STAT215: Assignment 3

Your Name

Due: March 3, 2020

Problem 1: *Variational inference.*

Standard VI minimizes $\mathrm{KL}(q(z) \,\|\, p(z \mid x))$, the Kullback-Leibler divergence from the variational approximation $q(z)$ to the true posterior $p(z \mid x)$. In this problem we will develop some intuition for this optimization problem. For further reference, see Chapter 10 of *Pattern Recognition and Machine Learning* by Bishop.

(a) Let $\mathcal{Q} = \{q(z) : q(z) = \prod_{d=1}^{D} \mathcal{N}(z_d \mid m_d, v_d^2)\}$ denote the set of Gaussian densities on $z \in \mathbb{R}^D$ with diagonal covariance matrices. Solve for

$$q^\star = \arg\min_{\mathcal{Q}} \mathrm{KL}\left(q(z) \,\|\, \mathcal{N}(z \mid \mu, \Sigma)\right),$$

where Σ is an arbitrary covariance matrix.

Your answer here.

(b) Now solve for $q^\star \in \mathcal{Q}$ that minimizes the KL in the opposite direction,

$$q^\star = \arg\min_{\mathcal{Q}} \mathrm{KL}\left(\mathcal{N}(z \mid \mu, \Sigma) \,\|\, q(z)\right).$$

Your answer here.

(c) Plot the contour lines of your solutions to parts (a) and (b) for the case where

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}.$$

Problem 2: *Variational autoencoders (VAE's)*

In class we derived VAE's as generative models $p(x, z; \theta)$ of observations $x \in \mathbb{R}^P$ and latent variables $z \in \mathbb{R}^D$, with parameters θ. We used variational expectation-maximization to learn the parameters θ that maximize a lower bound on the marginal likelihood,

$$\log p(x; \theta) \geq \sum_{n=1}^{N} \mathbb{E}_{q(z_n | x_n, \phi)} [\log p(x_n, z_n; \theta) - \log q(z_n \mid x_n, \phi)] \triangleq \mathscr{L}(\theta, \phi).$$

The difference between VAE's and regular variational expectation-maximization is that we constrained the variational distribution $q(z \mid x, \phi)$ to be a parametric function of the data; for example, we considered,

$$q(z_n \mid x_n, \phi) = \mathcal{N}\left(z_n \mid \mu(x_n; \phi), \mathrm{diag}([\sigma_1^2(x_n; \phi), \ldots, \sigma_D^2(x_n; \phi)])\right),$$

where $\mu : \mathbb{R}^P \to \mathbb{R}^D$ and $\sigma_d^2 : \mathbb{R}^P \to \mathbb{R}_+$ are functions parameterized by ϕ that take in a datapoint x_n and output means and variances of z_n, respectively. In practice, it is common to implement these functions with neural networks. Here we will study VAE's in some special cases. For further reference, see Kingma and Welling (2019), which is linked on the course website.

(a) Consider the linear Gaussian model factor model,

$$p(x_n, z_n; \theta) = \mathcal{N}(z_n; 0, I)\, \mathcal{N}(x_n \mid A z_n, V),$$

where $A \in \mathbb{R}^{P \times D}$, $V \in \mathbb{R}^{P \times P}$ is a diagonal, positive definite matrix, and $\theta = (A, V)$. Solve for the true posterior $p(z_n \mid x_n, \theta)$.

Your answer here.

(b) Consider the variational family of Gaussian densities with diagonal covariance, as described above, and assume that $\mu(x; \phi)$ and $\log \sigma_d^2(x; \phi)$ are linear functions of x. Does this family contain the true posterior? Find the member of this variational family that maximizes $\mathscr{L}(\theta, \phi)$ for fixed θ. (Hint: use your answer to Problem 1a.)

Your answer here.

(c) Now consider a simple nonlinear factor model,

$$p(x_n, z_n; \theta) = \mathcal{N}(z_n \mid 0, I) \prod_{p=1}^{P} \mathcal{N}(x_{np} \mid e^{a_p^\top z_n}, v_p),$$

parameterized by $a_p \in \mathbb{R}^D$ and $v_p \in \mathbb{R}_+$. The posterior is no longer Gaussian, since the mean of x_{np} is a nonlinear function of the latent variable.[1]

Generate a synthetic dataset by sampling $N = 1000$ datapoints from a $D = 1$, $P = 2$ dimensional model with $A = [1.2, 1]^\top$ and $v_p = 0.1$ for $p = 1, 2$. Use the reparameterization trick and automatic differentiation to perform stochastic gradient descent on $-\mathscr{L}(\theta, \phi)$.

