Quizz for the class on Natural Language Processing

Edouard Grave

Question 1. The PMI Matrix.

a. The Mutual Information (MI) between two variable x and y is:

1. $MI(x, y) = \log p(x, y) - \log p(x) + \log p(y)$
2. $MI(x, y) = \log p(x, y) + \log p(x) - \log p(y)$
3. $MI(x, y) = \log p(x, y) + \log p(x) + \log p(y)$
4. $MI(x, y) = \log p(x, y) - \log p(x) - \log p(y)$

b. The truncated singular value decomposition of the PMI matrix $\mathbf{M}$ to its d highest singular values is $\mathbf{U}_d,\ \mathbf{\Sigma}_d,\ \mathbf{V}_d^T$ =SVD($\mathbf{M}$). We get d dimensional distributed word vectors $\mathbf{W}$ by taking:

1. $\mathbf{W} = \mathbf{\Sigma}_d$
2. $\mathbf{W} = \mathbf{U}_d\mathbf{\Sigma}_d$
3. $\mathbf{W} = \mathbf{U}_d\mathbf{\Sigma}_d^{1/2}\mathbf{V}_d^T$
4. $\mathbf{W} = \mathbf{U}_d\mathbf{V}_d^T$

Question 2. The sigmoid σ is defined by $\sigma(x) = \frac{1}{1+\exp(-x)}$. What is the gradient of log sigmoid?

1. $\sigma(x)$
2. $1 - \sigma(x)$
3. $\sigma(x)(1 + \sigma(x))$
4. $\sigma(x)(1 - \sigma(x))$

Question 3. What is the tokenization of the sentence

The answer isn't correct.

1. `[The] [answer] [isn't] [correct.]`
2. `[The] [answer] [isn] ['t] [correct] [.]`
3. `[The] [answer] [is] [n't] [correct] [.]`
4. it depends on the tokenizer

Question 4. Let $\mathbf{s} \in \mathbb{R}^d$, and f is the softmax operator, such that $f_i(\mathbf{s}) = \dfrac{\exp(s_i)}{\sum_{j=1}^{d} \exp(s_j)}$. We have:

1. $\max_i f_i(\mathbf{s}) < \max_i f_i(10\,\mathbf{s})$
2. $\max_i f_i(\mathbf{s}) > \max_i f_i(10\,\mathbf{s})$
3. $\max_i f_i(\mathbf{s}) = \max_i f_i(10\,\mathbf{s})$
4. It depends on the value of $\mathbf{s}$.

Question 5. What is expression of the logistic loss function?

1. $\log(1 - \exp(x))$
2. $\exp(1 + \log(-x))$
3. $\log(1 + \exp(-x))$
4. $1 + \log(\exp(-x))$

Question 6. What is the stochastic gradient descent update of logistic regression?

1. $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta(1 - \sigma(\mathbf{w}_t^\top \mathbf{x}_i))\mathbf{x}_i$
2. $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta(1 - \sigma(y_i \mathbf{w}_t^\top \mathbf{x}_i))y_i \mathbf{x}_i$
3. $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta(1 - \sigma(y_i \mathbf{w}_t^\top \mathbf{x}_i))y_i \mathbf{w}_i$
4. $\mathbf{w}_{t+1} = \mathbf{w}_t + \eta(1 - \sigma(\mathbf{w}_t^\top \mathbf{x}_i))\mathbf{w}_i$

Question 7. Why is stochastic gradient descent useful in natural language processing?

1. because it has a faster convergence rate than gradient descent
2. because we don't need to optimize the train loss below the statistical error
3. because the noise in the gradient regularizes the model
4. because it guarantees the loss to decrease at each update

Question 8. When using dropout, hidden units are dropped at random during

1. training
2. testing
3. training and testing
4. it depends on the model

Question 9. The log perplexity of a sentence $W = (w_1, \ldots, w_T)$ is equal to:

4. $\log PP(W) = -\frac{1}{T} \sum_{t=1}^{T} \log P(w_t \mid w_{t-1}, \ldots, w_1)^{1/T}$
4. $\log PP(W) = -\frac{1}{T} \prod_{t=1}^{T} \log P(w_t \mid w_{t-1}, \ldots, w_1)$
4. $\log PP(W) = -\frac{1}{T} \log \left(\sum_{t=1}^{T} P(w_t \mid w_{t-1}, \ldots, w_1) \right)$
4. $\log PP(W) = -\frac{1}{T} \sum_{t=1}^{T} \log P(w_t \mid w_{t-1}, \ldots, w_1)$

