LING 572 Hw5 (MaxEnt decoder)
Due: 11pm on Feb 9, 2021

The example files are under /dropbox/20-21/572/hw5/examples/.

Q1 (5 points): Run the Mallet MaxEnt learner (i.e., the trainer's name is MaxEnt) with **train2.vectors.txt** as the training data and **test2.vectors.txt** as the test data.

- You can use *vectors2classify* or "*mallet train-classifier*" plus "*mallet classify-svmlight*".
- Save the model to a file called *q1/m1*.
- Convert the model into the text format with the following command: classifier2info --classifier q1/m1 > q1/m1.txt
- In your note file, write down the command you used, the training accuracy and the test accuracy.

Q2 (25 points): Write a MaxEnt classifier, called **maxent_classify.sh**, that classifies test data given a MaxEnt model learned from training data.

- The format is: maxent_classify.sh test_data model_file sys_output > acc_file
- test_data, sys_output, and acc_file have the same format as in Hw2-Hw4, except that sys_output and acc_file contain only the results for the TEST data, not the training data (since the training data is not available to the classifier).
- model_file has the same format as q1/m1.txt created in Q1.
- Run "maxent_classify.sh test2.vectors.txt q1/m1.txt q2/res > q2/acc". What is the test accuracy? Is it the same as the test accuracy in Q1?

Q3 (15 points): Write a script, calc_emp_exp.sh, to calculate empiricial expectation.

- The format is: calc_emp_exp.sh training_data output_file
- training_data has the same format as before.
- output_file has the format "class_label feat_name expectation raw_count" (c.f. emp_count_ex): *raw_count* is the number of training instances with that class_label and contains that feat_name; *expectation* is the empirical expectation.
- Run "calc_emp_exp.sh train2.vectors.txt q3/emp_count" and include q3/emp_count in your submission.

Q4 (30 points): Write a script, calc_model_exp.sh, to calculate model expectation.

- The format is: calc_model_exp.sh training_data output_file {model_file}
- training_data has the same format as before.
- output_file has the format "class_label feat_name expectation count" (e.g., **emp_count_ex**): *expectation* is the model expectation; *count* is *expectation* multiplied by the number of training instances. Note that the *count* is often a real number, not an integer, so outputing it as a real number.
- model_file is optional. If it is given, it has the same format as in Q2 (e.g., q1/m1.txt) and it is used to calculate $p(y|x_i)$. If it is not given, $p(y|x_i) = 1/|C|$, where $|C|$ is the number of class labels.
- Run "calc_model_exp.sh train2.vectors.txt q4/model_count q1/m1.txt" and include q4/model_count in your submission.
- Run "calc_model_exp.sh train2.vectors.txt q4/model_count2" and include q4/model_count2 in your submission.

Submission: Submit the following to Canvas:

- Your note file *readme.(txt | pdf)* that includes your answers to Q1 and Q2 and any notes that you want the TA to read.
- hw.tar.gz that includes all the files specified in dropbox/20-21/572/hw5/submit-file-list, plus any source code (and binary code) used by the shell scripts.
- Make sure that you run **check_hw5.sh** before submitting your hw.tar.gz.