Make the following plots:

[1] For this particular model, the expectations in $\mathscr{L}(\theta, \phi)$ can still be computed in closed form using the fact that $\mathbb{E}[e^z] = e^{\mu + \frac{1}{2}\sigma^2}$ for $z \sim \mathcal{N}(\mu, \sigma^2)$.

- A scatter plot of your simulated data (with equal axis limits).
- A plot of $\mathscr{L}(\theta, \phi)$ as a function of SGD iteration.
- A plot of the model parameters $(A_{11}, A_{21}, v_1, v_2)$ as a function of SGD iteration.
- The approximate Gaussian posterior with mean $\mu(x; \phi)$ and variance $\sigma_1^2(x; \phi)$ for $x \in \{(0,0), (1,1), (10,7)\}$ using the learned parameters ϕ.
- The true posterior at those points. (Since z is one dimensional, you can compute the true posterior with numerical integration.)

Comment on your results.

Your results here.

Problem 3: *Semi-Markov models*

Consider a Markov model as described in class and in, for example, Chapter 13 of *Pattern Recogntion and Machine Learning* by Bishop,

$$p(z_{1:T} \mid \pi, A) = p(z_1 \mid \pi) \prod_{t=2}^{T} p(z_t \mid z_{t-1}, A),$$

where $z_t \in \{1, \dots, K\}$ denotes the "state," and

$$p(z_1 = i) = \pi_i$$
$$p(z_t = j \mid z_{t-1} = i, A) = A_{ij}.$$

We will study the distribution of state durations—the length of time spent in a state before transitioning. Let $d \geq 1$ denote the number of time steps before a transition out of state z_1. That is, $z_1 = i, \dots, z_d = i$ for some i, but $z_{d+1} \neq i$.

(a) Show that $p(d \mid z_1 = i, A) = \mathrm{Geom}(d \mid p_i)$, the probability mass function of the geometric distribution. Solve for the parameter p_i as a function of the transition matrix A.

Your answer here.

(b) We can equivalently represent $z_{1:T}$ as a set of states and durations $\{(\tilde{z}_n, d_n)\}_{n=1}^{N}$, where $\tilde{z}_n \in \{1, \dots, K\} \setminus \{\tilde{z}_{n-1}\}$ denotes the index of the n-th visited state and $d_n \in \mathbb{N}$ denotes the duration spent in that state before transition. There is a one-to-one mapping between states/durations and the original state sequence:

$$(z_1, \dots, z_T) = (\underbrace{\tilde{z}_1, \dots, \tilde{z}_1}_{d_1 \text{ times}}, \underbrace{\tilde{z}_2, \dots, \tilde{z}_2}_{d_2 \text{ times}}, \dots \underbrace{\tilde{z}_N, \dots, \tilde{z}_N}_{d_N \text{ times}}).$$

Show that the probability mass function of the states and durations is of the form

$$p(\{(\tilde{z}_n, d_n)\}_{n=1}^{N}) = p(\tilde{z}_1 \mid \pi) \left[\prod_{n=1}^{N-1} p(d_n \mid \tilde{z}_n, A)\, p(\tilde{z}_{n+1} \mid \tilde{z}_n, A) \right] p(d_N \mid \tilde{z}_N, A),$$

and derive each conditional probability mass function.

Your answer here.

(c) *Semi-Markov* models replace $p(d_n \mid \tilde{z}_n)$ with a more flexible duration distribution. For example, consider the model,

$$p(d_n \mid \tilde{z}_n) = \mathrm{NB}(d_n \mid r, \theta_{\tilde{z}_n}),$$

where $r \in \mathbb{N}$ and $\theta_k \in [0, 1]$ for $k = 1, \dots, K$. Recall from Assignment 1 that the negative binomial distribution with integer r is equivalent to a sum of r geometric random variables. Use this equivalence to write the semi-Markov model with negative binomial durations as a Markov model on an extended set of states $s_n \in \{1, \dots, Kr\}$. Specifically, write the transition matrix for $p(s_n \mid s_{n-1})$ and the mapping from s_n to z_n.

Your answer here.