   # 31.37 for the iris petals
```

- Cluster numbers are just _____: 0 is not first or best. Clusters vs species: all 50 setosa together; ___ of 150 flowers disagree with their species.
- **Default** `n_clusters` is ___, so always set it yourself.

8. Choosing K: the elbow plot



- Inertia always goes _____ as K goes up. At K = 150 (one flower each), inertia = ___.
- Circle the elbow on the plot. Pick K where adding a cluster stops _____: here K = _____.
- Evaluation (labels or no labels): Dec 1.

9. Pull it together

In your own words: how is clustering different from classification?

Synthesize: why must we scale features for k-means, but not for a decision tree?

Exit question: you ran k-means twice and got different clusters. Name one reason why, and one setting that helps.