Machine Learning Principles and Methods **Due: 12 May, 2016**

Homework 5

Instructor: Ke Tran Email: m.k.tran@uva.nl

You are allowed to discuss with your colleagues but you should write the answers in *your own words*. If you discuss with others, write down the name of your collaborators on top of the first page. No points will be deducted for collaborations. If we find similarities in solutions beyond the listed collaborations we will consider it as cheating.
We will not accept any late submissions under any circumstances. The solutions to the previous homework will be handed out in the class at the beginning of the next homework session. After this point, late submissions will be automatically graded zero.

Problem 1. In this question we are interested in generating samples from a probability density $p(x)$ with $x \in \mathbb{R}^d$. We are given an approximation $q(x)$ of $p(x)$. We will denote unnormalized densities as $\tilde{p}$ and $\tilde{q}$.

a) Assume that you have a constant c such that $\tilde{q}(x) = cq(x)$ and $\tilde{q}(x) \geq p(x), \forall x$. Describe with pseudocode the "Rejection Sampler" algorithm.

b) Are the samples you generate independent from each other?

c) An "Importance Sampler" accepts all samples but weights them using weights w_n. Provide the expression for w_n in terms of $p(x_n)$ and $q(x_n)$.

d) An "Independence Sampler" uses a proposal distribution of the form $q(x_{t+1}|x_t) = q(x_{t+1})$ (i.e. the proposed new state is independent of the previous state) and subsequently accepts or rejects this proposed state as the next state of the Markov chain. Provide the expression for the Metropolis Hastings accept probability $\alpha(x_{t+1}, x_t)$ in terms of p and q for the Independence Sampler.

e) Are two subsequent samples from the Independence Sampler independent or dependent in general? Explain your answer.

f) Imagine we run the Independence sampler for 5 steps and during these 5 steps we propose the states x_1, x_2, x_3, x_4, x_5 (think of these represent as numeric values, e.g. 0.34, 3.5, 2.67, 0.82, 1.60). The MCMC procedure rejects the proposals x_2 and x_5. Which sequence of states will the Independence sampler generate after 5 steps?

g) Will any of the three samplers discussed above work in high-dimensional settings (e.g., $d > 20$)? Explain your answer by discussing how this "curse of dimensionality" will affect each of the three samplers discussed above.

Problem 2. Random walk

Consider a state space z consisting of the integers, with probability

$$\begin{aligned} p(z^{(r+1)} = z^{(r)}) &= 0.5 \\ p(z^{(r+1)} = z^{(r)} + 1) &= 0.25 \\ p(z^{(r+1)} = z^{(r)} - 1) &= 0.25 \end{aligned}$$

where $z^{(r)}$ denotes the state at step r. If the initial state is $z^{(1)} = 0$, prove that

$$\mathbb{E}\left[(z^{(r)})^2\right] = \frac{r}{2}$$

Problem 3. Bishop 11.13
Consider a simple 3-node graph shown in Figure 1 in which

$$\begin{aligned} x &\sim \mathcal{N}(x|\mu, \tau^{-1}) \\ \mu &\sim \mathcal{N}(\mu|\mu_0, s_0) \\ \tau &\sim \text{Gamma}(\tau|a, b) \end{aligned} \tag{1}$$

Derive Gibbs sampling for the posterior distribution $p(\mu, \tau|x)$.

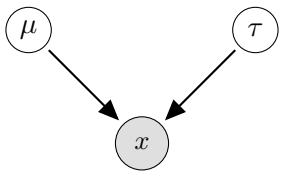

Figure 1: A graph involving an observed Gaussian variable x with prior distributions over its mean μ and precision τ

Problem 4.

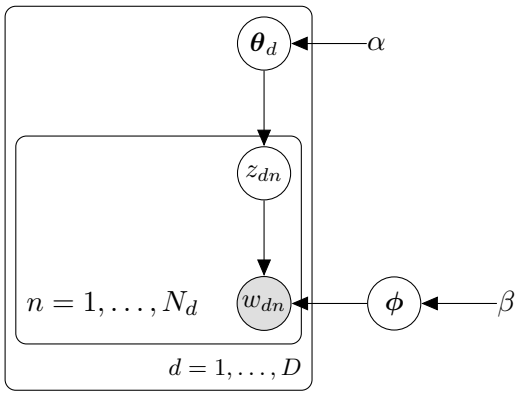

Figure 2: Graphical model representation of LDA

The generative process of LDA model is given as

(i) For $k = 1, \ldots, K$:

