Bag of N-Gram Document Classification

Alexandra Simonoff

2018/10/10

1 Introduction

For assignment 1 we were tasked with creating a binary classification of sentiment (pos/neg) for IMDB movie reviews using a bag of n-grams classifier. Data was accessed via `http://ai.stanford.edu/~amaas/data/sentiment/`, a source that provides a set of 25,000 rated movie reviews for training and 25,000 for testing. In order to achieve a strong model we performed hyperparameter tuning on several aspects of the model building and vocabulary generation processes. In addition to parsing the data we also chose to stem the words and remove stop words in hopes that it would remove noise from our n-gram tokens.

Full analysis and code can be found at `https://github.com/AMS889/DS-GA1011_Homework1`

2 Analysis of Results

The first hyperparameter we tuned was the learning rate used by the optimizer in our bag of n-grams model training. We held all other parameters constant and varied learning rate across [0.0001, 0.001, 0.01, 0.1, 1]. With the exception of 0.0001 all learning rates performed reasonably similarly with respect to validation accuracy as shown in Figure 1, however there was variability in how volatile the trends were by learning rate. We see the highest training accuracy for 0.01 and 0.001 and of the two 0.01 reached that level earlier (Figure 2). Looking over validation accuracy a learning rate of 0.01 appears more stable than 0.001 and ultimately validation accuracies are close enough that either seem a reasonable choice. Thus we move forward using a learning rate of 0.01. We also tested how learning rate annealing performed on our model using an initial learning rate of 0.01. We used the *torch.optim.lr_scheduler.ReduceLROnPlateau* scheduler to anneal our learning rate. It appears that the two options perform nearly identically on both validation and training data as shown in Figure 3 and Figure 4. We will move forward with a 0.01 learning rate that does not anneal as this trend has slightly higher average validation accuracy.

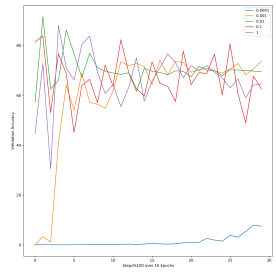

Figure 1: Learning Rate Validation Accuracy Training Curves

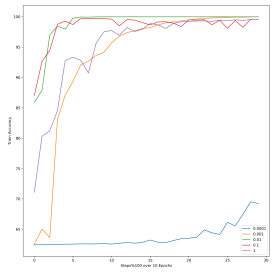

Figure 2: Learning Rate Training Accuracy Training Curves

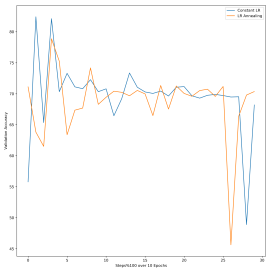

Figure 3: Learning Rate Annealing Validation Accuracy Training Curves

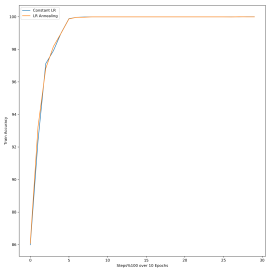

Figure 4: Learning Rate Annealing Training Accuracy Training Curves

Next we tested the performance of various embedding dimension sizes. Again we see high variability and rather consistent performance across various embedding dimensions in the validation accuracy (Figure

5). An embedding dimension of 1000 starts high, improves slightly and appears to steady over the 10 epochs. Given all dimensions hit 100% training accuracy (Figure 6) and the 1000 dimension appears to perform consistently well on the validation set we will move forward with a 1000 embedding dimension.

Now that we have a learning rate and an embedding dimension set, let's move back slightly in the modeling process to the vocabulary generation. While our vocabulary size will be affected by our n-gram value (n), we ran this optimization on bigrams initially with the knowledge that if our trigram model appeared to have strong performance we could adjust this value. In an ideal world (with a ton of computational power) we would have tested several vocabulary sizes for each n-gram possibility and taken the best combination. However given we have limited time and resources we will perform these optimizations in a microcosm. Just to see what an ill-sized vocabulary size looks like we ran the model on a vocabulary size of 5000. The validation accuracy varies greatly and the training accuracy is significantly lower than all other sizes (Figure 8). We find the most stable, strong performance in the vocabulary size of 80000 (Figure 7) with bigrams and as such will move forward with that size.

