DSC 80 Discussion 7 Worksheet

1 FA23 Final Exam Problem 8

Consider the following corpus:

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

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

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

Which words have a TF-IDF score of 0 for all four documents? Select all words that apply.

`[] yesterday [] rainy [] today [] sunny`

2 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.

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? **Select one.**

```
[] 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
```

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?

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'
```

X is the Term frequency-Inverse Document Frequency (TF-IDF) of the word `vacuum` in sentence 1. Chen replaces sentence 3 with the following new sentence/review.

```
New Sentence 3:  'Love Love Love i love this vacuum'
```

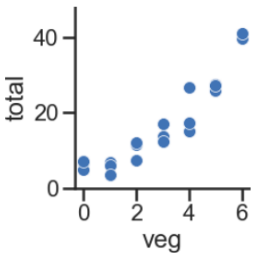
Y is the TF-IDF of the word `vacuum` in sentence 1 after the sentence 3 is replaced by the new sentence 3. Given the above information, which of the following statements is true? **Select one.**

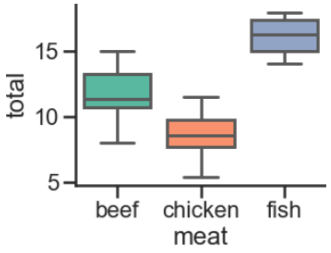
- $X = Y$
- $X > Y > 0$
- $X > 0$ and $Y = 0$
- $X > 0$ and $Y > X$

4 FA23 Final Problem 9

Every week, Lauren goes to her local grocery store and buys a varying amount of vegetable but always exactly one pound of meat. We use a linear regression model to predict her total grocery bill. We've collected a dataset containing the pounds of vegetables bought, the type of meat bought, and the total bill. Below we display the first few rows of the dataset and two plots generated using the entire training set.

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





Suppose we fit the following linear regression models to predict `total`. Based on the data and visualizations above, for each model, determine whether **each fitted model coefficient** w^* is positive, negative, or exactly 0. The notation "meat=beef", for example, refers to the one-hot encoded `meat` column with value 1 if the original value in the `meat` column was `beef` and 0 otherwise.

1. $H(x) = w_0$ $\quad w_0$ _____
2. $H(x) = w_0 + w_1 \cdot \text{veg}$ $\quad w_0$ _____ w_1 _____
3. $H(x) = w_0 + w_1 \cdot \text{meat} = \text{chicken}$ $\quad w_0$ _____ w_1 _____
4. $H(x) = w_0 + w_1 \cdot (\text{meat} = \text{beef}) + w_2 \cdot (\text{meat} = \text{chicken})$ $\quad w_0$ _____ w_1 _____ w_2 _____

Suppose we fit $H(x) = w_0 + w_1 \cdot \text{veg} + w_2 \cdot (\text{meat=beef}) + w_3 \cdot (\text{meat=fish})$, and find that $\vec{w}^* = [-3, 5, 8, 12]$.

What is the prediction of this model on the **first** point in our dataset? _____