Week 4 Homework

Samuel Cuthbertson SAMUEL.CUTHBERTSON@COLORADO.EDU

1. Questions About the Paper

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

Yes. According to the authors, they only 'exactly' translated sentences 5% of the time, but were able to translate and preserve meaning 43% of the time. I'd define this as accurate for 'phrases' because of how their translator works exponentially better for shorter sentences.

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

The Markov Assumption is being used when they rewrite $P(s_1 s_2 s_3 ... s_n)$ in section two. Using and n-gram model, the above is equal to $P(s_1)P(s_2|s_1)P(s_3|s_2, s_1)...P(s_n|s_1, s_2, ...s_{n-1})$. The authors use a bigram model, however, which is like rewriting

$$P(s_1 s_2 s_3 ... s_n) = P(s_2|s_1, Begin)P(s_3|s_2, s_1)...P(s_n|s_{n-2}, s_{n-1})$$

This uses the Markov Assumption in assuming that each word only has a relational history to the previous two words, not to any words before those two, and not to any words after the word in question.

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

A trigram model would extend that history mentioned in the previous part up to the previous three words, which more accurately models the dependent nature of language.

1.4. 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?

This model supposes that each and every input sentence T was generated from an English sentence S, and that studying other T and S pairs will enable them to put forth an S for any given T. This can be thought of in terms of the noisy channel model as there once was a sentence S, which was then put through a noisy channel and was transformed into T.

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

The prior is $P(S)$, the probability that sentence S exists. The posterior is $P(S|T)$, the probability that sentence S is the translation of sentence T.

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

The formula for finding the "best sentence" uses a stack search, computing the probabilities of first the first word, and a partial alignment translation of that word, and then proceeding down the sentence until $P(S|T)$ is at a maximum.

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

The parameters are the probabilities of an English word or words given a French word, and are generated using a method similar to an EM Algorithm. Simply, given an initial estimation of the parameters, a better estimation can be obtained by ranking all possible alignments of the parameters according to their probabilities relative to the initial estimate.

1.8. What is a "distortion?"

When, for example, on French word in the middle of a sentence turns into two English words at the beginning and ending of the translated sentence.

1.9. What is an "alignment?"

When a French word directly translate into one English word. See Figure 3 in the paper.

2. Miscellaneous Questions

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

The bias term sets at what input the logistic function will return $\frac{1}{2}$.

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

Stochastic gradient descent uses a new random data point for every iteration of updating the β terms, while gradient descent uses all the data points at every stage of updating the β terms.

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

We find the direction in the gradient which is most positive.

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

When a given example does not contain a feature, then the β value for that feature can be skipped.

2.5. What are the elements of the gradient vector?

All the partial derivatives of every dimension in our logistic function, which