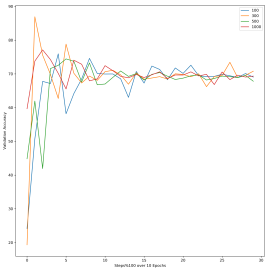

Figure 5: Embedding Dimension Validation Accuracy Training Curves

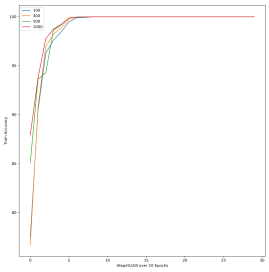

Figure 6: Embedding Dimension Training Accuracy Training Curves

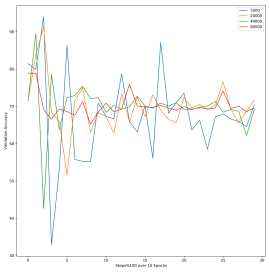

Figure 7: Vocab Size Validation Accuracy Training Curves

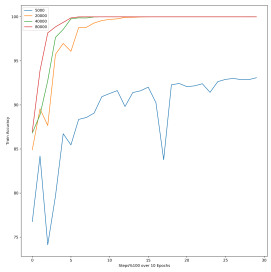

Figure 8: Vocab Size Training Accuracy Training Curves

Now we will move back even further in the modeling process to the choice of the n-gram value to use. Unsurprisingly the quadrigram appears to struggle to fit our dataset (Figure 10) given there are likely many quadrigrams and we are using an 80000 vocabulary size. Given enough time and computation power I would experiment with higher vocabulary sizes. Trigrams also appear to struggle fitting our data but unigrams and bigrams perform well on our validation set (Figure 9) with bigrams edging out unigrams in both validation and training accuracy. Therefore we will move forward with bigrams.

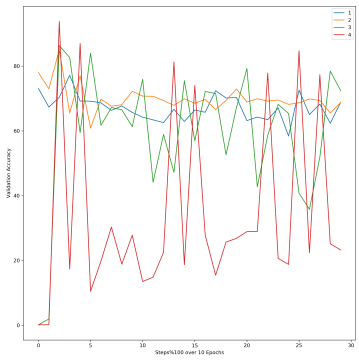

Figure 9: N-Gram Validation Accuracy Training Curves

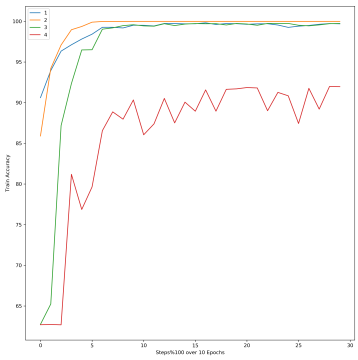

Figure 10: N-Gram Training Accuracy Training Curves

Next we'll look at the various tokenizers we could use. We tried the medium and small spacy tokenizers as well as the word tokenizer from NLTK. We find all three perform roughly similarly but nltk edges out the spacy tokenizers slightly on validation accuracy (Figure 11) so we will move forward with the NLTK tokenizer.

Our last hyperparameter to tune is the optimizer we choose to use. We saw a strange curve on our validation accuracy plots with both the SGD and Adam optimizers performing similarly but with the SGD pulling ahead slightly (Figure 13). With that said, the training accuracy did not move at all with the SGD optimizer (Figure 14). It seems the Adam optimizer does a better job fitting our data but a marginally worse job generalizing. Since the validation accuracies are within a percent of eachother we will move forward with the Adam optimizer.

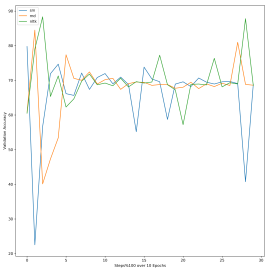

Figure 11: Tokenizer Validation Accuracy Training Curves

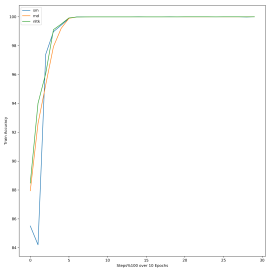

Figure 12: Tokenizer Training Accuracy Training Curves

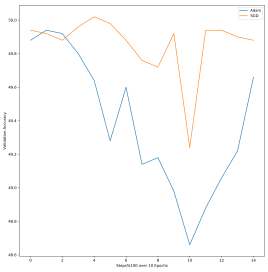

Figure 13: Optimizer Validation Accuracy Training Curves

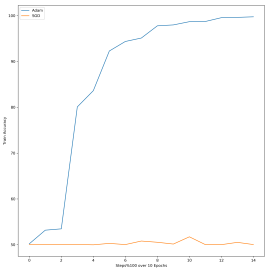

Figure 14: Optimizer Training Accuracy Training Curves

3 Final Model

After performing hyperparameter tuning we finalized our model to use an NLTK tokenizer on bigrams, an 80,000 vocabulary size, an embedding dimension of 1000, an Adam optimizer and a non-annealing learning rate of 0.01. Using these hyperparameters we were able to achieve a test accuracy of 81.78%.

Now let's take a look at 6 positive reviews- three of which are correctly classified as such while 3 are incorrectly classified as such.

