91.542 - Natural Language Processing
Problem Set 1

Chang Liu
chang_liu@student.uml.edu

September 28, 2015

1 Part I. Problem Setup, Language Modeling

1.1 Problem 1.[40 pts]

I)**Solution:**

a) the instances are the 300 articles / 100,000 text lines / 900,000 words, each article is an instance, we need to classify each instance into the starting marker, body, ending marker, for these three parts.

b) the labels are the '<article>' and '<\article>'

c) 300 * $\frac{2}{3}$ = 200 articles, and each article should have a starting and ending marker.

d) The five features for the label '<article>' can be:

1) article has started or not

2) article has ended or not

3) article started with a blank line or not

4) article ended with a blank line or not

5) article has another field that marks the starting or ending point(like the <title> marker)

II). **Solution:**

a) the instances are the 300 articles, each article is an instance.

b) There're two labels I want to assign, the first one is the 'article', the second one is the 'title'. By using these two labels, I can select out each article and its title with the marker '<article>' and '<title>'. I think maybe the problem is not very clear or my understanding is correct, but if we have already known each separated article without worrying about dividing them, then the classifier is just to predict the '<title>' labels.

c) 200 articles as the part I shows.

d) The five boolean features for the label '<title>' can be:

1) Whether they're all upper characters, form 'A' to 'Z'.

2) Whether they're the first word in the paragraph.

3) Whether they're followed by comma.
4) Whether their previous line is the blank line.
5) Whether they're the after/before the maker of the article start/end point.

1.2 Problem 2.[10 pts]

(a) When adding the sentence end-markers, there will be 2,000 more tokens there, and the total tokens should be 22,000. When using bigrams, the corpus select the neighbourhood two tokens as a bigrams, so the total number for bigrams should be **21,999**, which is just one less than the total token numbers.

(b) Using the same schemes, trigrams also has **21,998**, which is just two less than the total token numbers.

(c) In my understanding, the reliable counts should be the trigram that doesn't contain the sentence end-markers, so for these 2,000 markers, each marker will appear in 3 trigrams, so there have been 6,000 trigrams that don't have reliable counts, others are all reliable counts. So the result should be **15,998**.

1.3 Problem 3.[15 pts]

Solution: In order to prove that the distribution is valid, we just need to prove that the sum of these probabilities are 1. Sum up all the probability of $P_{add}(w_i)$, then we can get the following equation:

$$\begin{aligned}
\sum_w P_{add}(w_i|w_{i-(n-1)},...,w_{i-1}) &= \sum_w \frac{\sigma + count(w_{i-(n-1)},...,w_i)}{\sigma|V| + \sum_w count(w_{i-(n-1)},...,w_{i-1}w)} \\
&= \frac{\sum_w (\sigma + count(w_{i-(n-1)},...,w_i))}{\sigma|V| + \sum_w count(w_{i-(n-1)},...,w_{i-1}w)} \\
&= \frac{\sigma * \sum_w 1 + \sum_w count(w_{i-(n-1)},...,w_i)}{\sigma|V| + \sum_w count(w_{i-(n-1)},...,w_{i-1}w)} \\
&= \frac{\sigma|V| + \sum_w count(w_{i-(n-1)},...,w_{i-1}w)}{\sigma|V| + \sum_w count(w_{i-(n-1)},...,w_{i-1}w)} \\
&= 1
\end{aligned}$$

In the above proof, we know that $\sum_w 1 = |V|$, as the number of w for all the vocabulary is the $|V|$, and the sum of the $count(w_{i-(n-1)},...,w_{i-1}w_i)$ in the w field is just the total of the $count(w_{i-(n-1)},...,w_{i-1}w)$, so the sum of the probability is 1, which is a valid distribution.

1.4 Problem 4.[30 pts]

(a)**Solution:**

Using the chain rule as follows, just get the probability(add a starting and ending marker as <s> in each sentence):

$$
\begin{aligned}
P_a &= P(We|<s>) * P(seek|We) * P(a|seek) * P(solution|a) * P(that|solution) \\
&\quad * P(could|that) * P(be|could) * P(accepted|be) * P(by|accepted) * P(both|by) \\
&\quad * P(sides|both) * P(.|sides) * P(<s>|.) \\
&= \frac{1}{9} * \frac{1}{2} * \frac{1}{1} * \frac{1}{8} * \frac{1}{1} * \frac{3}{4} \\
&* \frac{2}{5} * \frac{1}{4} * \frac{1}{1} * \frac{1}{3} * \frac{1}{2} * \frac{1}{2} \\
&* \frac{1}{2} * \frac{1}{8} * \frac{9}{13} \\
&\approx 0.000003756
\end{aligned}
$$

If we don't add the starting and ending marker <s>, then the value should be larger, multiplied by 13 and should be 0.0000488.

