Week 6 Homework

Samuel Cuthbertson SAMUEL.CUTHBERTSON@COLORADO.EDU

1. Answers to Part A

1.1. One-vs-all classification decides which label to predict by using the most confident classifiers to vote on the correct answer.

True. One-vs-all does exactly that, in seeing which classifier is most probably correct (most confident) and then returning that classifier's answer.

1.2. Error correcting code classification constructs multiple binary classifiers for choosing a label.

True. It creates N binary classifiers where N is the number of features, and uses the output from all of those classifiers to find which class is "closest" (hamming distance) to the provided input.

1.3. Standard support vector machines produce probabilities, just like logistic regression and the perceptron.

False. SVMs simply return the predicted class, and have no probability associated with it. Something like the probability/confidence can be derived from finding how close the input vector is to the separator, but this is not a direct relationship.

1.4. Logistic regression can be viewed as a special case of the perception.

True. Logistic regression can be viewed as a single-layer perceptron network.

2. Answers to Part B

2.1. What is the objective of this paper?

To predict the final verb in a sentence before the sentence is fully interpreted, in order to enable better real time translation.

2.2. What kind of learning algorithm does this paper use for its computational results?

Logistic Regression.

2.3. What is a "baseline," and how is it used in this paper?

Here, a baseline is the accuracy given by simply choosing the most common class for every problem.

2.4. What multi-class classification scheme does the paper use for 50 verbs?

One-vs-All Classification.

2.5. What multi-class classification scheme does the paper use for the multiple choice scenario? Describe how this works.

It encodes the class into the feature vector, and uses only one binary classifier to classify new information. From what I understand, this works exactly like the classifier that was described (but never named) in lecture.

2.6. How does the computational performance on the multiple choice data compare with the human performance?

The accuracy compares as 90% for machine accuracy, and 81.1% for humans. This makes me wonder at how the model was trained to expect a certain *dialect* from the training questions, and then learned how to "test well", and might not be very robust. In other words, this sounds like the model is over fit.

2.7. What are the feature set and hypothesis space, respectively?

The feature set is both **context** and **verb** features, as described in section 3.1.1. The hypothesis space is all possible verb answers.

2.8. Why does the paper suggest that the baseline model doesn't work well?

It cannot reliably predict more than a handful of verbs.

2.9. What does the paper mean in Figure 3, where it describes the distribution as "Zipfian?"

It means that the distribution is highly concentrated: the most common item appears roughly twice as often as the next most common item, and so on.