Welcome to "How to win a data science competition"

Competition mechanics

Recap of main ML algorithms

Software/Hardware requirements

Feature preprocessing and generation with respect to models

Feature extraction from text and images

✓ **Video:** Bag of words
10 min

✓ **Video:** Word2vec, CNN
13 min

✓ **Practice Quiz:** Feature extraction from text and images
4 questions

✓ **Quiz:** Feature extraction from text and images
4 questions

✓ **Reading:** Explanation for quiz questions
10 min

Explanation for quiz questions

Question 1

Select true statements about n-grams.

Correct answers:

- N-grams can help utilize local context around each word. Correct, because ngrams encode sequences of words.
- N-grams features are typically sparse. Correct. Ngrams deal with counts of words occurrences, and not every word can be found in a document. For example, if we count occurrences of words from an english dictionary in our everyday speech, a lot of words won't be there, and that is sparsity.

Incorrect answers:

- N-grams always help increase significance of important words. No, ngrams deals with words occurrences and not their importance.
- Levenshteining should always be applied before computing n-grams. Although, there is Levenshtein distance, there is no such thing as Levenshteining.

Question 2

Select true statements.

Correct answers:

- Bag of words usually produces longer vectors than Word2vec. Correct! Number of features in Bag of words approach is usually equal to number of unique words, while number of features in w2v is restricted to a constant, like 300 or so.
- Semantically similar words usually have similar word2vec embeddings. Correct. This is one of the main benefits of w2v in competitions.

Incorrect answers:

- Meaning of each value in BOW matrix is unknown. Incorrect. Meaning of a value in BOW matrix is the number of a word's occurrences in a document.
- You do not need bag of words features in a competition if you have word2vec features. Incorrect. Both approaches are valuable and you should try to utilize both of them.

Question 3

Suppose in a new competition we are given a dataset of 2D medical images. We want to extract image descriptors from a hidden layer of a neural network pretrained on the ImageNet dataset. We will then use extracted descriptors to train a simple logistic regression model to classify images from our dataset.

We consider to use two networks: ResNet-50 with imagenet accuracy of X and VGG-16 with imageNet accuracy of Y (X < Y). Select true statements.

Correct answers:

- It is not clear what descriptors are better on our dataset. We should evaluate both. Correct! This depends on the a specific dataset and a specific task, so you should evaluate both!

Incorrect answers:

- With one pretrained CNN model you can get only one vector of descriptors for an image. Incorrect. With one CNN you can get different descriptors from different layers.
- Descriptors from ResNet 50 will always be better than the ones from VGG-16 in our pipeline. Incorrect. Although, ResNet50 shows better performance on Imagenet, this depends on the a specific dataset and a specific task.
- For any image descriptors from the last hidden layer of ResNet-50 are the same as the descriptors from the last hidden layer of VGG-16. Incorrect in general. Moreover it is hard to come up with an image that will have the

same descriptors in both networks.
- Descriptors from ResNet-50 and from VGG-16 are always very similar in cosine distance. Incorrect. This depends on the a specific dataset and a specific task.

Question 4

Data augmentation can be used at (1) train time (2) test time

Correct answer:

True, True. Data augmentation can be used (1) to increase the amount of training data and (2) to average predictions for one augmented sample.

✓ Completed

Go to next item

Like Dislike Report an issue