DD2434 Machine Learning, Advanced Course
Assignment 2 (V1.5)

Jens Lagergren

Deadline, see Canvas

Read this before starting

There are several commonalities between the problems and, consequently, it may be useful to read all of them before starting. Also think about the formulation and try to visualize the model. You are allowed to discuss the formulations, but have to make a note of the people you have discussed with. You will present the assignment by a written report, submitted before the deadline using Canvas. You must solve the assignment individually and it will automatically be checked for similarities to other students' solutions as well as documents on the web in general. Although you are allowed to discuss the problem formulations with others, you are not allowed to discuss solutions, and any discussions concerning the problem formulations must be described in the solutions you hand in. From the report it should be clear what you have done and you need to support your claims with results. You are supposed to write down the answers to the specific questions detailed for each task. This report should clearly show how you have drawn your conclusions and explain your derivations. Your assumptions, if any, should be stated clearly. Show the results of your experiments using images and graphs together with your analysis and add your code as an appendix.

Being able to communicate results and conclusions is a key aspect of scientific as well as corporate activities. It is up to you as a author to make sure that the report clearly shows what you have done. Based on this, and only this, we will decide if you pass the task. No detective work should be required on our side. In particular, neat and tidy reports please!

The grading of the assignment will be as follows,

E Correctly completed Problem 2.1, 2.2, 2.3, and 2.4.

D Correctly completed Problem 2.1, 2.2, 2.4, 2.5, and one of 2.6-2.9

C Correctly completed Problem 2.1, 2.2, 2.4, 2.5, and two of 2.6-2.9

B Correctly completed Problem 2.1, 2.2, 2.4, 2.5, and three of 2.6-2.9

A Correctly completed Problem 2.1, 2.2, and 2.4-2.9.

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E. **Notice that 2.3 is not required for grades higher than E. Moreover, if you fail the assignment due to having failed 2.3 or 2.4 but have successfully solved 2.1 and 2.2 it is sufficient to hand in a correct solution of 2.3 and 2.4 in the second round in order to receive an E on this assignment.**

Good Luck!

2.1 Knowing the rules

Question 2.1.1: *It is mandatory to read the above text. Have you read it?*

Question 2.1.2: *List all your collaborations concerning the problem formulations in this assignment.*

Question 2.1.3: *Have you discussed solutions with anybody?*

2.2 Dependencies in a Directed Graphical Model

. Consider the graphs shown in the below figures. In the below question, consider independence as independence in all distributions that factorize according to the graph and dependence as dependence in some such distribution. You merely have to answer "yes" or "no" to each question.

Question 2.2.4: *In the graphical model of Figure 1, is* $\mu_k \perp \tau_k$ *(not conditioned by anything)?*

Question 2.2.5: *In the graphical model of Figure 1, is* $\mu_k \perp \tau_k | X^1, \ldots, X^N$ *?*

Question 2.2.6: *In the graphical model of Figure 2, is* $\mu \perp \beta'$ *(not conditioned by anything) ?*

Question 2.2.7: *In the graphical model of Figure 2, is* $\mu \perp \beta' | X^1, \ldots, X^N$ *?*

Question 2.2.8: *In the graphical model of Figure 2, is* $X^n \perp S^n$ *(not conditioned by anything) ?*

Question 2.2.9: *In the graphical model of Figure 2, is* $X^n \perp S^n | \mu_k, \tau_k$ *?*

2.3 Likelihood of a tree GM only for E level.

Let T be a binary tree, with vertex set $V(T)$ and leaf set $L(T)$, and consider the graphical model T, Θ described as follows. For each vertex $v \in V(T)$ there is an associated random variable X_v that assumes values in $[K]$. Moreover, for each $v \in V(T)$, the CPD $\theta_v = p(X_v | x_{\mathrm{pa}(v)})$ is a categorical distribution. Let $\beta = \{x_l : l \in L(T)\}$ be an assignment of values to all the leaves of T.

Question 2.3.10: *Implement a dynamic programming algorithm that, for a given* T, Θ *and* β*, computes* $p(\beta | T, \Theta)$*.*

Question 2.3.11: *Apply your algorithm to the graphical model and data provided separately.*

2.4 Simple VI

Consider the model defined by Equation (10.21)-(10-23) in Bishop. We are here concerned with the VI algorithm for this model covered during the lectures and in the book.

