600.465 – Natural Language Processing
Assignment 2: Probability and Vector Exercises

Jinyi Guo

February 2016

1. (a) Let Z be the event space E, we have $p(Z) = 1$.
Since $Y \subseteq Z$, let $X = Y^c \cap Z$, we have $Y = X \cup Y$. Notice that X and Y are disjoint i.e. $X \cap Y = \emptyset$, we have $p(Z) = p(X) + p(Y)$ where $p(X) \geq 0$. Thus we have $p(Z) \geq p(Y)$.

   (b) By the definition of conditional probability we have $p(X|Z) = \frac{p(X \cap Z)}{p(Z)}$.
If X and Z are disjoint, then $p(X \cap Z) = 0$ and $p(X|Z) = 0$. If $X \subseteq Z$ and $Z \subseteq X$ then $p(X \cap Z) = 1$.
Hence $p(X \cap Z)$ falls in the range $[0, 1]$. Notice the fact that $(X \cap Z) \subseteq Z$ we have $p(X \cap Z) \leq p(Z)$.
Therefore $p(X|Z)$ always fall in the range $[0, 1]$.

   (c) Let X equals to the event space E, we have $p(X) = p(E) = 1$. Let $Y = \emptyset$, then obviously $p(X \cup Y) = p(X) + p(Y)$ since $X \cap Y = \emptyset$.
Notice that $X \cup Y = X$, we have $p(X) + p(Y) = p(X) = 1$. Thus $p(Y) = p(\emptyset) = 1 - p(X) = 0$.

   (d) Let $Y = E - X$, then we have $E = X \cup Y \Rightarrow p(E) = p(X) + p(Y) = 1$ since $X \cap Y = \emptyset$ and E is the event space. Hence $p(X) = 1 - p(Y) = 1 - p(\bar{X})$.

   (e) By the definition of conditional probability, we have

$$p(singing\ AND\ rainy|rainy) = \frac{p((singing\ AND\ rainy) \cap rainy)}{p(rainy)} \tag{1}$$

   Notice that $p((singing\ AND\ rainy) \cap rainy) = p(singing \cap rainy \cap rainy) = p(singing \cap rainy)$ since set intersection is associative. Then by applying the definition of conditional probability again we have

$$p(singing\ AND\ rainy|rainy) = \frac{p(singing \cap rainy)}{p(rainy)} = p(singing|rainy) \tag{2}$$

   (f) By the definition of conditional probability we have

$$p(X|Y) = \frac{p(X \cap Y)}{p(Y)} \tag{3}$$

$$p(\bar{X}|Y) = \frac{p(\bar{X} \cap Y)}{p(Y)} \tag{4}$$

   Summing up (3) and (4) we have

$$p(X|Y) + p(\bar{X}|Y) = \frac{p(X \cap Y) + p(\bar{X} \cap Y)}{p(Y)} = \frac{p(Y)}{p(Y)} = 1 \tag{5}$$

   Thus

$$p(X|Y) = 1 - p(\bar{X}|Y) \tag{6}$$

(g) First we have:

$$p(X|Y) \cdot p(Y) + p(X|\bar{Y}) \cdot p(\bar{Y}) = \frac{p(X \cap Y)}{p(Y)} \cdot p(Y) + \frac{p(X \cap \bar{Y})}{p(\bar{Y})} \cdot p(\bar{Y}) \tag{7}$$

$$= p(X \cap Y) + p(X \cap \bar{Y}) \tag{8}$$

$$= p(X) \tag{9}$$

Substituting the result into the original polynomial we have:

$$p(X) \cdot \frac{p(\bar{Z} \cap X)}{p(X) \cdot p(\bar{Z})} = p(X \mid \bar{Z}) \tag{10}$$

(h) Only when singing and being rainy are disjoint event (i.e.these two things never happen at the same time) the equation holds. In this case $S \cap R = \emptyset$ where S stands for singing and R stands for being rainy. Hence we have:

$$p(singing\ OR\ rainy) = p(S \cup R) = p(singing) + p(rainy)$$

(i) Only when the two things are independent does the equation hold. This is by the definition of independent event:

$$p(singing\ AND\ rainy) = p(S \cap R) = p(singing) \cdot p(rainy)$$

(j) Since $p(X|Y) = 0$, we have:

