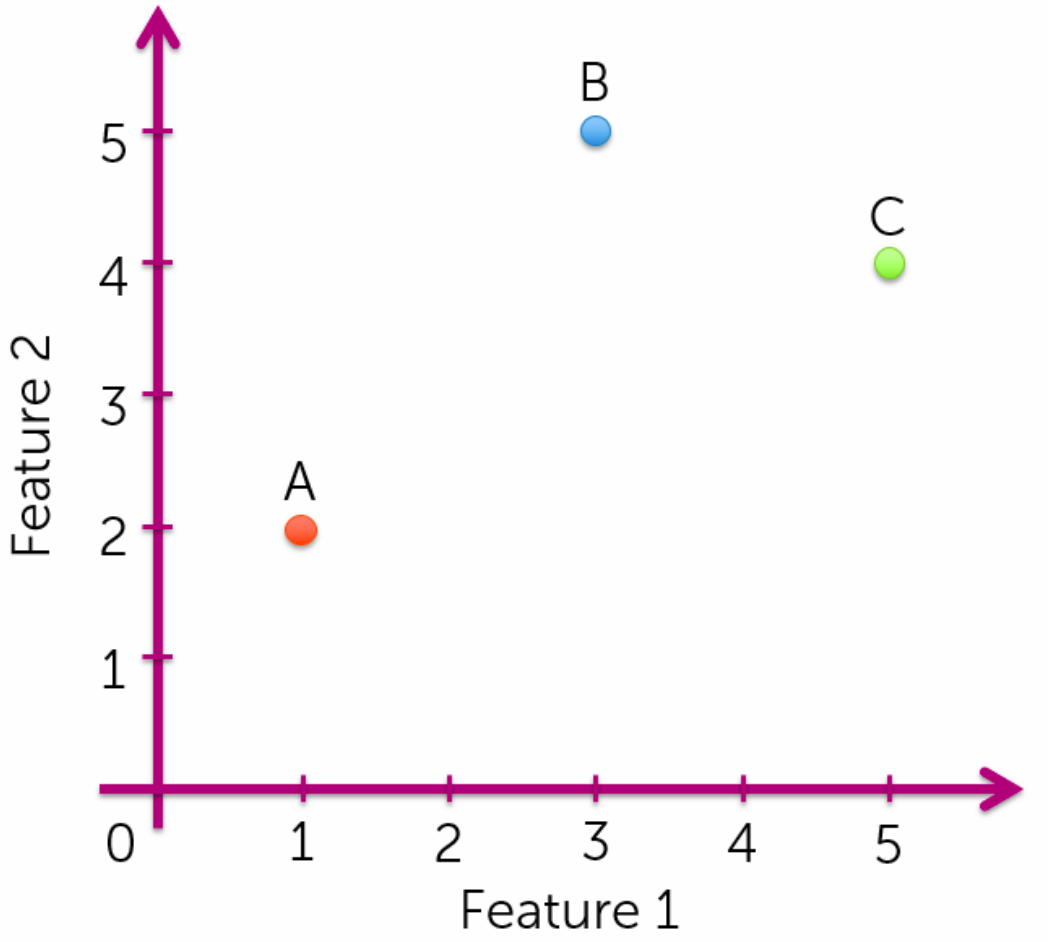
Representations and metrics

7 questions

1 point

1.
Consider three data points with two features as follows:



Among the three points, which two are closest to each other in terms of having the **smallest Euclidean distance**?