Question 10. In the continuous bag of words (cbow) model from word2vec, the context is represented by

1. using a recurrent neural network to encode the context
2. computing the pointwise mutual information between words of the context
3. sampling one word vector from the context
4. computing the average of the word vectors of the context

Question 11. In a recurrent neural network, such that $\mathbf{h}_t = \sigma(\mathbf{A}\mathbf{w}_t + \mathbf{R}\mathbf{h}_{t-1})$, we have

1. $\|\frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_{T-k}}\| \leq 0.25^k \lambda_{\text{max}}(\mathbf{R})^k$
2. $\|\frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_{T-k}}\| \geq 0.25^k \lambda_{\text{min}}(\mathbf{R})^k$
3. none of the above
4. both of the above

Question 12. The perceptron algorithm always converges to a local minimum

1. true
2. false

Question 13. Given label y_t in $\{-1, 1\}$ and x_t a vector in $\mathbb{R}^d$, the perceptron has the following decision function: $\hat{y}_t = \text{sign}(w_t^T x_t)$. What is the update rule of the perceptron algorithm, after making a mistake?

1. $w_{t+1} = w_t + y_t x_t$
2. $w_{t+1} = w_t - y_t x_t$
3. $w_{t+1} = w_t + x_t$
4. $w_{t+1} = w_t - x_t$

Question 14. To solve the exploding gradient issue in recurrent neural networks, we

1. use backpropagation through time (BPTT)
2. use gradient clipping
3. use dropout
4. use learning rate decay

Question 15. We have a dataset split for positive $(+)$ and negative $(-)$ restaurant reviews:

$+$	pasta were great
$-$	food was bad
$-$	pasta were not good
$+$	best food ever
$-$	pasta were terrible

We have a new review:

pasta were bad

We want to know what is the most likely label and the associated joint probability given by a Naive Bayes model?

1. The most likely label is $-$ and $P(-,$ pasta were bad$) = \frac{2}{1250}$
2. The most likely label is $+$ and $P(+,$ pasta were bad$) = \frac{1}{750}$
3. The most likely label is $-$ and $P(-,$ pasta were bad$) = \frac{3}{1250}$
4. The most likely label is $-$ and $P(-,$ pasta were bad$) = \frac{1}{250}$

Question 16. We have a dataset containing $N = 1200$ words in total with a vocabulary of 25 unique words. We want to estimate the probability of the following sentence:

<s> i study machine learning

with a counted based model. To do so, we provide the unigrams counts for 8 tokens:

i	want	study	math	<s>	machine	learning	like
40	132	30	174	50	36	60	64

as well as their bigram counts:

$w_1 \backslash w_2$	<s>	i	want	study	math	machine	learning	like
<s>	0	25	0	0	0	1	5	0
i	0	0	4	12	3	0	0	10
want	0	1	0	1	2	7	3	0
study	0	0	0	0	11	5	4	4
math	0	0	0	5	0	0	0	0
machine	0	0	0	6	0	0	12	5
learning	0	0	2	0	0	3	0	22
like	0	8	0	1	0	2	8	0

(For example $c(\text{i want}) = 4$ and $c(\text{want i}) = 1$)

We remind that, by convention, $P(\text{<s>}) = 1$.

a. What is the probability of the sentence given by a unigram model?

1. $\frac{1}{400000}$
2. $\frac{1}{600000}$
3. $\frac{1}{800000}$
4. 0

b. What is the probability of the sentence given by a bigram model?

1. $\frac{1}{60}$
2. $\frac{1}{120}$
3. $\frac{1}{250}$
4. 0

Question 17. What is the complexity of the newton method for optimizing a binary linear logistic regression of dimension d over n examples?

1. $O(n^2 + dn)$
2. $O(d^2 + dn)$
3. $O(n^3 + dn^2)$
4. $O(d^3 + nd^2)$

Bonus question. What are the recurrent equations of an LSTM?