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} = 0 \tag{11}$$

Thus we must have $p(X,Y) = 0$ here. Since $X \cap Y \cap Z \subseteq X \cap Y$, we have $p(X,Y,Z) \leq p(X,Y) = 0$. Hence,

$$p(X|Y,Z) = \frac{p(X,Y,Z)}{p(Y,Z)} = 0 \tag{12}$$

(k)

$$p(W|Y) = \frac{p(W,Y)}{p(Y)} = 1$$

$$p(W,Y) = p(Y) = p(W,Y) + p(\bar{W},Y)$$

Hence $p(\bar{W},Y) = 0$ and $p(\bar{W},Y) \geq p(\bar{W},Y,Z) = 0$, we have

$$p(W|Y,Z) = p(W,Y,Z)/p(Y,Z)$$

$$= \frac{p(Y,Z) - p(W,Y,\bar{Z})}{p(Y) - p(Y,\bar{Z})}$$

$$= \frac{p(Y,Z)}{p(Y,Z)} = 1$$

2. (a)

$$p(Actual = blue \mid Claimed = blue)$$

$$= p(Claimed = blue \mid Actual = blue) \cdot \frac{p(Actual = blue)}{p(Claimed = blue)}$$

(b) Prior probability: $p(Actual = blue)$.
Likelihood of the evidence: $p(Actual = blue \mid Claimed = blue)$.
Posterior probability:$p(Claimed = blue \mid Actual = blue)$.

(c)

$$p(Actual = blue) = 0.1$$
$$p(Actual = blue \mid Claim = blue) = 0.8$$

Using Bayes' theorem, we have:

$$p(Claimed = blue \mid Actual = blue) \tag{1}$$
$$= \frac{p(Actual = blue \mid Claimed = blue)}{p(Claimed = blue)} \cdot p(Actual = blue) \tag{2}$$

where

$$\begin{aligned} p(Claimed = blue) = p(Claimed = blue \mid Actual = blue) \cdot p(Actual = blue) + \\ p(Claimed = blue \mid Actual = red) \cdot p(Actual = red) \\ = 0.8 \cdot 0.1 + 0.2 \cdot 0.9 \\ = 0.26 \end{aligned}$$

Substituting this result into (2) we have

$$p(Claimed = blue \mid Actual = blue) = 4/13 = 0.31 \tag{3}$$

The judge should care about $p(Actual = blue \mid Claimed = blue)$ because it measures how reliable the witness's words are.

(d) By expanding both sides of the equation we can prove it as follows:

$$LHS = \frac{p(A,B,Y)}{p(B,Y)} \tag{4}$$
$$RHS = \frac{p(A,B,Y) \cdot p(A,Y) \cdot p(Y)}{p(A,Y) \cdot p(Y) \cdot p(B,Y)} = LHS \tag{5}$$

(e)

$$LHS = \frac{p(A,B,Y)}{p(B,Y)} \tag{6}$$
$$RHS = \frac{p(A,B,Y) \cdot p(A,Y)}{p(A,Y) \cdot p(Y)} \cdot \tag{7}$$
$$\frac{1}{[p(A,B,Y)/p(A,Y)] \cdot [p(A,Y)/p(Y)] + p(\bar{A},B,Y) \cdot p(\bar{A},Y)/p(\bar{A},Y) \cdot p(Y)} \tag{8}$$
$$= \frac{p(A,B,Y)}{p(A,B,Y) + p(\bar{A},B,Y)} \tag{9}$$
$$= \frac{p(A,B,Y)}{p(B,Y)} = LHS \tag{10}$$

where the final step is based on the fact that $(A \cap B \cap Y) \cup (\bar{A} \cap B \cap Y) = B \cap Y$.

(f)

$$\begin{aligned} p(Actual = blue \mid Claimed = blue, Baltimore) \\ = p(Claimed = blue \mid Actual = blue, Baltimore) \cdot p(Actual = blue \mid Baltimore) / \\ (p(Claimed = blue \mid Actual = blue, Baltimore) \cdot p(Actual = blue \mid Baltimore) + \\ p(Claimed = blue \mid Actual = red, Baltimore) \cdot p(Actual = red \mid Baltimore)) \end{aligned}$$

Replacing the numbers from the problem we get:

$$p(Actual = blue \mid Claimed = blue, Baltimore) = \frac{0.8 \cdot 0.1}{0.8 \cdot 0.1 + 0.2 \cdot 0.9} = 0.31$$

3. (a)

$$\sum_i p(cry_i \mid situation) = 1 \tag{11}$$

   (b) The joint probability table:

$p(cry, situation)$	Predator!	Timber!	I need help!	TOTAL
bwa	0	0	0.64	0.64
bwee	0	0	0.08	0.08
kiki	0.2	0	0.08	0.28
TOTAL	0.2	0	0.8	1

