LING 572 Hw1
Due: 11pm on Jan 11, 2022

Q1 (25 points): Let X and Y be two random variables.

(a)

$P(X = 1) = 0.1 + 0.05 = 0.15$
$P(X = 2) = 0.2 + 0.15 = 0.35$
$P(X = 3) = 0.3 + 0.2 = 0.5$

(b)

$P(Y = a) = 0.1 + 0.2 + 0.3 = 0.6$
$P(Y = b) = 0.05 + 0.15 + 0.2 = 0.4$

(c)

$P(X = 1 \mid Y = a) = \frac{0.1}{0.6} = \frac{1}{6}$
$P(X = 1 \mid Y = b) = \frac{1}{8}$
$P(X = 2 \mid Y = a) = \frac{1}{3}$
$P(X = 2 \mid Y = b) = \frac{3}{8}$
$P(X = 3 \mid Y = a) = \frac{1}{2}$
$P(X = 3 \mid Y = b) = \frac{1}{2}$

(d)

$P(Y = a \mid X = 1) = \frac{0.1}{0.15} = \frac{2}{3}$
$P(Y = a \mid X = 2) = \frac{4}{7}$
$P(Y = a \mid X = 3) = \frac{3}{5}$
$P(Y = b \mid X = 1) = \frac{1}{3}$
$P(Y = b \mid X = 2) = \frac{3}{7}$
$P(Y = b \mid X = 1) = \frac{2}{5}$

(e) X and Y are not independent. $P(X = 1 \mid Y = a) \neq P(X = 1)$

(f) $H(X) = -\sum p(x)logp(x) = -((0.15)log(0.15) + (0.35)log(0.35) + (0.5)log(0.5) = 1.4406$

(g) $H(Y) = -\sum p(y)logp(y) = 0.9710$

(h) $H(X, Y) = -\sum_x \sum_y p(x, y)logp(x, y) = 2.4087$

(i) $H(X \mid Y) = H(X, Y) - H(Y) = 1.4377$

(j) $H(Y \mid X) = H(X, Y) - H(X) = 0.9680$

(k) $MI(X, Y) = \sum_x \sum_y p(x, y)log\frac{p(x,y)}{p(x)p(y)} = 0.0029$

(l)

$KL(P(X, Y) \mid\mid Q(X, Y)) = \sum_x \sum_y p(x, y)log\frac{p(x,y)}{q(x,y)} = 0.1021$
$KL(Q(X, Y) \mid\mid P(X, Y)) = \sum_x \sum_y q(x, y)log\frac{q(x,y)}{p(x,y)} = 0.0765$
Not the same.

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) $H(X) = -p * log_2(p) - (1 - p) * log_2(1 - p)$.

(b) $p^* = 0.5$

(c) $H'(X) = log_2(1 - p^*) - log_2(p^*) = log_2((1 - 0.5)/0.5) = 0$

Q3 (25 points): Permutations and combinations:

(a) $\Pi_{i=3}^{n/2}(2i - 1)$

(b) $\frac{10!}{5!*3!*2!} = 2520$

(c) **(c1)** $\frac{N!}{\Pi_i^n(t_i!)}$

(c2) $\frac{N!}{\Pi_i^n(t_i!)} * \Pi_i P(X = w_i)^{t_i}$

Q4

(4a) $\Pi_{i=1} P(w_i \mid t_i) P(t_i \mid t_{i-2}, t_{i-1})$

(4b) Each state corresponds to a tag. There are T^2 states in the trigram model. Transition probability corresponds to $P(t_i \mid t_{i-2}, t_{i-1})$, and emission probability to corresponds to $P(w_i \mid t_i)$.

Q5

(a) $O(V^2 + T^2)$

(b) x is the current word. y is the POS tag for x.

(c) Mike NN PrevWord /s CurrentWord Mike NextWord likes SurroundWords /s_likes PrevTag BOS PrevTwoTag BOS_BOS
likes VBP PrevWord Mike CurrentWord likes NextWord cats SurroundWords Mike_cats PrevTag NN PrevTwoTag BOS_NN
cats NNS PrevWord likes CurrentWord cats NextWord /s¿ SurroundWords likes_/s¿ PrevTag VBP PrevTwoTag NN_VBP

Q6

(a) The input would be a document in the native orthography of the language with POS taggers. The output would be two groups of features, one alphabetic, and the other one morphosyntactic. Some good features that could be useful would be up to thirty-five alphabets/characters in the language. For languages with unique orthographies, it would be easy to identify through their writing system. For languages with a large number of common characters (i.e. Indo-European languages), up to thirty-five alphabets should help identify the language. One more

feature to add to the alphabets is special punctuation. Morphosyntactic features include word features (unigrams, bigrams, trigrams for Roman alphabets and characters for character-based languages). Other features include whether there exists double negation in one sentence, whether the language is VOS/SOV/SVO..., or whether there is a case system.

(b) Literature type of documents (some forms of literature such as poems might not follow the standard grammar); Orthography (some languages might have more than one digital writing system); Bias in training data (the training data is likely to be well-documented languages)