Homework 3: Smoothed Language Modeling

Meili Liu

October 9, 2022

Exercise 1.Perplexities and corpora

on switchboard-small corpus - perplexity per word for three sample files:
$sample1 : 2^{(-(-9785.28/ln(2)/(1358+164)))} = 619.693$
$sample2 : 2^{(-(-5768.78/ln(2)/(762+108)))} = 758.074$
$sample3 : 2^{(-(-5997.96/ln(2)/(785+101)))} = 871.056$

change to large switchboard corpus:
$sample1 : 2^{(-(-9332.37/ln(2)/(1358+164)))} = 460.194$
$sample2 : 2^{(-(-5626.62/ln(2)/(762+108)))} = 643.794$
$sample3 : 2^{(-(-5869.42/ln(2)/(785+101)))} = 753.424$

Exercise 2. textcat program

compare two model's posterior probability
if these two probabilities are same, then choose the model with higher prior probability

Exercise 3. Evaluating a text classifier

(a) the result on gen:
179 files were more probably gen.model (99%)
1 files were more probably spam.model (1%)
the result on spam:
68 files were more probably gen.model (76%)
22 files were more probably spam.model (24%)
for gen: the error rate is 0.01
for spam: the error rate is 0.76
the total error rate is 0.26
thus about 26% files are classified incorrectly

(b) train the model use laplace 0.01.

the result on en.model
106 files were more probably en.model (88%)
14 files were more probably sp.model (12%)
the result on sp.model
9 files were more probably en.model (8%)
110 files were more probably sp.model (92%)
the total error rate is 0.096
only 9.6% of dev files were classified incorrectly

(c) 0
(d) try different λ for gen model to get a best cross-entropy per token

λ	cross-entropy
2	10.73
1	10.45
0.5	10.15
0.1	9.51
0.02	9.11
0.01	9.043
0.009	9.039
0.008	9.036
0.007	9.037
0.006	9.039
0.005	9.046
0.001	9.29

when λ =0.008, we can achieve minimum cross-entropy at 9.036

try different λ for spam model to get a best cross-entropy per token

λ	cross-entropy
2	10.78
1	10.54
0.5	10.27
0.1	9.66
0.02	9.23
0.01	9.13
0.009	9.121
0.008	9.113
0.007	9.104
0.006	9.099
0.005	9.096
0.004	9.098
0.001	9.256
0.0001	9.96

Thus, when $\lambda = 0.005$, the best cross-entropy is 9.096
(e) total tokens on dec/gen/* 48198+180=48378 total tokens on dev/spam/* is 39284+90 = 39374

λ	cross-entropy
0.008	(9.036*48378+9.113*39374)/(48378+39374)=9.071
0.007	(9.037*48378+9.104*39374)/(48378+39374)=9.067
0.006	(9.039*48378+9.099*39374)/(48378+39374)=9.066
0.005	(9.046*48378+9.096*39374)/(48378+39374)=9.068

so, when $\lambda = 0.006$, the minimum cross-entropy is 9.066
(f)
in the figure, I labeled: gen - gen as 0; gen-spam as 1; spam -gen 2; spam-spma as 3

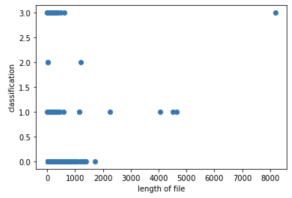

Figure 1: correlation between file length and classification accuracy

therefore, compare label0 and label1, label2 and label3 (except the outlier), we can know that with the increase of file length, the accuracy goes down.
(h)

$training_data$	overall error rate
gen	(3+30)/(180+90) = 0.122
gen-times2	(2+17)/(180+90) = 0.0704
gen-times4	(7+11)/(180+90) = 0.0667
gen-times8	(2+16)/(180+90) = 0.067

no,the accuracy will not go to 0 as the increase of training zero. it keeps steady.

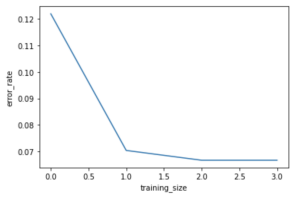

Figure 2: learning curve

Exercise 4. Analysis

(a) then the total sum probability of all words will exceed 1
same result for the add-λ estimate.
(b)if we set $\lambda = 0$, the model would totally copy the training data,so it could have big variance and cannot generalize well on test data.

(c)no.
$\hat{p}(z \mid xy) = \frac{\lambda V \cdot \hat{p}(z|y)}{c(xy)+\lambda V}$

where $\hat{p}(z' \mid xy) = \frac{\lambda V \cdot \hat{p}(z'|y)}{c(xy)+\lambda V}$
therefore, they are not equal unless $\hat{p}(z \mid y) == \hat{p}(z' \mid y)$
the answer is same for $c(xyz) = x(xyz') = 1$

(d)when increase λ, the cross-entropy and perplexity will also increase. when λ is big enough, the probability model will work like a uniform distribution.