   (c) i. $p(predator \mid kiki)$.

   ii.

$$\frac{p(predator, kiki)}{p(kiki)}$$

   iii. 5/7

   iv.

$$\frac{p(kiki \mid predator) \cdot p(predator)}{p(kiki \mid predator) \cdot p(predator) + p(kiki \mid timber) \cdot p(timber) + p(kiki \mid help) \cdot p(help)}$$

   v. The result is:

$$\frac{0.2}{0.2 + 0 \cdot 0 + 0.08}$$
$$= 5/7 = 0.714$$

4. (a)

$$p(\vec{w}) = \frac{c(w_{-1}w_0w_1)}{c(w_{-1}w_0)} \cdot \frac{c(w_0w_1w_2)}{c(w_0w_1)} \cdot \ldots \frac{c(w_{n-1}w_nw_{n+1})}{c(w_{n-1}w_n)}$$

   $c(BOS\ BOS)$ means the number of sentence in the corpus while $c(BOS\ BOS\ \texttt{i})$ is the count of cases where the sentence starts by the word "i". $c(\texttt{new york}\ EOS)$ means the bigram `new york` ends the sentence.

   (b) This is because normally a sentence ending with determiner is ungrammatical, and a good language model is less likely to be trained from a dataset containing ungrammatical sentences, making the likelihood of these ungrammatical sentences very low.

   The following parameters are responsible: $p(\texttt{so} \mid \texttt{you}, \texttt{think})$, $p(< /\texttt{s} > \mid \texttt{think}, \texttt{so})$.

   (c) (A) matches (2), for there is no sentence delimiter appear in this expression. This means the probability of the three words appear together in any part of the sentence other than EOS or BOS.

   (B) matches (1). This expression counts the conditional probabilities from BOS to EOS, meaning the probability that a complete sentence "Do you think" appears.

   (C) matches (3). This expression starts from a BOS but ends by the word `think`, so any word or EOS coould appear after `think`.

   By the definition of a trigram language model, the quantity in (B) is $p(\vec{w})$.

   (d) By the definition we have:

$$p(\vec{w}) = p(w_1 \mid BOS) \cdot p(w_2 \mid BOS, w_1) \cdot \ldots \cdot p(w_n \mid w_{n-2}, w_{n-1}) \cdot p(EOS \mid w_{n-1}, w_n) \tag{12}$$

$$= \frac{c(BOS\ BOS\ w_1)}{c(BOS\ BOS)} \cdot \frac{c(BOS\ w_1\ w_2)}{c(BOS\ w_1)} \cdot \frac{c(w_1w_2w_3)}{c(w_1w_2)} \cdot \ldots \cdot \frac{c(EOS\ w_{n-1}\ w_n)}{c(w_{n-1}w_n)} \tag{13}$$

$$p_{reverse}(\vec{w}) = p(BOS \mid w_1, w_2) \cdot \ldots \cdot p(w_n \mid w_{n+1}, w_{n+2}) \tag{14}$$

$$= \frac{c(BOS\ w_1w_2)}{c(w_1w_2)} \cdot \ldots \cdot \frac{c(w_{n-2}w_{n-1}w_n)}{c(w_{n-1}w_n)} \cdot \frac{c(w_{n-1}w_n\ EOS)}{c(w_n\ EOS)} \cdot \frac{c(w_n\ EOS\ EOS)}{c(EOS\ EOS)} \tag{15}$$

By inspecting the two equations and canceling out the common elements (assuming $c(w_{i-1}w_iwi+1 \neq 0)$ we have $p(\vec{w}) = p_{reverse}(\vec{w})$ iff

$$\frac{c(BOS\ BOS\ w_1)}{c(BOS\ BOS) \cdot c(BOS\ w_1)} = \frac{c(w_n\ EOS\ EOS)}{c(EOS\ EOS) \cdot c(w_n\ EOS)}$$

Since $c(BOS\ BOS\ w_1) = c(BOS\ w_1)$, $c(w_n\ EOS\ EOS) = c(w_n\ EOS)$, and $c(BOS\ BOS)$ and $c(EOS\ EOS)$ are equal for they both denote the number of sentence of the corpus. This equation holds. Thus $p(\vec{w}) = p_{reverse}(\vec{w})$.