From the training dataset, I can calculate the probability for each of the items, and the multiply them together, by using this function, I can get the final result is **0.000003756**.

For more details, check the code, I print out the counter for bigrams and their base counter, which forms the above fractional number.

(b) **Solution:**

The same rules goes for this question, except that when calculating the probability that we need to do some smoothing with absolute discounting, for each value minus the 0.03 for the fractional value, the final result is

$$\frac{0.78374335943}{250453.135692} = \mathbf{0.0000031293}.$$

Check the code for details.

2 Part II. Combinatorics, Probability, Information Theory

2.1 Problem 5. [30 pts]

(a) **Solution:** Assume that $S_k = \sum_{k=1}^{K} k * r^k$, then $S_{k+1} = \sum_{k=1}^{K} (k+1) * r^{k+1}$, then we need to split the equation of S_{k+1}, and get the following equation:

$$
\begin{aligned}
S_{k+1} &= r * \sum_{k=1}^{K} k * r^k + r * \sum_{k=1}^{K} r^n \\
&= r * S_k + r^2/(1-r)
\end{aligned}
$$

After further transformation, we can conclude that $S_k - r^2/(1-r)^2$ is a geometrical sequence, which meets the basic form of

$$\frac{S_{k+1} - r^2/(1-r)^2}{S_k - r^2/(1-r)^2} = r$$

Then using the equation to calculate the items for a geometrical sequence, we can get the general form for S_k. By combining the limitation of the $\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$, we can get the general item:

$$S_n - r^2/(1-r)^2 = r * \frac{1}{1-r}$$

So we can get the value of S_n, as follows:

$$S_n = \frac{r}{(1-r)^2}$$

(b) **Solution:** Assume that $P(X)$ is the probability of the X, then we can conclude that:

$$P(X = k) = 1/2^k$$

So we can get the table as follows:

X	1	2	3	...	k
P(X)	1/2	1/4	1/8	...	$1/2^k$.

Then according to the entropy equation, calculate the entropy:

$$\begin{aligned} H(X) &= -(\frac{1}{2} * \log\frac{1}{2} + \frac{1}{4} * \log\frac{1}{4} + \ldots + \frac{1}{2^k} * \log\frac{1}{2^k}) + \ldots \\ &= \log 2 * (\frac{1}{2} + \frac{2}{2^2} + \ldots + \frac{k}{2^k} + \ldots) \end{aligned}$$

So divide the $H(X)$ by 2, then I will get another equation:

$$\frac{H(X)}{2} = \log 2 * (\frac{1}{2^2} + \frac{2}{2^3} + \ldots + \frac{k}{2^{k+1}} + \ldots)$$

Subtract the above two equations and will get a geometrical sequence:

$$\begin{aligned} \frac{H(X)}{2} &= \log 2 * (\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots + \frac{1}{2^k} + \ldots) \\ &= \log 2 * (\frac{1}{2} + \frac{1}{2}) \\ &= \log 2 \end{aligned}$$

So the value of $H(X)$ is $\underline{2\log 2}$

2.2 Problem 6. [20 pts]

Solution: According to the definition of KL-Divergence, just get the $D(p||q)$ using this equation:

$$\begin{aligned} D(p||q) &= \sum_{x \in X} p(x) * \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in X} p(x) * (\log p(x) - \log q(x)) \\ &= (1-r) * \log \frac{1-r}{1-s} + r * \log \frac{r}{s} \\ &= \log \frac{1-r}{1-s} + r * \log \frac{r(1-s)}{s(1-r)} \end{aligned}$$

Using the same equation, we can get the $D(q||p)$:

$$\begin{aligned} D(q||p) &= \sum_{x \in X} q(x) * \log \frac{q(x)}{p(x)} \\ &= (1-s) * \log \frac{1-s}{1-r} + s * \log \frac{s}{r} \\ &= \log \frac{1-s}{1-r} + s * \log \frac{s(1-r)}{r(1-s)} \end{aligned}$$