Exercise 5. backoff smoothing implementation

$p(z \mid xy) = \frac{c(xyz)+\lambda V p(z|y)}{c(xy)+\lambda V)}$
$p(z \mid y) = \frac{c(yz)+\lambda V p(z)}{c(y)+\lambda V)}$
$p(z) = \frac{c(z)+\lambda}{N+\lambda V}$

Exercise 6.sampling from language models

the first model is gen model based on add-lambda smoothing with $\lambda = 0.006$
the second model is gen-backoff model based on add-lambda backoff smoothing with $\lambda = 0.006$

SUBJECT: Re : nothing Medicine complex suspected unless parser Furthermore done know proposed improve speech Channel brand disgusted listen especially ...
SUBJECT: Re VICTORY Music close online leaving is NOTHING beyond does given few convenient analysis Smallest notice global brought chocolate ...
SUBJECT: Re : (EMAIL) , NAME NAME Models corresponding Eat correctly happen interesting pm crucial bio e-mail surprised ...
SUBJECT: OOV details for Entities bored fuel next drama revise small size direct Private privileged 's? Your important scared Need ...
SUBJECT: NAME whose Inc. status machines Babe ahead Lane spoil nourished known enjoying distribution context alarm July gone Approach environment ...
SUBJECT: Re : Regarding university simply races So models individuals Tuesdays calls chips identical guitar pretty Next Text Dinner allowed ...
SUBJECT: Regarding unless included Obviously Who order exercise visiting 's? fare tutorial staying drink Students entertained Over cellular Trust write ...
SUBJECT: summary unfortunately environment discovered implementation loads other mail disposable plan upon examinations auld AHMED first experiment photograph netscape taking ...
SUBJECT: meeting Tuesday cash finishes repaying teams private rolling numerical accept worth steps fucking 5th life give someone Careers reported ...
SUBJECT: NAME Club completed general kin grace length babe OUT ability visitors library girls busy parents – spilt BORDER Imagination ...

SUBJECT: in some pictures " . If you 're talking NUM NAME : NUM) , so please keep it

...
SUBJECT: OOV practicals French n't know you all know . The difference exists between the left after me in a ...
SUBJECT: Re : Arrangements for Saturday evening and have a business this weekend . Please reply usual name : NUM ...
SUBJECT: reminder that the long be at 6pm I do n't succeed would like to tell me by risk , ...
SUBJECT: [Fwd been proposed will have to arrive at NUM EOS
SUBJECT: Your to others rather and NUM people per hour as vacuum NAME , I will do you soon ! ...
SUBJECT: Training NAME and NAME AT OOV OOV than NUM " <END_QUOTE> EOS
SUBJECT: NAME Dear NAME NAME . I mean I would like to have a fab you cannot Maybe I have ...
SUBJECT: Re : Arrangements happy . love , see vengeance : NUM NUM , NUM NAME : NUM NUM Speech ...
SUBJECT: ?) , NAME , or in your development go . This message . The circuits immediately . Have ...

1. these sentence generated by back-off model are more like real emails.
2. sentences in backoff model contains lots of high-frequency words and phases, such as NUM, NAME, I, you, in, me, I will, you're, punctuation and etc. Becasue they back off to bigram and unigram.
3. sentences in backoff model are shorter, because the frequency use of punctuation, which actually make the sentence more meaningful, such as "so please keep it", "so please keep it". Therefore, the backoff model take advantage of trigram bigram and unigram.
4. by contrast, the sentences generated by only add-lambda model are hard to read and seems totally meaningful. I guess if we use more data to train it, the generative quality would be better

Exercise 7.implementing a log-linear model

7(b)
gen_spam problem lr = 0.1

c value	cross-entropy(per token)	error_rate
c=1	8197.85 and 7319.82	(137+22)/270 = 0.59
c=0.5	7813.46 and 7513.05	(60+36)/270 = 0.36
c=0.1	7013.46 and 6613.05	(60+36)/270 = 0.35
c=0.01	7052.10 and 6728.5	(9+81)/270 = 0.33
c=0.05	7088 and 7526	(69+46)/270 = 0.43

therefore, the $c^{\cdot}$ = 0.1 for lr=0.1 however, when use the learning rate to 0.0001, change the l2-regularization did not affect the cross entropy and text classification accuracy

lay with different embedding dimensions and report the results.
with the increase of dimensions, the error rate goes down

How and when did you use the training, development, and test data?
training data is used for training model.
development data is used for adjust hyperparameters, such as C, prior probability and early stop
only use test data when report test result.

compare to add-λ backoff smoothing
the error rate of backoff smoothing is (6+11)/270 = 0.063
backoff smoothing model has much lower error rate

7(d) I have added (1)shuffling (2)OOV feature and (3)Unigram log-probability feature (4) early stop (5) convergent SGD
early stop: I use the cross entropy as the evaluation metric. if the cross entropy on the development data failed to improve 2 times, then stop.

Exercise 8. Speech recognition

we want to know $p(w/u)$,which is the probability of sentence w, given the utterance. According to Bayes's Theorem: $p(w/u) = \frac{p(u/w)\cdot p(w)}{p(u)}$
we skip $p(u)$, since the utterance has been given, then we can just compute $p(w/u) = argmax(p(u/w) \cdot p(w))$
$p(w) = p(w_1 w_2 w_3 ... w_n)$ can be computed by our language model