CS 270
Final Exam Review

Jake Cahoon

Table of Contents

Chapter 1 **Introduction** Page 3

1.1 Note From Jake 3

Chapter 2 **Data Mining Process Model** Page 4

2.1 The Process 4

2.2 The Cycle Picture 5

Chapter 3 **Bayesian Learning** Page 6

3.1 Bayes Theorem 6

Notes — 6

3.2 Bayesian Classifiers 6

3.3 Maximum A Posteriori (MAP) Estimation 7

3.4 Bayes Optimal Classifier 7

3.5 Naive Bayes Classifiers 7

Notes — 7

Chapter 4 **Ensembles** Page 8

4.1 Bias and Variance 8

4.2 Why are Ensembles Helpful? 8

Four Important Criteria — 8

4.3 Voting Ensemble 8

Notes — 9

4.4 Bagging 9

Random Forest — 9 • Notes — 9

4.5 Boosting 9

AdaBoost — 10 • Gradient Boosting — 10 • Notes — 10

4.6 Stacking 11

Notes — 11

Chapter 5 **Clustering** Page 12

5.1 K-Means 12

5.2 Hierarchical Clustering 12

Walkthrough — 12 • Closeness — 14

5.3 Silhouette Score 14
Bois and Gurls to Write Down — 14 • Walkthrough — 14

Chapter 6 Reinforcement Learning Page 16
6.1 Q-Learning 16

Chapter 7 Basic Precision/Recall Page 17

Chapter 8 CNNs Page 18
8.1 Structure 18
Convolutional Layers — 18 • Pooling Layers — 18 • Fully Connected Layers — 18

Chapter 9 Other Deep Learning Topics Page 19
9.1 GANs 19
9.2 RNNs 19
9.3 LSTMs 19
9.4 Transformers 19

Chapter 1

Introduction

1.1 Note From Jake

I hope the last review document was helpful, and I hope this one is as well. To make these, I've gone through the Exam Study Guide topics posted on Learning Suite. With the hope that I don't miss topics that will be on the final, I went through every slide and included all that I deem noteworthy. Despite this endeavor, I will miss topics, so I suggest you use this as a supplement to your study rather than relying solely on it.

I've included a table of contents this time around, so you can jump to the topics you need to study. That being said, let's friggin' do this.

