CMPT 413/825: Natural Language Processing **(Due: 09/30/20)**

Homework #1

Name: Coulton Fraser
SFU Email: coultonf@sfu.ca, *ID:* 301405411 *Students discussed with:*

Course Policy: Read all instructions carefully before you start working on the assignment, and before you make a submission. The course assignment policy is available at https://angelxuanchang.github.io/nlp-class/.

Problem 1. Language models

a)

$$\begin{aligned} P(store|computer) &= \frac{C(w_{computer} w_{store})}{C(w_{computer})} \\ &= \frac{4}{10} \\ &= 0.4 \\ P(monitor|computer) &= \frac{C(w_{computer} w_{monitor})}{C(w_{computer})} \\ &= \frac{4}{10} \\ &= 0.4 \end{aligned} \tag{0.1}$$

No word is more likely than the other.

b)

$$\begin{aligned} P(store|computer) &= \frac{1 + C(w_{computer} w_{store})}{|V| + C(w_{computer})} \\ &= \frac{5}{34} \\ &= 0.147 \\ P(monitor|computer) &= \frac{1 + C(w_{computer} w_{monitor})}{|V| + C(w_{computer})} \\ &= \frac{5}{34} \\ &= 0.147 \end{aligned} \tag{0.2}$$

No word is more likely than the other.

c)

$$\begin{aligned}
P(store|computer) &= \lambda_1 \frac{C(w_{computer} w_{store})}{C(w_{computer})} + \lambda_2 \frac{C(w_{store})}{N} \\
&= 0.6\frac{4}{10} + 0.4\frac{6}{11} \\
&\approx 0.4582 \\
P(monitor|computer) &= \lambda_1 \frac{C(w_{computer} w_{monitor})}{C(w_{computer})} + \lambda_2 \frac{C(w_{monitor})}{N} \\
&= 0.6\frac{4}{10} + 0.4\frac{5}{11} \\
&\approx 0.4218
\end{aligned} \tag{0.3}$$

Store is slightly more likely than monitor by about 3

Problem 2. Log-linear models

a) $f_1(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\text{=a and } pos\text{=A} \\ 0 & \text{otherwise} \end{array} \right\}$

$f_2(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\text{=bear and } pos\text{=N} \\ 0 & \text{otherwise} \end{array} \right\}$

$f_3(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\text{=eats and } pos\text{=V} \\ 0 & \text{otherwise} \end{array} \right\}$

$f_4(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\neg \text{ [a, bear, eats] and } pos\text{=A} \\ 0 & \text{otherwise} \end{array} \right\}$

$f_5(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\neg \text{ [a, bear, eats] and } pos\text{=N} \\ 0 & \text{otherwise} \end{array} \right\}$

$f_6(word, pos) = \left\{ \begin{array}{ll} 1 & \text{if } word\neg \text{ [a, bear, eats] and } pos\text{=V} \\ 0 & \text{otherwise} \end{array} \right\}$

b)

$$\begin{aligned} p(A|cat) &= \frac{e^{w_A \cdot f_4(cat,A)}}{1 + \sum_{j=1}^{k} e^{w_j \cdot f_j(x,c)}} \\ p(N|walks) &= \frac{e^{w_N \cdot f_5(walks,N)}}{1 + \sum_{j=1}^{k} e^{w_j \cdot f_j(x,c)}} \\ p(A|bear) &= \frac{1}{1 + \sum_{j=1}^{k} e^{w_j \cdot f_j(x,c)}} \end{aligned} \tag{0.4}$$

c) $ln\frac{P(X)}{1-P(X)} = b + w_1X...w_nX$
Given a bias term of 0, the coefficients of f_n is the log odds of P as when $f_n(x, c) = 1$ all other features are zero. Therefore...

$$\begin{aligned} f_1, f_2, f_3 &= ln(\frac{0.7}{1-0.7}) \\ &= 0.3680 \\ f_4, f_5 &= ln(\frac{0.4}{1-0.4}) \\ &= -0.1761 \\ f_6 &= ln(\frac{0.2}{1-0.2}) \\ &= -0.6021 \end{aligned} \tag{0.5}$$

Problem 3. Naive Bayes

Add One Smoothing Counts

	what	is	the	title	of	book
C(s)	1	2	3	1	2	3
C(q)	3	3	3	1	1	3

$|V| = 22$
$\sum count(w,c) = 32$
$\hat{P}(W_i|c) = \frac{count(w_i,c)+1}{(\sum count(w,c))+|V|}$

$$\begin{aligned}
P(w_{what}, c_{statement}) &= \frac{1+1}{32+22} \\
&= \frac{2}{54} \\
P(w_{is}, c_{statement}) &= \frac{2+1}{32+22} \\
&= \frac{3}{54} \\
P(w_{the}, c_{statement}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(w_{title}, c_{statement}) &= \frac{1+1}{32+22} \\
&= \frac{2}{54} \\
P(w_{of}, c_{statement}) &= \frac{2+1}{32+22} \\
&= \frac{3}{54} \\
P(w_{book}, c_{statement}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(c = statement|d_8) &= 2(\frac{2}{54})2(\frac{3}{54})3(\frac{4}{54}) \\
&= 1.829 \times 10^{-3}
\end{aligned} \tag{0.6}$$

$$\begin{aligned}
P(w_{what}, c_{question}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(w_{is}, c_{question}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(w_{the}, c_{question}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(w_{title}, c_{question}) &= \frac{1+1}{32+22} \\
&= \frac{2}{54} \\
P(w_{of}, c_{question}) &= \frac{1+1}{32+22} \\
&= \frac{2}{54} \\
P(w_{book}, c_{question}) &= \frac{3+1}{32+22} \\
&= \frac{4}{54} \\
P(c = question|d_8) &= 5(\frac{4}{54})2(\frac{2}{54}) \\
&= 0.0274
\end{aligned} \tag{0.7}$$

The probability of D_8 is in favour of it being a question classification (0.0274).

Problem 4. Evaluation

No, accuracy is not a sufficient metric for choosing the better model. This is because accuracy doesn't represent the false positives and false negatives nearly as well as F1 score. In the case of tumours, priority of the false positives and false negatives is critical. This is because if a model produces a lot of false positives then an large amount of patients will be given harmful treatment, and in the other case of false negatives, many patients will not be receiving the treatment they need.

$$
\begin{aligned}
Accuracy &= \frac{TP + FP}{TP + FP + TN + FN} \\
Precision &= \frac{TP}{TP + FP} \\
Recall &= \frac{TP}{TP + FN} \\
F1 &= 2\left(\frac{Precision * Recall}{Precision + Recall}\right)
\end{aligned}
\tag{0.8}
$$

Problem 5. Decision boundaries

a.
(C)

b.
(A)
Regularization reduces the variance of a model by decreasing the coefficients, therefore the outliers have less effect. Plot A shows a model where the 'X' data points are not curving the line like the other plots, therefore the regularization must be maximum to create the greatest reduction in variance. If the model is over fitting, then there must not be a lot of regularization.