5. By conditional independence assumption we have:

$$p(w_1w_2w_3w_4) = p(w_1 \mid BOS) \cdot p(w_2 \mid w_1) \cdot p(w_3 \mid w_2) \cdot p(w_4 \mid w_3)$$

Taking the topic into account, we have:

$$p(w_1w_2w_3w_4) = \sum_a p(w_1, w_2, w_3, w_4, a)$$

$$= \sum_a p(w_1 \mid w_2, a) + \sum_a p(w_2 \mid w_1, a) + \sum_a p(w_3 \mid w_2, a) + \sum_a p(w_4 \mid w_3, a)$$

8. (a) The most similar words to `seattle` are: seahawks, spokane, tacoma, florida, atlanta.
The most similar words to `dog` are: badger, dogs, hound, cat, borzoi.
The most similar words to `communist` are: socialist, communists, comintern, bolshevik, leftist.
The most similar words to `jpg` are: png, svg, galleria, gif, fuji.
The most similar words to `the` are: its, in, entire, of, which.
The most similar words to `google` are: com, yahoo, faq, flickr, web.

Examples work well: we have city or state name for similar words of `seattle`, and animal words for `dog`. They all fall into the same category. Bad examples include the word "fuji" is found similar to `jpg`, and the similar words of `the` are not of the same part-of-speech with "the", they only share some common points like they don't refer to a specific thing.

Patterns noticed: the model works find with words with specific meaning. e.g. nouns that refer to a specific thing like "dog" and "cat". For those without specific meaning or the meaning is not well defined, the model may only find out words that tend to appear together or look similar but not necessarily related.

(b) Examples work well include "queen" is found for `king - man + woman`, and the case that "aircraft" for `car - road + air`, or "children", "parents" and "infants" for `child - goose + geese` because the plural and family member property is retained. Usually words being added are retained while properties of words being subtracted are somehow reduced.

Bad examples include "bomber" for `car - road + air` for it has neither opposite property of "road" or similar property of "car" and "air".

When we switch to $d = 10$ the result seems more deviated from the property we assumed to have. i.e. most of the cases the model works poorly, the words found are often not related to the given 3 words at all.

`king - man` is the vector that represents all the properties/features of word "king" subtracted by the properties/features of "man".

These vectors quantify the meaning of each word from many perspectives. For example, "king" may have both the meaning of a male and meaning of a person with great power. That's why when we "lower" its meaning of male by subtracting the vector of "man" and add the vector of "woman", we have "queen" as the most similar word.

9. Using the chain rule we have:

$$p(A, B, C, D) = p(A) \cdot p(D|A) \cdot p(C|A, D) \cdot p(B|A, C, D)$$

Notice that A, C and B, D are two independent pair, we can backoff the above equation as follows:

$$p(A,B,C,D) = p(A) \cdot p(D|A) \cdot p(C|D) \cdot p(B|A,C) \tag{16}$$

By applying the chain rule again to $p(B|A,C)$ and canceling out the independent variables we have:

$$\begin{aligned} p(A,B,C,D) &= p(A) \cdot p(D|A) \cdot p(C|D) \cdot p(C) \cdot p(B|C) \cdot p(A|B,C)/(p(A) \cdot p(C)) \\ &= p(A|B) \cdot p(B|C) \cdot p(C|D) \cdot p(D|A) \end{aligned}$$

10. By (a) and the result in 1k we have

$$p(\neg shoe \mid \neg nail) = p(\neg shoe \mid \neg nail, \neg horse) = 1 \tag{17}$$

$$p(\neg shoe \mid \neg nail) = \frac{p(\neg shoe, \neg nail)}{p(\neg nail)} = 1 \Rightarrow \tag{18}$$

$$p(\neg shoe, \neg nail) = p(\neg nail) \tag{19}$$

By equation (b) we have:

$$p(\neg horse \mid \neg shoe) = p(\neg horse \mid \neg shoe, \neg nail) = 1 \Rightarrow \tag{20}$$

$$p(\neg shoe, \neg nail, \neg horse) = p(\neg shoe, \neg nail) \tag{21}$$

By (18), (20) we have:

$$p(\neg nail) = p(\neg shoe, \neg nail, \neg horse) \tag{22}$$

Hence we have:

$$p(\neg horse, \neg shoe \mid nail) = \frac{p(\neg horse, \neg shoe, \neg nail)}{p(\neg nail)} = 1 \tag{23}$$

And again by the conclusion in 1a we have:

$$p(\neg horse \mid nail) \geq p(\neg horse, \neg shoe \mid nail) = 1$$

By applying the same method to equations (c) and (d) we have:

$$p(\neg fortune \mid \neg nail) \geq p(\neg fortune, \neg race, \neg horse, \neg shoe \mid \neg nail) = 1$$