Chapter 2

Data Mining Process Model

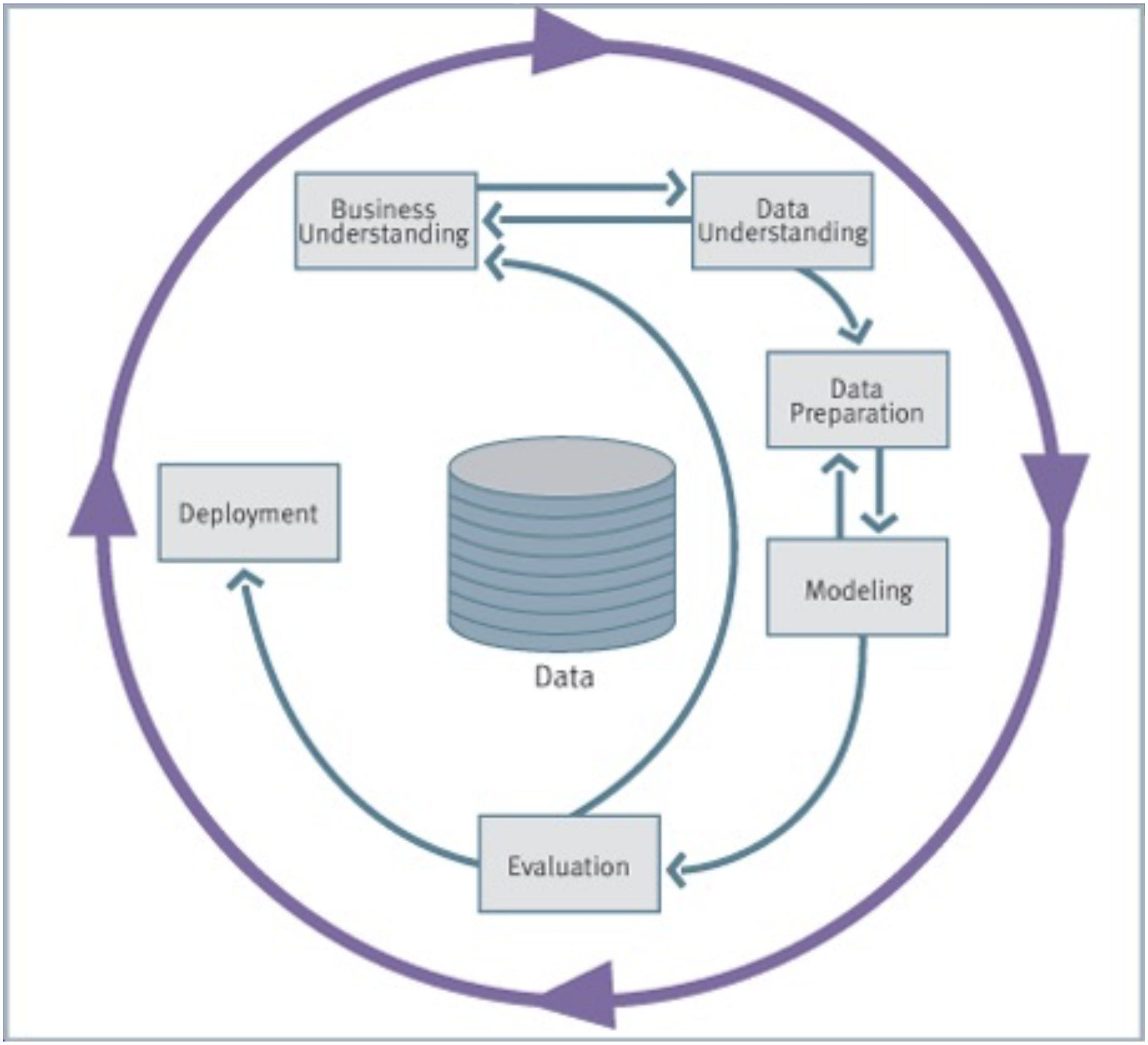
2.1 The Process

Note:-

I'm pulling straight from the slides here. Know this process at a high level.

1. Identify and define the task (business understanding)
    - Understand the context, audience, and problem
    - Tell the story
2. Gather and prepare the data
    - Build a dataset for the task
    - Select/transform/derive features
    - Conduct exploratory data analysis
    - Clean the data
3. Build and evaluate model
4. Deploy the model
    - Evaluate business related results
5. Iterate and improve the model

2.2 The Cycle Picture



Chapter 3

Bayesian Learning

There are two approaches to statistical learning: frequentist and Bayesian. Frequentist statistics is based on the idea of repeated sampling, while Bayesian statistics is based on the idea of starting with prior beliefs and then updating beliefs based on new information.

3.1 Bayes Theorem

Definition 3.1.1

Let C and A be events. Then

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}$$

Note:-

The right side of the equation is based on our current data, while the left side is what we want to find.

3.1.1 Notes

- Prior probabilities are based on prior knowledge. They are the initial beliefs.
- Posterior probabilities are the updated beliefs based on new information.
- $P(C)$ is the prior probability of the Class.
- $P(A)$ is the prior probability of the Attribute.
- $P(A|C)$ is the likelihood of the Attribute given the Class.
- $P(C|A)$ is the posterior probability of the Class given the Attribute.

3.2 Bayesian Classifiers

Given a set of attributes $\{A_1, A_2, \ldots, A_n\}$ and a class C, we can use Bayes Theorem to find the output class C that maximizes $P(C|A_1, A_2, \ldots, A_n)$. For each output class C, do

$$P(C|A_1, A_2, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_n|C)P(C)}{P(A_1, A_2, \ldots, A_n)}$$

3.3 Maximum A Posteriori (MAP) Estimation

Definition 3.3.1

Let D be a dataset and let H be the set of all hypotheses. Then

$$\hat{h}_{MAP} = \underset{h \in H}{\operatorname{argmax}}\, P(h|D)$$

where $\hat{h}_{MAP}$ is the maximum a posteriori hypothesis.

