Weekly Homework 4

Alex Ring

1 Part A

1. Does this model accurately translate phrases longer than a single word? Explain.

It can translate phrases longer than a single word, but accuracy depends on how what you would consider accurate. The model can translate whole sentences, but instead of having an exact translation, most of the time you are left with something that makes sense, and conveys the same message, but is not an exact translation. Because the model assumes that words are independent of one another it struggles with certain phrases being tied to single words between languages. It also struggles with sentence structure. That being said, if you consider the message being maintained as accurate, the model functions about 48% of the time. They considered things like: ”Yet it is very simple” -¿ ”It is still very simple” to be a success, because the original meaning of the message is preserved.

2. How is the Markov assumption being used in this paper?

The Markov Assumption assumes that the future and present state depend only on the present state, rather than the series of events that lead up to it. In the language translation model used here, they talk about how assuming the state of a word given all previous words in the sentence would be impossible because there are just too many probabilities. Instead they put sentences into equivalence classes and use an n-gram model. The n-grams agree if the histories end in the same word. In this way they assume a kind of word independence similar to that described in the Markov Assumption.

3. They suggest that a trigram model would improve the performance of their system. Why would this be?

They mention that the current bigram model is not very good at recognizing things like phrases that get translated as a single unit. They acknowledge that these phrases stay together even if they move around in sentences. As an example with their current model, ”aller” which is ”to go” in English, is being translated from just ”go” and not ”to”. The current model assumes that words are independent of one another, which is not the case a lot of the time. This new trigram model aims to solve these issues by focusing on the positions of the target words produced by a particular source word depending on the identity of the source word and on the positions of the target words produced by the previous source word. In other terms, this new trigram model will take the words around it into account, instead of assuming word independence.

4. What is the ”generative story” of the translation model? That is, how does the model suppose that translations are generated, and how does this relate to the noisy channel model?

The translation model uses a system of multiplying probabilities of words, their direct translations, and the fertility of those translations. It attempts to map the words in an English sentence to one in French directly. It considers how words in a French sentence could be shuffled around from the English sentence in much the same way that the noisy channel model considers a mispelled word as being shuffled around from a correctly spelled word. In much the same way the model also considers missing words in the French translation such as ”does” in ”John does”, just like the noisy channel model considers missing letters such as ”better” and ”beter”.

5. What are the prior and posterior in the first equation, and what do they signify?

The prior, Pr(T|S), is the probability that a translator will produce T in the target language when presented with S in the source language. The posterior, P(S|T), is the probability sentence S in the source language, is the translation of sentence T from the translator. Basically the posterior is the probability that the translation is correct.

6. What is the formula that describes the process of selecting the ”best sentence” in section 4?

P(T|S) = Pr(n|e)Pr(f|e)Pr(i|j,l)

7. What are the parameters in the translation model, and how are they estimated?

The translation model uses a set of fertility probabilities Pr(n|e), a set of translation probabilities Pr(f|e) and a set of distortion probabilities Pr(i|j, l).

Fertility is the number of French words that an English word produces in a given alignment. They define an alignment as indicating the origin in the English sentence of each of the words in the French sentence. John is aligned with Jean. The probabilities are calculated by multiplying the probability of each fertility of each word in the sentence. Pr(n|e) for each fertility n from 0 to some moderate limit.

Distortion is a measurement of how far the words in an English sentence are from the words in a French sentence. Like adjectives coming before words in English and after words in

French. Distoriton is calculated using the length of the target sentence and the position of the source word. Pr(i|j, l) where i is a target position, j a source position, and l the target length.

The translation probabilities Pr(f|e), one for each element f of the French vocabulary and each member of the English vocabulary.

Translation is just a measurement of how likely a word **8. What is a "distortion?"**

Distortion refers to how in different languages words do not always map to one another in the sentence structure. The example they give in the paper is how, in English, adjectives precede nouns they modify, but in French they follow them. They refer to the effect of words moving around in the sentence structure between languages as "distortion" and account for it in the model.

9. What is an "alignment?"

An alignment indicates the origin in the source language sentence of each of the words in the translation target language sentences. It is a mapping system. In the paper they give the example of "John" in English being aligned with "Jean" in French. A word in the source language is aligned with the word that it produces in the target language.

2 Part B

1. Graphically, what effect does the bias term have on the logistic function in logistic regression?

The bias parameter in logistic regression is a scalar parameter that will shift the decision boundary by a constant amount. Graphically this means that the line that makes up the decision boundary is just being scaled up and down.

2. Describe the difference(s) between stochastic gradient descent and gradient descent.

Both of these are methods for updating a set of parameters through iteration. The main difference lies in how much of the training data is used in each pass to do the update. Gradient descent uses all of the training data each pass to update the parameters. While Stochastic Gradient Descent uses only one training example each pass to update the parameters. SGD converges faster than GD but is less minimized to error than normal GD. It is a tradeoff between time and correctness. With large datasets normal GD is going to take too much time. Most of the time SGD is good enough for prediction, even if the error function is less minimized than normal GD.

3. Given a convex problem surface, what do we use to find the direction of steepest ascent?

To find the direction of steepest ascent we use the gradient. The gradient is actually equal to the direction of the steepest ascent. It is simply a vector that points in that direction.

4. Which beta terms can be skipped during the update stage of logistic regression with SGD?

When updating the beta terms in the update stage of logistic regression you can skip updating the bias term.

5. What are the elements of the gradient vector?

The gradient vector is a vector that represents the direction of maximum change of a parameter. It is a vector that tells us how far we are going to move from data point to another on the gradient. In a 3D plane it could be represented as $\triangledown F = (F_x, F_y, F_z)$