CS 533: Natural Language Processing **(Due: 04/21/20)**

Assignment 5

Instructor: Karl Stratos

- 2 problems: total 20 points (11 + 9)
- No collaboration
- Due by 11:59pm of the due date, no late submission accepted
- Use the provided LaTeX assignment template to write the answers. Upload the code as well.

Problem 1: EM

(1 + 5 + 5 = 11 points)

Consider the following variant of the bigram language model with parameters

- $p(z|x)$: conditional probability of $z \in \{1 \ldots m\}$ given $x \in V \cup \{*\}$
- $p(x'|z)$: conditional probability of $x' \in V$ given $z \in \{1 \ldots m\}$

where $m \geq 1$ is an integer, V denotes the vocabulary, and $*$ a special beginning-of-sentence symbol. Given a sequence of words $x_1 \ldots x_T \in V$ and a corresponding sequence of integers $z_1 \ldots z_T \in \{1 \ldots m\}$, the model defines the joint probability

$$p(x_1 \ldots x_T, z_1 \ldots z_T) = \prod_{t=1}^{T} p(z_t|x_{t-1}) \times p(x_t|z_t)$$

where we assume $x_0 = *$.

1. Given $x_1 \ldots x_T \in V$ and $z_1 \ldots z_T \in \{1 \ldots m\}$, what is the maximum likelihood estimate (MLE) of the model parameters? That is, write down as a function of the data

$$p(z|x) = \qquad\qquad \forall z \in \{1 \ldots m\}\,,\ x \in V \cup \{*\}$$

$$p(x'|z) = \qquad\qquad \forall z \in \{1 \ldots m\}\,,\ x' \in V$$

   that maximize $\log p(x_1 \ldots x_T, z_1 \ldots z_T)$ under the constraint that $p(z|x)$ and $p(x'|z)$ are proper conditional distributions.

2. Given only $x_1 \ldots x_T \in V$, we wish to compute the MLE of the model parameters. That is, we wish to find conditional distributions $p(z|x)$ and $p(x'|z)$ that maximize

$$\log \sum_{z_1 \ldots z_T \in \{1 \ldots m\}} p(x_1 \ldots x_T, z_1 \ldots z_T)$$

   Recall that this objective no longer has a closed-form solution but EM can be used to iteratively optimize the objective (without guarantees) by alternating the E step and the M step each of which does have a closed-form solution. Specifically, we first initialize the parameters somehow (e.g., random distributions) and repeat the two steps. Give the **E step** which calculates the posterior distribution for each data point as a function of the current model parameters $p(z|x)$ and $p(x'|z)$:

$$p(z|x_{t-1}, x_t) = \qquad\qquad \forall z \in \{1 \ldots m\}\,,\ t \in \{1 \ldots T\}$$

3. Give the **M step** which computes new parameter values $p^{\text{new}}(z|x)$ and $p^{\text{new}}(x'|z)$ as a function of the per-datum posterior probabilities $p(z|x_{t-1}, x_t)$:

$$p^{\text{new}}(z|x) = \qquad\qquad \forall z \in \{1 \dots m\}\,,\ x \in V \cup \{*\}$$

$$p^{\text{new}}(x'|z) = \qquad\qquad \forall z \in \{1 \dots m\}\,,\ x' \in V$$

Problem 2: VAE (3 + 3 + 3 = 9 points)

Consider a latent-variable generative language model which defines

- $p_Y(y) = \mathcal{N}(0_d, I_d)$: prior probability of $y \in \mathbb{R}^d$
- $p_Z(z) = \mathcal{N}(0_d, I_d)$: prior probability of $z \in \mathbb{R}^d$
- $p^\theta_{X|YZ}(\boldsymbol{x}|y,z)$: conditional probability of any sentence $\boldsymbol{x} = (x_1 \dots x_T) \in V^T$ given $y, z \in \mathbb{R}^d$. θ denotes the learnable parameters of the distribution. This can be defined in a number of ways, for instance

$$p^\theta_{X|YZ}(\boldsymbol{x}|y,z) = \prod_t \text{softmax}_{x_t}(\text{RNN}([e(x_{t-1}), z], [h_t, y]))$$

  where an RNN cell predicts the next token conditioning on y, z as well as its hidden state and the previous word embedding. In this case θ refers to word embeddings and all parameters of the RNN.

The model defines the joint probabilty of any $y, z \in \mathbb{R}^d$ and sentence $\boldsymbol{x}$ by

$$p^\theta_{YZX}(y,z,\boldsymbol{x}) = p_Y(y) \times p_Z(z) \times p^\theta_{X|YZ}(\boldsymbol{x}|y,z)$$

Given a single sentence $\boldsymbol{x}$ as training data, the MLE objective is to find parameters θ that maximize

$$J^{\text{MLE}}(\theta) = \log \int_{y \in \mathbb{R}^d} \int_{z \in \mathbb{R}^d} p^\theta_{YZX}(y,z,\boldsymbol{x})$$

1. We introduce a variational model ϕ that defines the posterior distribution with a conditional independence assumption: $q^\phi_{YZ|X}(y,z|\boldsymbol{x}) = q^\phi_{Y|X}(y|\boldsymbol{x}) \times q^\phi_{Z|X}(z|\boldsymbol{x})$. Give the corresoponding VAE objective $J^{\text{ELBO}}(\theta,\phi)$ (which is a lower bound on $J^{\text{MLE}}(\theta)$ for all ϕ). The objective must take the form of (1) an expectation of the reconstruction term under $p^\theta_{X|YZ}$ with respect to $q^\phi_{Y|X}$ and $q^\phi_{Z|X}$, minus (2) the KL divergences associated with $q^\phi_{Y|X}$ and $q^\phi_{Z|X}$.

$$J^{\text{ELBO}}(\theta,\phi) =$$

2. Re-express $J^{\text{ELBO}}(\theta,\phi)$ in the previous question as a differentiable function of θ and ϕ by using the (single-sample) reparameterization trick on the reconstruction term and the closed-form formula for the KL divergence between Gaussian distributions. To be specific, assume that

$$q^\phi_{Y|X}(\cdot|\boldsymbol{x}) = \mathcal{N}(\mu_{\phi,Y}(\boldsymbol{x}), \text{diag}\left(\sigma^2_{\phi,Y}(\boldsymbol{x})\right))$$
$$q^\phi_{Z|X}(\cdot|\boldsymbol{x}) = \mathcal{N}(\mu_{\phi,Z}(\boldsymbol{x}), \text{diag}\left(\sigma^2_{\phi,Z}(\boldsymbol{x})\right))$$

   where $\mu_{\phi,Y}, \mu_{\phi,Z}, \sigma^2_{\phi,Y}, \sigma^2_{\phi,Z}$ are differentiable functions parameterized by ϕ that encode a sentence to a d-dimensional vector.

3. Draw the computation graph underlying the loss in the previous question.