This is guaranteed to be "best", but it is computationally expensive and impractical for large hypothesis spaces.

3.4 Bayes Optimal Classifier

TODO: Figure out what to add here.

3.5 Naive Bayes Classifiers

A simple classifier that assumes that the attributes are conditionally independent given the class. This is a naive assumption, but it works well in practice. We assume that:

$$P(A_1, A_2, \dots A_n|C) = P(A_1|C) \cdot P(A_2|C) \cdot \dots \cdot P(A_n|C)$$

In other words, the probability of all the attributes given the class is the product of the probabilities of each attribute given the class. Then for each $A_i \in A$ and for each $C_j \in C$ we can estimate $P(A_i|C_j)$, which you did when you calculated the probabilities in the HW.

Once we have the probabilities, we can classify a new instance X by

$$\hat{C} = \underset{C_j \in C}{\operatorname{argmax}} \left(P(C_j) \cdot \prod_{i=1}^{n} P(A_i|C_j) \right)$$

3.5.1 Notes

- Various $P(C_j)$ and $P(A_i|C_j)$ are estimated from the training data.
- Stores the probabilities in a table.
- For a new instance X, the classifier calculates the probability of each class given the attributes.
- $\hat{C}$ is the class with the highest probability.
- The true probability is the normalized probability.
- Independence assumption may not hold for some attributes.

Chapter 4

Ensembles

> Two heads are better than one, not because either is infallible, but because they are unlikely to go wrong in the same direction. - C.S. Lewis

4.1 Bias and Variance

- Bias is the error due to overly simplistic assumptions in the learning algorithm.
- Variance is the error due to the algorithm's sensitivity to fluctuations in the training data.

Bias and variance are inversely related. As one goes up, the other goes down. The goal is to minimize both.

4.2 Why are Ensembles Helpful?

By using ensembles, we can reduce the bias and variance of our models. Ensembles combine multiple models to create a stronger model. The idea is that the models will make different errors, and by combining them, we can reduce the overall error. See Dr. Snell's slides for examples.

4.2.1 Four Important Criteria

1. *Independence:* The models should be independent.
2. *Diversity:* The models should be different enough to make different errors.
3. *Decentralization:* The models should be trained on different subsets of the data.
4. *Aggregation:* The models should be combined in a way that reduces error.

4.3 Voting Ensemble

Let T be the training set, $A = \{A_1, A_2, \dots, A_n\}$ be the set of models, and C be the set of classes. Define a function δ as follows:

$$\delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

Then the voting ensemble is defined as

1. For $k = 1$ to N, h_k is a model trained on T using learning algorithm A_k.
2. For a new instance X, the ensemble predicts

$$\hat{C} = \underset{c \in C}{\operatorname{argmax}} \sum_{k=1}^{N} \delta(c, h_k(X))$$

4.3.1 Notes

- Key issues: diversity and independence (too small of a set A and too similar models will not help).
- The models should be trained on different subsets of the data.

4.4 Bagging

Let T be the training set, A be the learning algorithm, N be the number of samples (or bags) of size d drawn from T, C be the set of classes, and δ be defined as in the Voting Ensembles section. Then the bagging ensemble is defined as

1. For $k = 1$ to N, $S_k \subseteq T$ is a sample of size d drawn from T, with replacement and h_k is a model trained on S_k using learning algorithm A.
2. For a new instance X, the ensemble predicts