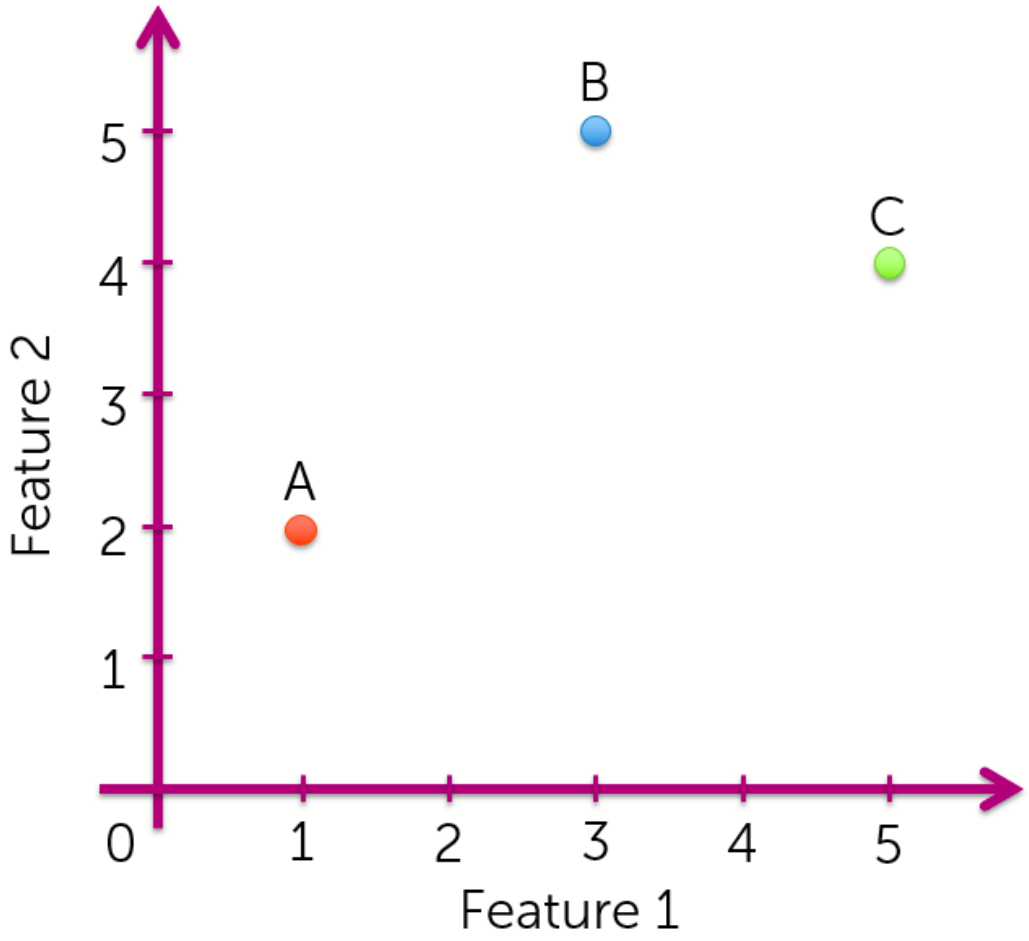
- A and B

A and C

B and C

1 point

2.
Consider three data points with two features as follows:



Among the three points, which two are closest to each other in terms of having the **largest cosine similarity** (or equivalently, **smallest cosine distance**)?

A and B

A and C

B and C

1 point

3.
Consider the following two sentences.

- Sentence 1: The quick brown fox jumps over the lazy dog.
- Sentence 2: A quick brown dog outpaces a quick fox.

Compute the Euclidean distance using word counts. To compute word counts, turn all words into lower case and strip all punctuation, so that "The" and "the" are counted as the same token. That is, document 1 would be represented as

$$x = [\#\text{ the}, \#\text{ a}, \#\text{ quick}, \#\text{ brown}, \#\text{ fox}, \#\text{ jumps}, \#\text{ over}, \#\text{ lazy}, \#\text{ dog}, \#\text{ outpaces}]$$

where $\#\text{ word}$ is the count of that word in the document.

Round your answer to 3 decimal places.

3.606

---

1 point

4.
Consider the following two sentences.

- Sentence 1: The quick brown fox jumps over the lazy dog.
- Sentence 2: A quick brown dog outpaces a quick fox.

Recall that

cosine distance = 1 - cosine similarity = $1 - \frac{x^T y}{||x||||y||}$

Compute the **cosine distance** between sentence 1 and sentence 2 using word counts. To compute word counts, turn all words into lower case and strip all punctuation, so that "The" and "the" are counted as the same token. That is, document 1 would be represented as

$$x = [\#\text{ the}, \#\text{ a}, \#\text{ quick}, \#\text{ brown}, \#\text{ fox}, \#\text{ jumps}, \#\text{ over}, \#\text{ lazy}, \#\text{ dog}, \#\text{ outpaces}]$$

where $\#\text{ word}$ is the count of that word in the document.

Round your answer to 3 decimal places.

0.565

1 point

5.
(True/False) For positive features, cosine similarity is always between 0 and 1.

- True
- False

1 point

6.
Using the formula for TF-IDF presented in the lecture, complete the following sentence:

A word is assigned a zero TF-IDF weight when it appears in ____ documents. (N: number of documents in the corpus)

- N - 1
- N/2
- N
- 0.1*N
- 100

1 point

7.
Which of the following does **not** describe the word count document representation?

- Ignores the order of the words

- [ ] Assigns a high score to a frequently occurring word
- [ ] Penalizes words that appear in every document

Submit Quiz