  (a) $\boldsymbol{\phi}_k \in \mathbb{R}^{|V|} \sim \text{Dir}(\beta, \ldots, \beta)$

(ii) For each document $\mathbf{w}_d \in \mathcal{D}$

  (a) Draw a topic distribution $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha, \ldots, \alpha)$

  (b) For each of the N word w_n in the document:

i. $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$.
ii. $w_{dn}|z_{dn}, \phi_{dn} \sim \text{Mult}(\phi_{dn})$

Assume our data consists of D documents, a vocabulary of size V, and we model with K topics. Let $A_{dk} = \sum_{n=1}^{N_d} \delta(z_{dn} = k)$ be the number of z_{dn} variables taking on value k in document $\mathbf{w}_d$, and $B_{kw} = \sum_{d=1}^{D} \sum_{i=n}^{N_d} \delta(w_{dn} = w)\delta(z_{dn} = k)$ be the number of times word w is assigned to topic k, where N_d is the total number of words in document $\mathbf{w}_d$, and let $M_k = \sum_w B_{kw}$ be the total number of words assigned to topic k.

1. Write down the joint probability over the observed data and latent variables.
2. Integrate out the parameters $\boldsymbol{\theta}_d$'s and $\boldsymbol{\phi}_k$'s from the joint probability. Express this result in terms of the counts N_d, M_k, A_{dk}, and B_{kw}.
3. Derive the Gibbs sampling updates for z_{di} with all parameters integrated out.

Problem 5. Consider a multivariate Bernoulli distribution

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1 - \mu_i)^{1-x_i}$$

where $\mathbf{x} = (x_1, \ldots, x_D)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$, with $\mu_i \in [0, 1], x_i \in \{0, 1\}$ for $i = 1, \ldots, D$.

a) What is the mean of $\mathbf{x}$ under this distribution?

b) What is the covariance matrix of $\mathbf{x}$ under this distribution?

Now consider a mixture of K of these multivariate Bernoulli distributions

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1 - \mu_{ki})^{1-x_i}$$

c) What is the mean of $\mathbf{x}$ under this mixture distribution?

Suppose we are given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$.

d) Write down the log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression.

e) Why doesn't standard maximum-likelihood work here?

We will use the Variational EM algorithm to learn the parameters of the model. For each datapoint $\mathbf{x}_n$, introduce a latent variable $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})$ which is a one-of-K coded binary vector that indicates the latent class of that datapoint. In other words: the latent variable $\mathbf{z}_n$ has K components, all of which are 0 except for the k'th one that is 1, where k is the latent class for data point $\mathbf{x}_n$. Using these conventions, for data point $\mathbf{x}_n$ and associated latent class $\mathbf{z}_n$, we can write:

$$p(\mathbf{x}_n, \mathbf{z}_n | \boldsymbol{\mu}, \boldsymbol{\pi}) = p(\mathbf{z}_n | \boldsymbol{\pi}) p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\mu}) = \prod_{k=1}^{K} \pi_k^{z_{nk}} p(\mathbf{x}_n | \boldsymbol{\mu}_k)^{z_{nk}}$$

f) Write down the complete-data log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression.

g) Draw the corresponding graphical model using plate notation. Clearly distinguish observed variables, latent variables, parameters, and make clear which variable subscripts are "looped over" if you use plates.

h) Write down an explicit expression for the VEM objective function $\mathcal{B}(\{q_n(\mathbf{z}_n)\}, \boldsymbol{\mu}, \boldsymbol{\pi})$ for this model.

i) Include Lagrange multipliers for all constraints in the model and construct the Lagrangian $\tilde{\mathcal{B}}$ from $\mathcal{B}$. Make the Lagrangian as explicit as possible.

j) Work out the details of the E-step, i.e., optimize $\tilde{\mathcal{B}}$ with respect to q_n for all $n = 1, \ldots, N$. Solve the equation. What is the interpretation of $q_n(\mathbf{z}_n)$?

k) Work out the details of the M-step for $\boldsymbol{\pi}$, i.e., optimize $\tilde{\mathcal{B}}$ with respect to π_k for all k. Solve the equation.