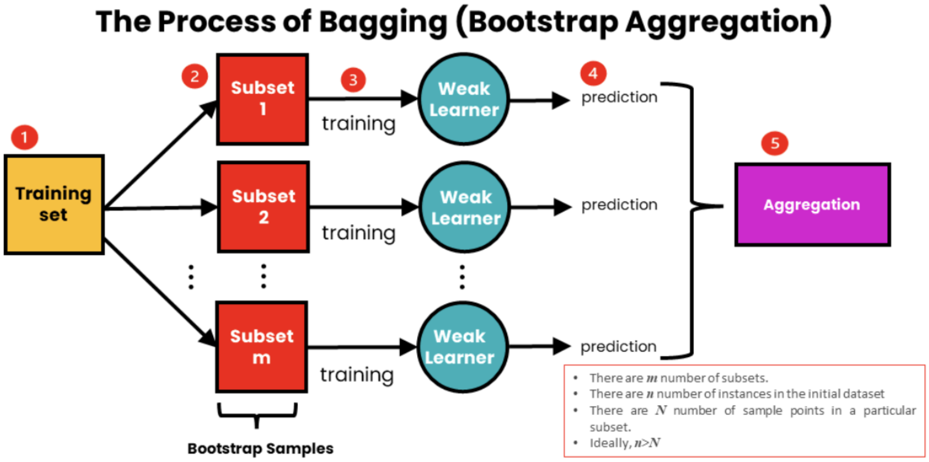
$$\hat{C} = \operatorname*{argmax}_{c \in C} \sum_{k=1}^{N} \delta(c, h_k(X))$$

4.4.1 Random Forest

- A bagging ensemble extension of decision trees
- Each tree is trained on a different bootstrap sample
- Two random variables: the bootstrap sample and the feature subset

4.4.2 Notes

- Train N models on N different bootstrap samples
- Combines the outputs by voting (δ function)
- Decreases error by reducing variance due to unstable learning algorithms
- Homogeneous models are used



The Process of Bagging (Bootstrap Aggregation)

- There are m number of subsets.
- There are n number of instances in the initial dataset
- There are N number of sample points in a particular subset.
- Ideally, n>N

4.5 Boosting

"Boost weak learners into strong learners."

4.5.1 AdaBoost

AdaBoost learns from mistakes by increasing the weights of the misclassified instances. His slides go quite in depth on AdaBoost (adaptive boosting), so I would recommend reviewing them.

1. Start with uniform weights
2. Train a weak learner
3. Update the weigthts based on the performance of the weak learner
4. Repeat

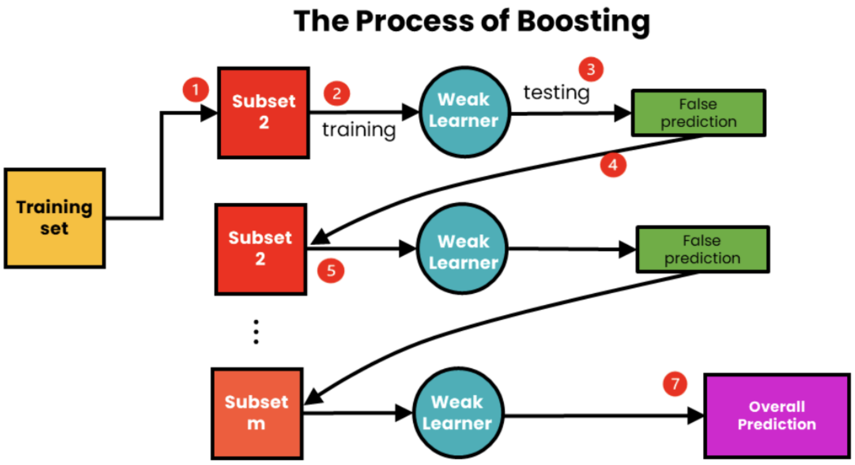
4.5.2 Gradient Boosting

Unlike AdaBoost, Gradient Boosting learns from residual errors, rather than directly updating the weights.

