GENERAL ASSEMBLY

CHOOSING AN UNSUPERVISED MODEL

Mason Gallo, Data Scientist

AGENDA

- Unsupervised overview
- Surprising uses of each algorithm
- Intuition for when to use
- Try it out on your own

OBJECTIVES

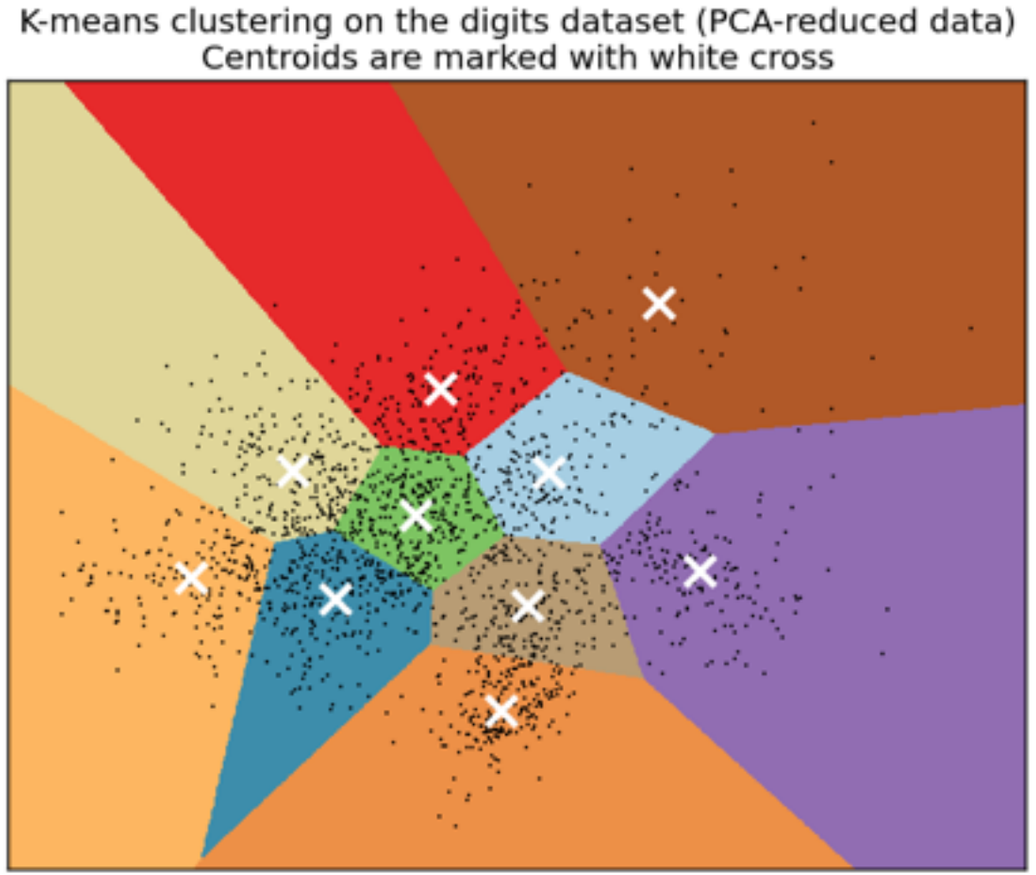
- Intuition on when to use Clustering vs. PCA
- Understanding of real-world implications
- Know how Unsupervised Learning may affect your models
- Implementation in Python

NOTE FOR TODAY'S CLASS

- Unsupervised learning can be challenging because there isn't always a rubric
- Art vs. Science

MOTIVATING EXAMPLE: APPLY ALL YOU'VE LEARNED

UNLEASH YOUR UNSUPERVISED ARSENAL



WHAT ARE THE EFFECTS OF USING CLUSTERING AND PCA ON THE DIGITS DATA?

UNSUPERVISED OVERVIEW

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

	continuous	*categorical*
supervised	*regression*	*classification*
unsupervised	*dimension reduction*	*clustering*

WE DON'T HAVE EXAMPLES OF THE CORRECT ANSWER

WHAT THEY DO

Clustering splits our data into similar groups

Dimensionality Reduction simplifies our data to make it easier to understand

SURPRISING USES OF EACH ALGORITHM

THEY DO WHAT?

