# LING 572 Hw1
# Author: Dhwani Serai (dserai)

**Q1** The values for P(X,Y) and Q(X,Y) are shown in Table 1 and 2 respectively

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 |

**(a):** $P(X)$

$P(X) = \sum_{Y=a}^{Y=b} P(X,Y)$

Using marginal probability,

Table 3: probability P(X)

| X=1 | X=2 | X=3 |
|---|---|---|
| 0.15 | 0.35 | 0.5 |

Also note that $\sum P(X) = 1$

**(b):** $P(Y)$

$P(Y) = \sum_{X=1}^{X=3} P(X,Y)$

table for P(Y)

Also note that $\sum P(Y) = 1$

Table 4: probability P(Y)

| | |
|---|---|
| Y=a | 0.6 |
| Y=b | 0.4 |

**(c):** $P(X \mid Y)$
By Baye's theorem we know that $P(X \mid Y) = P(X,Y)/P(Y)$

Table 5: The conditional probability $P(X \mid Y)$

| | X=1 | X=2 | X=3 |
|---|---|---|---|
| Y=a | 0.167 | 0.333 | 0.5 |
| Y=b | 0.125 | 0.375 | 0.5 |

**(d):** $P(Y \mid X)$
By Baye's theorem we know that $P(Y \mid X) = P(X,Y)/P(X)$

Table 6: The conditional probability $P(Y \mid X)$

| | X=1 | X=2 | X=3 |
|---|---|---|---|
| Y=a | 0.667 | 0.571 | 0.6 |
| Y=b | 0.333 | 0.429 | 0.4 |

**(e):** Are X and Y independent? Why or why not?
Ans. Since we can see that $P(X \mid Y)$ is not equal to P(X) and $P(Y \mid X)$ is not equal to P(Y) we can say that X and Y are not independent

**(f):** $H(X)$
$H(X) = -\sum p(x) log p(x)$ = 0.15*log(0.15)+0.35*log(0.35)+0.5*log(0.5)
H(X) = 1.441

**(g):** $H(Y)$
$H(Y) = -\sum p(y) log p(y) = 0.971$

**(h):** $H(X,Y)$
$H(X,Y) = -\sum_x \sum_y p(x,y) log p(x,y) = 2.409$

**(i):** $H(X \mid Y)$
$H(X \mid Y) = H(X,Y) - H(Y)$
$H(X \mid Y) = 1.438$

**(j):** $H(Y \mid X)$
$H(Y \mid X) = H(X,Y) - H(X)$
$H(X \mid Y) = 0.968$

**(k):** $MI(X,Y)$
$MI(X,Y) = H(X) + H(Y)$ - $H(X,Y)$
$MI(X,Y) = 1.441 + 0.971$ - $2.409 = 0.003$

**(l):** 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?

Ans.
$H_c(X,Y) = -\sum p(x,y) log q(x,y)$ = 0.10*log(0.10)+0.20*log(0.20)+0.3*log(0.4)
0.05*log(0.01)+0.15*log(0.09)+0.2*log(0.2) =2.5109
$KL(P(X,Y) \mid\mid Q(X,Y))$ =
$H_c(X,Y) - H(X,Y)$ = 0.1019
$KL(P(X,Y) \mid\mid Q(X,Y))$ =
$H(X,Y) - H_c(X,Y)$ = -0.1019
Therefore, $KL(P(X,Y) \mid\mid Q(X,Y))$ and $KL(Q(X,Y) \mid\mid P(X,Y))$ are not the same

**Q2:** 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)** formula for $H(X)$.
H(X) = -$\sum_x p(x) log p(x)$ = -plog p - (1-p) log(1-p)

**(b)** Let $p^* = arg\ max_p\ H(X)$; that is, What is $p^*$?
maximal value of H(x) is highest uncertainty
if coin is fair uncertainty is maximized i.e. p=0.5
Therefore, $p^* = arg\ max_p\ H(X)$ when p=0.5
H(X)=-log 0.5 = 1

**(c)** Prove that the answer you give in (b) is correct.

From part a) We know, that H(X) = -plog p - (1-p)log(1-p),
To find the maximal value we need to find the value of p when the derivative of H(X) is 0.
$d(H(X)/dp$ = -(log (ep) - log(1-p) - log (e) )=0
i.e. $-log(ep/e(1-p))$ =0
i.e. p/(1-p) = 1
Therefore, p= 0.5