1. Train a weak learner
2. Compute the residuals (errors) on the training set
3. Train a new weak learner to predict the residuals
4. Repeat from step 2 until the residuals are small

4.5.3 Notes

- Great for reducing bias
- Combines the outputs by weighted voting/averaging
- Homogeneous models are used
- Weak learners need to be better than random guessing
- Construct a strong learner by weighted voting of the weak learners



4.6 Stacking

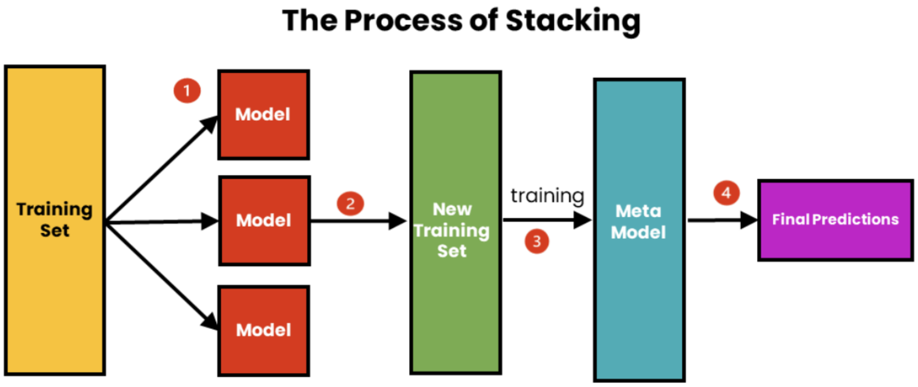
Let T be the base-level training set, N be the number of base-level learning algorithms, $A = \{A_1, A_2, \ldots, A_n, A_{\text{meta}}\}$ be the set of base-level learning algorithms, and A_{meta} be the chosen meta-level learner. Then the stacking ensemble is defined as

1. For $i = 1$ to N, h_i is a model trained on T using learning algorithm A_i.
2. Let T_m be the meta-level training set. $T_m = \emptyset$.
3. For $k = 1$ to $|T|$, $E_k = \{h_1(X_k), h_2(X_k), \ldots, h_N(X_k), y_k\}$.
4. $T_m = T_m \cup E_k$.
5. h_{meta} is a model trained on T_m using learning algorithm A_{meta}.
6. For a new instance X, the ensemble predicts

$$\hat{C} = h_{\text{meta}}(h_1(X), h_2(X), \ldots, h_N(X))$$

4.6.1 Notes

- Improves accuracy by combining the outputs of multiple models
- Hetereogeneous models are used at the base level
- The meta-level model is trained on the outputs of the base-level models



Chapter 5

Clustering

Clustering is an important type of unsupervised learning (PCA is unsupervised). The goal of clustering is to group similar data points together in a cluster.

5.1 K-Means

K-Means is a simple and popular clustering algorithm. The goal is to partition the data into k clusters. The algorithm works as follows:

5.2 Hierarchical Clustering

1. Start with an $n \times n$ adjacency matrix
2. Initialize each point as its own cluster
3. Merge the two “closest” clusters (closeness is either single, complete, or average linkage)
4. Repeat until there is only one cluster

5.2.1 Walkthrough

Example 5.2.1

Compute the clusters for the following adjacency matrix using single linkage.

$$\begin{array}{cc} & \begin{matrix} a & b & c & d & e \end{matrix} \\ \begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix} & \begin{bmatrix} 0 & 1.4 & 0.2 & 1.3 & 1.2 \\ 1.4 & 0 & 1.6 & 0.1 & 0.4 \\ 0.2 & 1.6 & 0 & 1.5 & 1.4 \\ 1.3 & 0.1 & 1.5 & 0 & 0.3 \\ 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{bmatrix} \end{array}$$

Note:-

I recommend x-ing out the values in the clusters you’ve already merged.