Clustering and PCA can improve the results of Supervised models!

THEY DO WHAT?

Clustering and PCA can improve the results of Supervised models!

NOTICE I SAID CAN AND NOT WILL ALWAYS...

THEY DO WHAT?

Clustering can improve Supervised models if we have multiple populations

THEY DO WHAT?

Clustering can improve Supervised models if we have multiple populations

Ex: You want to know average square feet of apartments

What happens when you ask people in NYC but then ask people in Florida?

THEY DO WHAT?

Clustering can improve Supervised models if we have multiple populations

Ex: You want to know average square feet of apartments

What happens when you ask people in NYC but then ask people in Florida?

YOU HAVE MULTIPLE POPULATIONS IN THE SAME DATASET!

THEY DO WHAT?

PCA can improve Supervised models if we have extreme multicollinearity

THEY DO WHAT?

PCA can improve Supervised models if we have extreme multicollinearity

MORE ON THIS IN A MINUTE...

INTUITION FOR USE OF EACH ALGORITHM

SO HOW DO I USE THEM?

Clustering groups your ROWS or OBSERVATIONS into classes

SO HOW DO I USE THEM?

Clustering groups your ROWS or OBSERVATIONS into classes

PCA groups your COLUMNS or FEATURES into numerical components

SO HOW DO I USE THEM?

Clustering groups your ROWS or OBSERVATIONS into classes

PCA groups your COLUMNS or FEATURES into numerical components

THINK: DO I NEED TO WORK WITH MY ROWS OR MY COLUMNS?

WHEN SHOULD I USE CLUSTERING?

- You think there are multiple groups/populations in your dataset
- You need a way to split your dataset into groups in order to explain it
- You want to find similarity among your observations or rows

CLUSTERING PROCESS

- Scale your features to have mean 0 and variance 1
- Run k-means on your training data (if you have train/test)
- Check your results based on intuition and metrics
- Try different values for k and check the counts for each cluster
- If the goal isn't prediction, drop features that are hard to explain and rerun
- Apply your k-means from training data to testing data (if you have train/test)
- Done once you have a model that is easy to explain or improves prediction

WHEN SHOULD I USE PCA?

- You need to visualize multiple dimensions in a single graphic
- You need to group your features to find latent variables to explain
- You are seriously concerned about multicollinearity for your algorithm (NB)

PCA PROCESS

- Scale your features to have mean 0 and variance 1
- Choose number of components < number of features
- Run PCA on training data (if you have test/train)
- Check your variance explained and look for intuition
- If satisfied with results, apply PCA to test data (if you have test/train)
- If predicting, use PCA features as predictors rather than original features

FAQ

DIMENSIONALITY REDUCTION VS FEATURE SELECTION

What's the difference? Aren't they doing the same thing?

- PCA focuses on explaining variance - doesn't look at the response variable at all
- Feature Selection chooses features that help in predicting the response
- PCA is very slow, so this would be a very bad idea if many features

PCA SHOULD PROBABLY NOT BE USED FOR FEATURE SELECTION

WHEN DOES IT MATTER?

Can you give me a real example?

- Naive Bayes assumes that all features aren't related
- PCA guarantees that all features are not correlated!
- DONE

SO YOU MIGHT WANT TO TRY PCA FOR YOUR FEATURES WHEN USING NB

THE RIGHT NUMBER OF CLUSTERS

How do I know?

- “I know it when I see it”
- Usually the most important part is interpretability
- Look at metrics but don’t let them rule decisions
- Up to your judgement based on domain knowledge

MORE ART THAN SCIENCE

WHAT'S THE ANSWER

Can you just tell me which is best to use?

- No. (not that I would if I could - how would you learn?)
- There is no free lunch

YOU NEED TO TEST OUT POTENTIAL SOLUTIONS – NO GUARANTEES!

GIMME NUMBERS BRO

IF YOU MUST...

- k-Means is typically measured with silhouette score
- PCA is typically measured with variance explained

OF COURSE, THIS IS HIGHLY DEPENDENT ON YOUR DOMAIN AND APPLICATION

LET'S CODE!