CS269: Machine Learning in Natural Language Processing **Fall 2017**

Midterm

Nov. 9^{th}, 2017

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **three** problems.
- You have 30 minutes to earn a total of 25 points.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**

Good Luck!

Name and ID: (1 Point)

Short Questions		/10
Graphical Model Assumptions		/5
Representation Learning		/4
Total		/20

Short Questions [10 points]

(a) [3 points] Determine if the statement is true or false, and justify your answer.

- The difference between MEMM and CRF is that the probability distribution in CRF is normalized locally.

  False. the probability distribution in CRF is normalized globally.

- Both HMM and MEMM models are discriminative models because in the decoding time we are estimating

$$\arg\max_{Y} P(Y \mid X).$$

  False. Despite HMM estimates $max_Y P(Y \mid X)$ in the decoding time, it is a generative model as it models $P(X, Y)$.

- Adding more hidden layers will solve the vanishing gradient problem for a 3 layer neural network

  False. Adding more layers cannot solve the vanishing gradient problem.

(b) [2 points] In the soft-max function, we can use a temperature parameter σ to control the distribution of the output:

$$P(Y = y; \sigma) = \frac{\exp(s(y)/\sigma)}{\sum_{y' \in 1,2,3...K} \exp(s(y)/\sigma)}$$

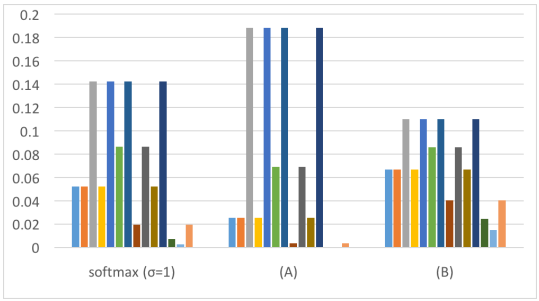

Figure 1: softmax output with different σ

Figure 1 shows the probability distribution over 14 labels. If the bar graph on the left is the distribution of softmax output with $\sigma = 1$, which one (A or B) is the distribution of softmax output with $\sigma = 2$? (B).

(c) [2 points] Choose from the following options to fill the blanks:
(a) $P(Y|X)$ (b) $P(X|Y)$ (c) $P(X, Y)$ (d) $P(X)$ (e) $P(Y)$

Given the observations X and labels Y, a generative model makes independent assumptions to specify (c) $P(X, Y)$, while a discriminative classifier explicitly models (a) $P(Y|X)$.

(d) [2 points] Figure 2 describes the Viterbi algorithm. However, the algorithm suffers from arithmetic underflow when n is large. To avoid this issue, we can do all the calculations in log space. Please describe how to modify the Viterbi algorithm such that the computations are done in the log space (Hint: $\log(xy) = \log(x) + \log(y)$)

1. Initial: For each state s, calculate

$$\text{score}_1(s) = P(s)P(x_1|s) = \pi_s B_{x_1,s}$$

2. Recurrence: For i = 2 to n, for every state s, calculate

$$\begin{aligned}\text{score}_i(s) &= \max_{y_{i-1}} P(s|y_{i-1})P(x_i|s)\text{score}_{i-1}(y_{i-1}) \\ &= \max_{y_{i-1}} A_{y_{i-1},s} B_{s,x_i} \text{score}_{i-1}(y_{i-1})\end{aligned}$$

3. Final state: calculate

$$\max_{\mathbf{y}} P(\mathbf{y}, \mathbf{x}|\pi, A, B) = \max_s \text{score}_n(s)$$

π: Initial probabilities
A: Transitions
B: Emissions

Figure 2: Viterbi algorithm

First, pre-compute $\log \pi, \log A, \log B$ (optional). Then, modify the initial step in the algorithm as:

$$\text{score}_1(s) = \log \pi_s + \log B_{x_1,s}$$

.