First correct classification: I saw his film at the Ann Arbor Film Festival. I am a film student at the Univeristy of Michigan so I know a thing or two about film. And Crispin Glover's film is outrageous. He basically exploits the mentally challenged. Not only is Shirly Temple the anti-Christ (which I admit is a little funny) telling the mentally challenged to kill each other, but there is an obsession with killing snails. Crispin also plays with the idea of being in love with one of his actors who is as they all are, mentally challenged. PETA and Human Rights should be all over this thing. It's not 'counter-culture' as Crispin stated at the Ann Arbor Film Festival, it's exploitation.

Second correct classification: Bettie Page was a icon of the repressed 1950s, when she represented the sexual freedom that was still a decade away, but high in the hopes and dreams of many teenagers and young adults. Gretchen Mol does a superb job of portraying the scandalous Bettie, who was a small town girl with acting ambitions and a great body. Her acting career went nowhere, but her body brought her to the peak of fame in an admittedly fringe field. Photogrsphed in black and white with color interludes when she gets out of the world of exploitation in New York, this made-for-TV (HBO) film has good production values and a very believable supporting cast. The problem is, it's emotionally rather flat. It's difficult to form an attachment to the character, since Bettie is portrayed as someone quite shallow and naive given the business she was in. The self-serving government investigations are given a lot of screen time, which slows down the film towards the end. But it's definitely worth watching for the history of the time, and to see the heavy-handed government repression that was a characteristic of the fifties. 7/10.

Third correct classification: I've read up a little bit on Che before watching this film and you wanna know something, he was a real hero for the people because he only wanted to see equality for everyone and that he hated what the oppressive forces were doing to his people as well as all other Latin Americans in general! Now, I don't know about others, but to me he did the right thing by wanting socialism so that everyone had to pay their fair share. However, the powerful elite obviously weren't going to go for that. So, rather than understanding what Che Guevera wanted, they were forced to kill him in attempting to suppress the revolution. It didn't work since there were too many of his other followers who only picked up where he left off. A good example of this was

when Castro continued his leadership in Cuba. As far as I'm concerned and as Che said it himself right before he died: "If you kill me, that's fine. But you're only killing a man, you'll NEVER kill the cause!" I couldn't have said it any better myself.¡br /¿¡br /¿But ... ANYWAYS.... that's why I give this film a 7 out of 10.

First incorrect classification: Ten out of ten stars is no exaggeration. This documentary provides the viewers with unique footage about the 2003 coup in Venezuela. This great film is now the minimum knowledge requirement if you want to express a competent opinion about Venezuela or Hugo Chavez.¡br /¿¡br /¿The dramatic, electrified atmosphere, the unique footage will allow you to experience a true historic moment. You'll feel like you're in the middle of the situation. ¡br /¿¡br /¿The film will help you gain unique insight in the happenings of 2003 and will help you hear a side you will rarely hear on TV. It's something you shouldn't miss.

Second incorrect classification: The material in this documentary is so powerful that it brought me to tears. Yes, tears I tell you. This popular struggle of a traditionally exploited population should inspire all of us to stand up for our rights, put forth the greater good of the community and stop making up cowardly excuses for not challenging the establishment. Chavez represents the weak and misfortunate in the same way Bush is the face of dirty corporations and capitalism ran amok. Indeed, Latin America is being reshaped and the marginalized majority is finally having a voice in over five centuries. Though, in the case of Mexico, the election was clearly stolen by Calderon. Chavez is not perfect, far from it. He's trying to change the constitution to allow him to rule indefinitely. That cannot be tolerated. Enough with the politics and back to the movie; The pace is breath taking at moments, and deeply philosophical at others. It portrays Chavez as a popular hero unafraid to challenge the US hegemony and domination of the world's resources. If you think the author is biased in favour of Chavez, nothing's stopping you from doing your homework. One crucial message of the film is questioning info sources, as was clearly demonstrated by the snippers casualties being shamefully blamed on Chavez's supporters. Venezuela puts American alleged democracy to shame. Hasta la revolucion siempre.

Third incorrect classification: This is something new.¡br /¿¡br /¿There's a coup d'etat and a couple of irish documentary filmmakers are right inside of it.¡br /¿¡br /¿A democratically elected president who uses his power to bring literacy to his people and encourages them to read the constitution is being slandered by the private media openly as dictator, mentally unstable, new hitler, etc. without repercussion from the governments side (like, say, silencing them via bullets and other traditional dictatorial methods). Oh, and they still claim that they are being suppressed, of course.¡br /¿¡br /¿See how the media gloats about their own role in the coup d'etat on TV after they toppled the government with the help of rouge generals (how much more stupid can you get??).¡br /¿¡br /¿And see how the people of Venezuela march to the palace, holding the constitution in their hands, and reinstall their elected government.¡br /¿¡br /¿This sounds like a Hollywood fairytale, but it happened for real, against the explicit wishes of the USA. The documentary is a historical masterpiece, shot from the center of the action, acute and totally embarrassing for the prime supporters of the coup: The good, democratic, freedom loving, benevolent USA (who still channel large amounts of money to Chavez' political opponents).¡br /¿¡br /¿Also highly entertaining and exciting. 10 points.