**Q3 (25 points):** Permutations and combinations:

**(a)** The class has n students to be divided in teams of 2

Ans. for dividing n students into teams of 2 we get $({}^nC_2 + {}^{n-2}C_2 + ....{}^4C_2 + {}^2C_2) * 1/(n/2)!$
Simplifying this form we get,

$$\frac{n!}{2^{n/2}(n/2)!}$$

**(b)** 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?

number of different color sequences in this case will be

$$\frac{10!}{5!3!2!}$$

which is equal to 2520 sequences

**(c):** 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 $\sum_i t_i = N$. For instance, suppose the vocabulary $\Sigma$ is ["a", "cat", "chases", "dog", "sheep", "eat"], and the document is "a cat chases a dog", then the document length $N = 5$, the vocabulary size $n = 6$, and $[t_1, t_2, t_3, t_4, t_5, t_6]$=[2, 1, 1, 1, 0, 0].

**(c1):** Ans. vocab size n=6 document length N=5 and [t1, t2, t3, t4, t5, t6]=[2, 1, 1, 1, 0, 0]
Therefore,

$$\frac{5!}{2!1!1!1!}$$

which is equal to 60

**(c2)** Ans.

$$\frac{N!}{\prod_{i=1}^{n} t_i!} * \prod_{i=1}^{n} P(w_i)^{t_i}$$

Here, n= number of words in vocabulary
N=length of document
P is probability of a word

**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.

**(4a) 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 $w_i$.

Ans. $P(w_1, ..., w_n, t_1, ..., t_n) = \prod_{i=1}^{n} q(t_i \mid t_{i-1}, t_{i-2}) \prod_{i=1}^{n} e(w_i \mid t_i)$
Here q is the transition probability for a tag i given that the previous tags are $t_{i-1}$ and $t_{i-2}$
e is the emission probability for the word being $w_i$ given that tag is $t_i$

**(4b)** HMM to implement a trigram POS tagger.

- each state in HMM? How many states?

  Ans. Each state in the trigram HMM corresponds to a tag pair. The number of states in the HMM is equal to the number of tag pairs which will be equal to $T^2$ if T is the number of tags in the tagset(vocab of tags)

- What probabilities in the formula for Q4(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 $o_k$.

  Ans. $a_{ij} = P(s_j \mid s_i)$ where s is a state in an HMM;
  Let $s_i$=(t1,t2) and $s_j$=(t2,t3); therefore $a_{ij} = P(t3 \mid t, t2)$

  Similarly, $b_{jk} = P(o_k, s_j) = P(o_k, t3)$

  $a_{ij}$ corresponds to $q(t_i \mid t_{i-1}, t_{i-2})$ in 4a)
  and $b_{ij}$ corresponds to $e(w_i \mid t_i$ in 4a)

**Q5**

**(a)** $O(VT + T^2)$.

**(b)** A classifier predicts class label y given the input x.
Ans. In this task, y is the word x is the tag

**(c)** **Mike/NN likes/VBP cats/NNS**, write down the feature vector for each word in the sentence.
The feature vector has the format Mike NN $w^{-1}$ ¡s¿ $w^{+1}$ likes $w^{-1}w^{+1}$ BOS likes $t_{-1}$ BOS $t_{-2}t_{-1}$ BOS BOS.
likes VBP $w^{-1}$ Mike $w^{+1}$ cats $w^{-1}w^{+1}$ Mike$_c ats t_{-1}$ NN $t_{-2}t_{-1}$ BOS NN
cats NNS $w^{-1}$ likes $w^{+1}$ EOS $w^{-1}w^{+1}$ likes EOS $t_{-1}$ VBP $t_{-2}t_{-1}$ NN VBP

**Q6** Ans.

**(a)** We can frame the language identifier as a classification problem for ML. Because we have labeled data, we can provide input x as a vector of multiple features extracted from each document provided, and output y will be the corresponding label, i.e. the language code.

Good features would be any property of the document that is not common across all languages and can be categorized into varying buckets to ensure correct classification.

Features used could be -
- n-grams (unigrams, bigrams, etc.)
- Ascii value range for characters
- Doc ID
- k most frequent words in the doc
- Embeddings of words
- POS tags

**(b)** Factors that could affect system performance would be-

- Size of word embeddings
- Classifier model chosen for task
- Hyperparameters chosen for task
- Preprocessing of data

**Q7 (10 "free" points):**