Discussion 7

DSC 80

2024-05-17

1. FA23 Final Exam Problem 8
2. WI23 Final Problem 7
3. SP23 Final Problem 5.3
4. FA23 Final Problem 9
5. Attendance

Section 1

FA23 Final Exam Problem 8

FA23 Final Exam Problem 8

Document number	content
1	yesterday rainy today sunny
2	yesterday sunny today sunny
3	today rainy yesterday today
4	yesterday yesterday today today

Bag of Words

Using a bag-of-words representation, which two documents have the largest dot product?

- What is a bag-of-words representation?
- What do we need in order to compute a dot product?

Bag of Words Representation

Document number	yesterday	rainy	today	sunny
1	1	1	1	1
2	1	0	1	2
3	1	1	2	0
4	2	0	2	0

Now, how do we compute a dot product between two documents?

Solution

The largest dot product is between documents 3 and 4, with a dot product of $(1 \cdot 2) + (1 \cdot 0) + (2 \cdot 2) + (0 \cdot 0) = 6$.

Question: Why might the dot product, by itself, not be a good document similarity metric?

Cosine Similarity

Using a bag-of-words representation, what is the *cosine similarity* between documents 2 and 3?

What's the formula for cosine similarity?

Solution

The cosine similarity between two vectors $\vec{a}$ and $\vec{b}$ is $\frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||}$.

So, we just need to calculate the dot product of documents 2 and 3, and the magnitude of each vector.

Solution

The dot product of documents 2 and 3 is $1 + 0 + 2 + 0 = 3$, and the magnitude of both documents is the same value, $\sqrt{1^2 + 0^2 + 1^2 + 2^2} = \sqrt{6}$.

So, the cosine similarity is $\frac{3}{\sqrt{6} \cdot \sqrt{6}} = \frac{1}{2}$.

TF-IDF

Which words have a TF-IDF of 0 for all four documents?

When does TF-IDF equal 0?

TF-IDF

TF-IDF multiplies a TF and IDF term, with the IDF term for a given word t defined as:

$$\log(\frac{\text{documents}}{\text{documents containing t}})$$

So, if a term t appears in every document, this fraction is 1, and $\log(1) = 1$.

Section 2

WI23 Final Problem 7

WI23 Final Problem 7

We decide to build a classifier that takes in a state's demographic information and predicts whether, in a given year, a state's mean math score was greater than its mean verbal score (1), or a state's mean math score was less than or equal to its mean verbal score (0).

The simplest possible classifier we could build is one that predicts the same label (1 or 0) every time, independent of all other features.

Problem

If a $>$ b, then the constant classifier that maximizes training accuracy predicts 1 every time; otherwise, it predicts 0 every time.

For which combination of a and b is the above statement **not guaranteed to be true**?

Question: Before we even look at the options, what does this statement even mean?

Options

```
a = (sat['Math'] > sat['Verbal']).mean(); b = 0.5

a = (sat['Math'] - sat['Verbal']).mean(); b = 0

a = (sat['Math'] - sat['Verbal'] > 0).mean(); b = 0.5

a = ((sat['Math'] / sat['Verbal']) > 1).mean() - 0.5; b = 0
```

Solution

The solution is option 2, since it's the only one that doesn't directly compare the values of `Math` and `Verbal` – we only care about which one is larger, not how different they are on average.

Part II

Suppose we train a classifier that achieves an accuracy of $5/9$ on our training set. Typically, RMSE is used as a performance metric for regression models, but mathematically, nothing is stopping us from using it for classification models as well. What is the RMSE of our classifier on our training set?

- What is the definition of RMSE? (If you know what it stands for, that's basically it)
- Since predictions are 0 or 1, what is the magnitude of a single prediction error?

Solution

An accuracy of $5/9$ means that we made errors on $4/9$ data points. Each error has the same magnitude, of 1, and each correct prediction has a magnitude of 0.

Since $1^2 = 1$, the mean squared error is $((5/9) \cdot 0) + ((4/9) \cdot 1) = 4/9$, so the RMSE is just the square root of this, or $2/3$.

Section 3

SP23 Final Problem 5.3

SP23 Final Problem 5.3

Chen downloaded 4 reviews of a new vacuum cleaner from Amazon (as shown in the 4 sentences below).

`Sentence 1: 'if i could give this vacuum zero stars i would'`

`Sentence 2: 'i will not order again this vacuum is garbage'`

`Sentence 3: 'Love Love Love i love this product'`

`Sentence 4: 'this little vacuum is so much fun to use i love it'`

TF-IDF

X is the TF-IDF of the word "vacuum" in sentence 1 with the original dataset.

Chen then replaces sentence 3 with `'Love Love Love i love this vacuum'`.

Y is the *new* value of TF-IDF for vacuum in sentence 1.

Solution

- What do we know about the value X?

Solution

- What do we know about the value X?
- What changes when we switch the old sentence for the new one?

Solution

- What do we know about the value X?
- What changes when we switch the old sentence for the new one?
- What do we know about the value of Y?

Section 4

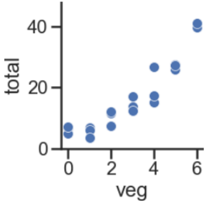
FA23 Final Problem 9

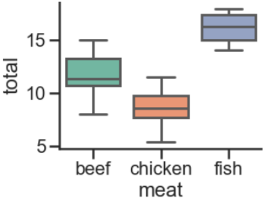
FA23 Final Problem 9

We're going to build a linear regression model to predict the total price of groceries given two features – the amount of vegetable (veg) and which meat is purchased, of beef, chicken, and fish.

Figures

veg	meat	total
1	beef	13
3	fish	19
2	beef	16
0	chicken	9





Problem

We're going to look at four different potential models we could use, and answer whether each model coefficient w is positive (+), negative (-), or 0.

Note: Not sure if you've seen one-hot encoded features, but essentially, think of the term ($meat = chicken$), for example, as being a feature that is 1 if the meat bought is chicken, and 0 otherwise.

Part 1

$H(x) = w_0$

- What would this model look like on a graph?
- Using the least-squares method, what should be the value of w_0? Is that positive, negative, or 0?

Part 2

$H(x) = w_0 + w_1 \cdot \text{veg}$

- What would this model look like on a graph?
- What information from the figures can we use for this?

Part 3

$H(x) = w_0 + w_1 \cdot \text{meat} = \text{chicken}$

- We don't have a graph for this anymore, and we have a binary (one-hot encoded) feature!
- What would this model predict in different situations?
- What data from the figures can we use to figure out the sign of the coefficients?

Part 4

$$H(x) = w_0 + w_1 \cdot (\text{meat} = \text{beef}) + w_2 \cdot (\text{meat} = \text{chicken})$$

- What would this model predict in different situations?
- What data from the figures can we use to figure out the sign of the coefficients?

Section 5

Attendance

Attendance

Once I give you a number, fill out the following Google form:
https://forms.gle/iwMXdxcxiwqTMiCJ8