LING 572 Hw1
Due: 11pm on Jan 16, 2019

Q1 (25 points): Let X and Y be two random variables. The values for P(X,Y) are shown in Table 1, H(X) is the entropy of X, and MI(X,Y) is the mutual information of X and Y. Please write down the formulas and the results for the following.

(a) 1 pt: $P(X)$

(b) 1 pt: $P(Y)$

(c) 2 pt: $P(X \mid Y)$

(d) 2 pt: $P(Y \mid X)$

(e) 2 pts: Are X and Y independent? Why or why not?

(f) 2 pts: $H(X)$

(g) 2 pts: $H(Y)$

(h) 2 pts: $H(X, Y)$

(i) 2 pts: $H(X \mid Y)$

(j) 2 pts: $H(Y \mid X)$

(k) 2 pts: $MI(X, Y)$

(l) 5 pts: The value for $Q(X, Y)$ are shown in Table 2. What is the value for $KL(P(X, Y) \mid\mid Q(X, Y))$? What is the value for $KL(Q(X, Y) \mid\mid P(X, Y))$? Are they the same?

Table 1: The joint probability P(X,Y)

	X=1	X=2	X=3
Y=a	0.10	0.20	0.30
Y=b	0.05	0.15	0.20

Table 2: The joint probability Q(X,Y)

	X=1	X=2	X=3
Y=a	0.10	0.20	0.40
Y=b	0.01	0.09	0.20

Q2 (10 points): Let X be a random variable for the result of tossing a coin. $P(X = h) = p$; that is, p is the possibility of getting a head, and $1 - p$ is the possibility of getting a tail.

(a) 1 pt: $H(X)$ is the entropy of X. Write down the formula for $H(X)$.

(b) 2 pts: Let $p^* = arg\ max_p\ H(X)$; that is, p^* is the p that results in the maximal value of $H(X)$. What is p^*?

(c) 7 pts: Prove that the answer you give in (b) is correct. Hint: recall how you calculate the optimal solution for a function $f(x)$ in your calculus class. In this case, $H(X)$ is a function of p.

Q3 (25 points): Permutations and combinations:

(a) 6 pts: The class has n students, and n is an even number. The students are forming teams to work on their homework. Each team has exactly 2 students and each student has to appear in exactly one team. How many distinct ways are there to form the teams for the class? Write down the formula. Hint: when n=4, there are 3 ways. For instance, if student #1 and #2 are in the same team, students #3 and #4 would have to be in the same team too.

(b) 5 pts: There are 10 balls: 5 are red, 3 are blue, and 2 are white. Suppose you put the balls in a line, how many different color sequences are there?

(c) 14 pts: Suppose you want to create a document of length N by using only the words in a vocabulary $\Sigma = \{w_1, w_2, ..., w_n\}$. Let $[t_1, t_2, ..., t_n]$ be a list of non-negative integers such that $\sum_i t_i = N$.

(c1) 7 points: How many different documents are there which satisfy the condition that, for each w_i in the vocabulary Σ, the occurrence of the word w_i in the document is exactly t_i? That is, how many different word sequences are there which contain exactly t_i w_i's for each w_i in Σ?

Hint: The answer to (c1) is very similar to the answer to (b).

(c2) 7 pts: Let P(X) be a unigram model on the vocabulary Σ; that is, $P(X = w_i)$ is the probability of a word w_i, and $\sum_{w_i \in \Sigma} P(X = w_i) = 1$.

Suppose a document of length N is created with the following procedure: for each position in a document, you pick a word from the vocabulary according to P(X); that is, the probability of picking w_i is $P(X = w_i)$. What is the probability that you will end up with a document where the occurrence of the word w_i (for each $w_i \in \Sigma$) in the document is exactly t_i?

Hint: As (c1) shows, there will be many documents that contain exactly t_i w_i's. The answer to (c2) should be the sum of the probabilities of all these documents.

Q4 (10 points): Suppose you want to build a trigram POS tagger. Let T be the size of the tagset and V be the size of the vocabulary.

(a) 2 pts: Write down the formula for calculating $P(w_1, ..., w_n, t_1, ..., t_n)$, where w_i is the i-th word in a sentence, and t_i is the POS tag for w_i.

(b) 8 pts: Suppose you will use an HMM package to implement a trigram POS tagger.

- What does each state in HMM correspond to? How many states are there?

- What probabilities in the formula for (a) do transition probability a_{ij} and emission probability b_{jk} correspond to? $a_{i,j}$ is the transition probability from state s_i to s_j, and b_{jk} is the probability that State s_j emits symbol o_k.

Q5 (10 points): In a POS tagging task, let V be the size of the vocabulary (i.e., the number of words), and T be the size of the tagset. Suppose we want to build a classifier that predicts the tag of the current word by using the following features:

1. Previous word w_{-1}
2. Current word w_0
3. Next word w_{+1}
4. Surrounding words w_{-1} w_{+1}
5. Previous tag t_{-1}
6. Previous two tags t_{-2} t_{-1}

(a) 3 pts: How many unique features are there **in total**? You just need to give the answer in the Big-O notation (e.g., $O(V^3)$).

(b) 2 pts: A classifier predicts class label y given the input x. In this task, what is x? what is y?

(c) 5 pts: For the sentence **Mike/NN likes/VBP cats/NNS**, write down the feature vector for each word in the sentence. The feature vector has the format "InstanceName classLabel featName1 val1 featName2 val2 ....". For the instanceName, just use the current word.

Q6 (10 points): Suppose you want to build a language identifier (LangID) that determines the language code of a given document. The training data is a set of documents with the language code for each document specified. The test data is a set of documents, and your LangID needs to determine the language code of each document.

(a) 7 pts: How do you plan to build the LangID system? For instance, if you want to treat this as a classification problem, what would x (the input) be? what would y (the output) be? What would be good features? Name at least five types of features (e.g., one feature type is the word unigrams in the document).

(b) 3 pts: What factors (e.g., the amount of training data) could affect the system performance? Name at least three factors, excluding the amount of training data.

Q7 (10 "free" points): If you are not familiar with Mallet, please go over the Mallet slides at the course website[1]

Set up the package in your patas environment, run some experiments. We will use Mallet in later assignments. If you do not have a patas account, you should contact me right away.

[1] .The third link at `https://canvas.uw.edu/courses/1257221/pages/slides-from-prior-ling570`

Submission: In your submission, include the following:

- readme.(txt|pdf) that includes your answers to Q1-Q6. No need to submit anything for Q7.
- Since this assignment does not require programming, there is no need to submit hw.tar.gz, and no need to run check_hwX.sh script.