Solution:

1. Merge: b and d since the distance between them is 0.1.

$$\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & 0.2 & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & 0.1 & 0.4 \\ c & 0.2 & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & 0.1 & 1.5 & 0 & 0.3 \\ e & 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{array} \implies \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & 0.2 & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \boxed{0.1} & 0.4 \\ c & 0.2 & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \boxed{0.1} & 1.5 & 0 & 0.3 \\ e & 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{array}$$

2. Merge: a and c since the distance between them is 0.2.

$$\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & 0.2 & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \times & 0.4 \\ c & 0.2 & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & 0.3 \\ e & 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{array} \implies \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & \boxed{0.2} & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \times & 0.4 \\ c & \boxed{0.2} & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & 0.3 \\ e & 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{array}$$

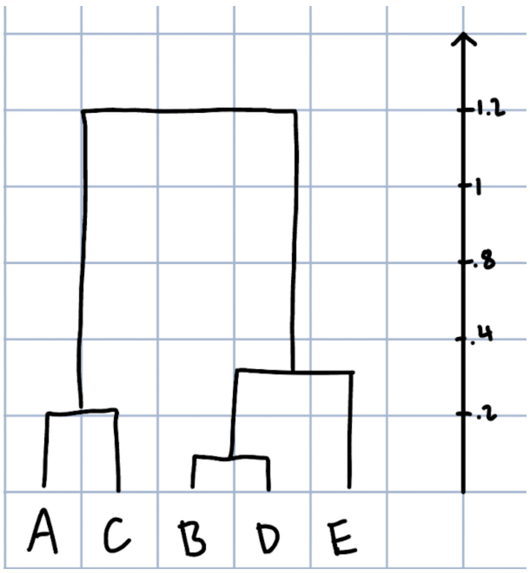
3. Merge: $\{b, d\}$ and e since the distance between them is 0.3.

$$\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & \times & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \times & 0.4 \\ c & \times & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & 0.3 \\ e & 1.2 & 0.4 & 1.4 & 0.3 & 0 \end{array} \implies \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & \times & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \times & 0.4 \\ c & \times & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & \boxed{0.3} \\ e & 1.2 & 0.4 & 1.4 & \boxed{0.3} & 0 \end{array}$$

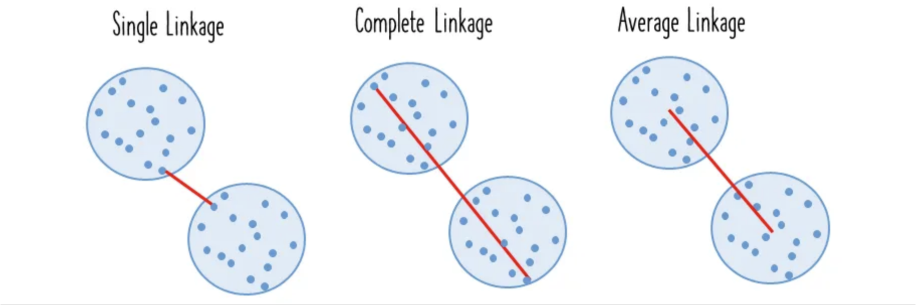
4. Merge: $\{a, c\}$ and $\{b, d, e\}$ since the distance between them is 1.2.

$$\begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & \times & 1.3 & 1.2 \\ b & 1.4 & 0 & 1.6 & \times & \times \\ c & \times & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & \times \\ e & 1.2 & \times & 1.4 & \times & 0 \end{array} \implies \begin{array}{cccccc} & a & b & c & d & e \\ a & 0 & 1.4 & \times & 1.3 & \boxed{1.2} \\ b & 1.4 & 0 & 1.6 & \times & \times \\ c & \times & 1.6 & 0 & 1.5 & 1.4 \\ d & 1.3 & \times & 1.5 & 0 & \times \\ e & \boxed{1.2} & \times & 1.4 & \times & 0 \end{array}$$