Finally, modify the recurrence step as:

$$\text{score}_i(s) = \max_y \log A_{y_{i-1},s} + \log B_{s,x_i} + \text{score}_{i-1}(y).$$

Graphical Model Assumptions [5 points]

Given one binary random variable $Y = \{0, 1\}$, we can use one parameter ($\gamma_1 \equiv P(Y = 1)$) to specify its discrete probability distribution. Without any assumption, to specify the joint probability of two binary random variables X, Y, we need three parameters ($\gamma_1 \equiv P(X = 1, Y = 1), \gamma_2 \equiv P(X = 1, Y = 0), \gamma_3 \equiv P(X = 0, Y = 1)$. Note that we don't need to specify $P(X = 0, Y = 0)$ because it can be derived from $1 - \gamma_1 - \gamma_2 - \gamma_3$.

Suppose $Y_1, Y_2, \ldots, Y_T$ are T binary random variables (i.e., Y_i can take value either 1 or 0.) in a sequence. Assume $T > 2$.

(a) [1 points] How many parameters will we need at least to completely describe the joint distribution $P(Y_1, Y_2, \ldots, Y_T)$, without any assumptions on the relations between them? $2^T - 1$

(b) [1 points] Suppose a domain expert tells us that the value of Y_{i+2} is influenced only by the previous two variables Y_{i+1} and Y_i for $i = 1, 2, \ldots, T-2$ and that the dependency of any other previous variables on Y_{i+2} can be ignored and further that Y_2 is influenced only by Y_1. Write down the joint probability $P(Y_1, Y_2, \ldots, Y_T)$ based on the above independent assumptions.

$P(Y_1, Y_2, \ldots, Y_T) =$ $P(Y_1)P(Y_2|Y_1)\prod_{i=1}^{T-2} P(Y_{i+2} \mid Y_{i+1}, Y_i)$

(c) [1 points] How many parameters will we need at least to specify the joint distribution under the above independent assumption? We need 4 parameters to specify the conditional probability of $P(Y_{i+2} \mid Y_{i+1}, Y_i)$, 2 parameters for $P(Y_2|Y_1)$ and 1 parameter for $P(Y_1)$. Therefore, we need $4(T-2)+2+1 = 4T-5$ in total. (0.5 point for any answer in $O(T)$).

(d) [1 points] Suppose the domain expert further tells us all the conditional probabilities $P(Y_{i+2}|Y_{i+1}, Y_i), i = 1, 2, \ldots T-2$ share the same set of parameters (for example, $P(Y_3 = 0 \mid Y_2 = 0, Y_1 = 1) = P(Y_4 = 0 \mid Y_3 = 0, Y_2 = 1)$). Based on the above independent assumptions, how many parameters do you need now (hint: don't forget the initial probability $P(Y_1)$ and $P(Y_2 \mid Y_1)$)? We need $4 + 2 + 1 = 7$ in total. (0.5 point for any answer in $O(1)$).

(e) [1 points] Discuss your observations from (a)-(d)

We need less number of parameters to specify a model when a stronger independent assumptions are made.

Representation Learning [4 points]

You are consulting for a law firm. They provide you a collection of legal documents and ask you to classify their categories. The company also provides you with the true annotations of the document types. A quick analysis shows that the total vocabulary size is 10,000.

(a) [1 point] If we want to represent each word in a document using a one-hot vector, describe how to do it. First assign each word a unique index. Then create a 10,000-dimensional vector to represent each word. To encode the word with the index i, set i-th element to be 1 and all other components of the vector to be 0.

(b) [1 point] What are the advantages and disadvantages of using word-embedding in comparison with using one-hot vectors (provides at least 2 differences).

1. One-hot vectors are high-dimensional and sparse, while word embeddings are low-dimensional and dense.
2. Similar words have similar vectors in word embedding model, which is not the case in one-hot vectors.

(c) [2 points] Determine if the statement is true or false.

- Recurrent neural network can handle sentences with various length, while a 3-layer feed forward neural network cannot. True. (full credit given to answer False with a good explanation).

- A 2-layer feed forward neural network can only handle non-linear data. If the data is linearly separable, then it cannot be separated by a 2-layer feed forward neural network.

False. Model that can separate non-linear data can also separate linear data.