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

Assignment 5

Name: Madhu Sivaraj*, NetID:* ms2407

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

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

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

where we assume $x_0 = *$.

1. Given $x_1 ... x_t \in V$ and $z_1 ... z_t \in \{1...m\}$, the maximum likelihood estimate (MLE) of the model parameters can be found below. For $p(z|x)$, we sum over the total number of x and for $p(x'|z)$, we sum over z.

   $p(z|x) = \frac{count(z|x)}{\sum_{x'' \in V \cup \{*\}} count(z|x'')}$, $\qquad \forall z \in \{1...m\}, x \in V \cup \{*\}$

   $p(x'|z) = \frac{count(x'|z)}{\sum_{t=1}^{T} count(x'|z_t)}$, $\qquad \forall z \in \{1...m\}, x' \in V$

2. The expectation-maximization (EM) algorithm is an iterative method to find maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables. It iteratively optimizes the objective by alternating the E step and the M step each of which does have a closed-form solution.The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

   Below is the posterior distribution of z, calculated by the E step.

   $p(z|x_{t-1}, x_t) = \frac{p(z|x_{t-1})p(x_t|z)}{\sum_{z'} p(z'|x_{t-1})p(x_t|z')}$, $\qquad \forall z \in \{1...m\}, t \in \{1...T\}$

3. The M step computes parameters maximizing the expected log-likelihood found on the E step. We compute new parameter values $p^{new}(z|x)$ and $p^{new}(x'|z)$ as a function of the perdatum posterior probabilities $p(z|x_{t-1}, x_t)$.

   $p^{new}(z|x) = \frac{\sum_{x_i \in V} p(z|x,x_i)p(z|x)}{\sum_{x_i \in V} \sum_{z' \in \{1...m\}} p(z'|x,x_i)p(z'|x)}$, $\qquad \forall z \in \{1...m\}, x \in V \cup \{*\}$

   $p^{new}(x'|z) = \frac{\sum_{x_i \in V \cup \{*\}} p(z|x_i,x')p(x'|z)}{\sum_{x_i \in V \cup \{*\}} \sum_{z' \in \{1...m\}} p(z'|x_i,x')p(x'|z')}$, $\qquad \forall z \in \{1...m\}, x' \in V$

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

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

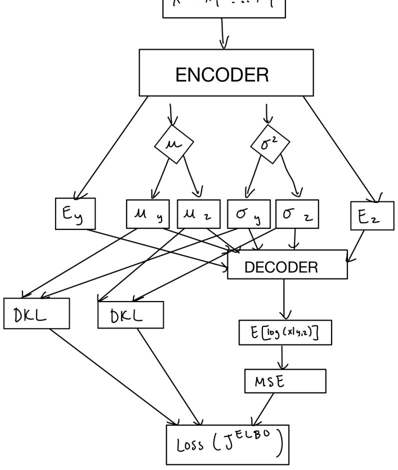
1. We introduce a variational model ϕ that defines the posterior distribution with a given conditional independence assumption. The objective takes the form of an expectation of the reconstruction term minus the KL divergences (D_{KL}). See below for the corresponding VAE objective $J^{ELBO}(\theta, \phi)$, which is a lower bound on $J^{MLE}(\theta)$ for all ϕ.

$$J^{ELBO}(\theta,\phi) = E_{y\sim q^\phi(y|x), z\sim q^\phi(z|x)} \text{ [log } p^\theta_{X|YZ}(x|y,z) \text{] - } D_{KL}(q_c(y|x)||p_\theta(y)) - D_{KL}(q_c(z|x)||p_\theta(z))$$

2. We re-express $J^{ELBO}(\theta, \phi)$ in the previous question as a differentiable function of θ and ϕ by using the singlesample reparameterization trick on the reconstruction term and the closed-form formula for the KL divergence between Gaussian distributions. We use the reparameterization trick to express a gradient of an expectation and as an expectation of a gradient. Without the reparameterization trick, we have no guarantee that sampling large numbers a sample will help converge to the right estimate. See re-expression below.

$$J^{ELBO}(\theta,\phi) = E[log(p(x)\mu(x) + \sigma(x)\cdot E, \mu(x) + \sigma(x)\cdot E)] - \frac{1}{2}[\sum_{i=1}^m \sigma^2_{\phi,y}(x) + \mu^2_{\phi,y}(x) - 1 - log\sigma^2_{\phi,y}(x)] - \frac{1}{2}[\sum_{i=1}^m \sigma^2_{\phi,z}(x) + \mu^2_{\phi,z}(x) - 1 - log\sigma^2_{\phi,z}(x)]$$

3. Below is the computation graph underlying the loss in 2.2.


X → x₁ ...x_T
ENCODER
μ
σ²
Ey
μ y
μ z
σ y
σ z
Ez
DECODER
DKL
DKL
E[log(x|y,z)]
MSE
Loss (J^ELBO)