5. Draw the dendrogram dude



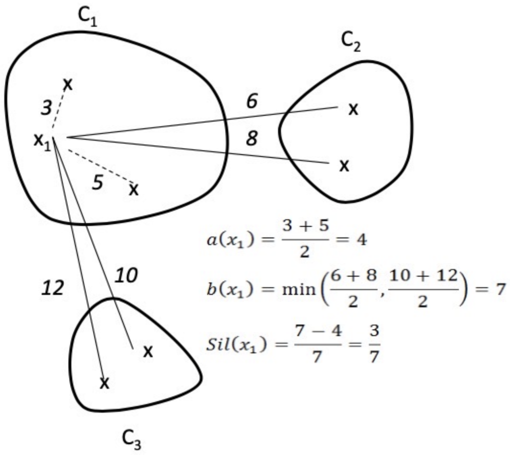
5.2.2 Closeness



- Single linkage can lead to long chained clusters
- Complete linkage finds more compact clusters
- Average linkage is used less because it is computationally expensive

5.3 Silhouette Score

- A measure of how similar an object is to its own cluster compared to other clusters
- Ranges from -1 to 1 with 1 being the best
- A score of 1 means the object is well matched to its own cluster and poorly matched to neighboring clusters
- A score of 0 means the object is on the boundary of two clusters (ambiguous)
- A score of -1 means the object is probably in the wrong cluster


C₁
C₂
C₃
x
3
x₁
6
x
8
x
5
x
12
10
x
x
a(x₁) = (3 + 5)/2 = 4
b(x₁) = min((6 + 8)/2, (10 + 12)/2) = 7
Sil(x₁) = (7 − 4)/7 = 3/7

5.3.1 Bois and Gurls to Write Down

Let i be a point in a cluster. Then

- $a(i)$: The average distance (dissimilarity) between i and all other points in the same cluster (minimize a)
- $b(i)$: The minimum average distance between i and all points in any other cluster (maximize b)
- Once we have a and b, we can calculate the silhouette score by

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

5.3.2 Walkthrough

Example 5.3.1

Compute the silhouette scores for each point, each cluster, and the overall score for the following:

Cluster	Point	x	y
1	a	0.8	0.7
1	b	0.9	0.8
2	c	0.6	0.6
2	d	0	0.2
2	e	0.2	0.1

Note:-

I would highly recommend making an adjacency matrix for easy calculation.

Solution: I recommend drawing a table and filling it out. Here's what it should look like:

Cluster	Point	x	y	$a(i)$	$b(i)$	s(i)
1	a	0.8	0.7	0.2	0.93	0.785
1	b	0.9	0.8	0.2	1.13	0.832
2	c	0.6	0.6	0.95	0.4	-0.579
2	d	0	0.2	0.65	1.4	0.536
2	e	0.2	0.1	0.6	1.3	0.538

I'll walk through two points, a and c.

$$\begin{aligned} a(a) &= \frac{0.1+0.1}{1} \\ &= 0.2 \\ b(a) &= \frac{0.3+1.3+1.2}{3} \\ &= 0.93 \\ s(a) &= \frac{0.93-0.2}{\max(0.2, 0.93)} \\ &= \frac{0.73}{0.93} \\ &\approx 0.785 \end{aligned}$$

$$\begin{aligned} a(c) &= \frac{1+0.9}{2} \\ &= 0.95 \\ b(c) &= \frac{0.5+0.3}{2} \\ &= 0.4 \\ s(c) &= \frac{0.4-0.95}{\max(0.95, 0.4)} \\ &= \frac{-0.56}{0.95} \\ &\approx -0.579 \end{aligned}$$

Now take the average of each cluster for the cluster scores:

$$\begin{aligned} s(1) &= \frac{0.785+0.832}{2} \approx 0.804 \\ s(2) &= \frac{-0.579+0.536+0.538}{3} \approx 0.165 \end{aligned}$$

Finally, take the average of the cluster scores for the overall score:

$$s = \frac{0.785+0.832-0.579+0.536+0.538}{5} \approx 0.421$$

Chapter 6

Reinforcement Learning

6.1 Q-Learning

Chapter 7

Basic Precision/Recall

Chapter 8

CNNs

8.1 Structure

8.1.1 Convolutional Layers

8.1.2 Pooling Layers

8.1.3 Fully Connected Layers

Chapter 9

Other Deep Learning Topics

9.1 GANs

9.2 RNNs

9.3 LSTMs

9.4 Transformers