Question 2.4.12: *Implement the VI algorithm for the variational distribution in Equation (10.24) in Bishop.*

Question 2.4.13: *What is the exact posterior?*

Question 2.4.14: *Compare the inferred variational distribution with the exact posterior. Run the inference on data points drawn from iid Gaussians. Do this for three interesting cases and visualize the results. Describe the differences.*

2.5 Mixture of trees with observable variables

Consider the mixture model $\mathcal{M} = (\pi, \tau)$, where π is a categorical distribution on $[K]$ and $\tau = \{(T_k, \Theta_k) : k \in [K]\}$ is a set of K graphical models that each is a tree with vertices V and root r. All CPDs are binary and all variables are observable. There is an EM algorithm that, for given data $\mathcal{D} = \{x^n : n \in [N]\}$, estimates $\mathcal{M}$ by iteratively performing the following steps w.r.t. to a current $\mathcal{M} = \pi, \tau$.

1. For each n, k, compute the responsibilities
$$r_{n,k} = \pi_k p(x^n | T_k, \Theta_k)/p(x^n).$$
2. Set $\pi'_k = \sum_{n=1}^{N} r_{n,k}/N$.
3. For each k, let G_k be a directed graph with edge weights defined by $w(st) = I_q(X_s, X_t)$, where $I_{q^k}(X_s, X_t)$ is the mutual information between X_s and X_t under the distribution q^k, i.e.,
$$I_{q^k}(X_s, X_t) = \sum_{a,b \in \{0,1\}} q^k(X_s = a, X_t = b) \log \frac{q^k(X_s = a, X_t = b)}{q^k(X_s = a) q^k(X_t = b)},$$
and q^k is defined by
$$q^k(X_s = a, X_t = b) = \frac{\sum_{n \in [N]: X_s^n = a, X_t^n = b} r_{n,k}}{\sum_{n \in [N]} r_{n,k}}.$$
Moreover, any term in $I_{q^k}(X_s, X_t)$ for which $q^k(X_s = a, X_t = b) = 0$ is considered to be 0.
4. Let T'_k be a maximum spanning tree in G_k.
5. Let $\Theta'_k(X_r) = q^k(X_r)$ and $\Theta'_k(X_s = a | X_t = b) = q^k(X_s = a | X_t = b)$.

The root stays the same; it is facilitating our computations, but any root would give the same result. Initialize the EM algorithm randomly, independently of the data, and use sieving. If you run into problems with values being zero, you can change them slightly using the Python "sys" module that contains a function called float_info. (e.g., *sys.float_info.epsilon: Value: $2e^{-16}$ or *sys.float_info.min: Value: $2e^{-308}$.)

Question 2.5.15: *Implement this EM algorithm.*

Question 2.5.16: *Apply your algorithm to the provided data and show how well you reconstruct the mixtures. First, compare the real and inferred trees with the* unweighted Robinson-Foulds *(aka* symmetric difference*) metric. Do the trees have similar structure (don't worry if the inferred trees don't match with the real trees)? Then, compare the likelihoods of real and inferred mixtures. Finally, simulate more data and analyse the results (try to find some interesting and more challenging cases).*

Question 2.5.17: *Simulate new tree mixtures with different number of nodes, samples and clusters. Try to find some interesting cases. Analyse your results as in the previous question.*

2.6 Super epicentra – EM

We have seismographic from an area with frequent earthquakes emanating from K super epicentra. Each super epicentra is modeled by a 2-dimensional Gaussian determining the location of a an earthquake and a Poisson distribution determining its strength. As shown in Figure 1 the entire model is a mixture of K such components. The variable Z^n is a class variable that follows a categorical distribution π and determines the super epicentra of the nth observation, i.e., the parameters of the Gaussian distribution that X_n is sampled from, $\mu_k = (\mu_{k,1}, \mu_{k,2})$ and $\tau_k = (\tau_{k,1}, \tau_{k,2})$ where $\tau_{k,1}$ and $\tau_{k,2}$ are diagonal elements of the diagonal precision matrix, as well as the intensity of the Poission distribution that S^n is sampled from, λ_k .

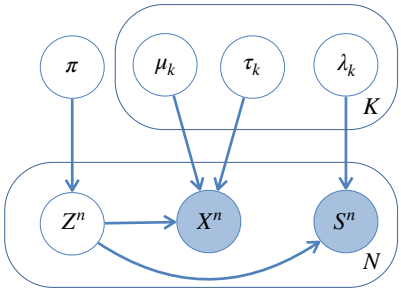

Figure 1: Mixture of components modeling location and strengths of earthquakes associated with a super-epicentra. In the figure, $\mu_k = (\mu_{k,1}, \mu_{k,2})$ and $\tau_k = (\tau_{k,1}, \tau_{k,2})$.

Question 2.6.18: *Derive an EM algorithm for the model.*

Question 2.6.19: *Implement your EM algorithm.*

Question 2.6.20: *Apply it to the data provided separately, give an account of the success, and provide visualizations for a couple of examples.*

2.7 Super epicentra – VI

As in Problem 2.6 , we have seismographic from an area with frequent earthquakes emanating from K super epicentra. In fact, the core of the present model is the model described in Problem 2.6 but now the parameters also have conjugate prior distributions. As shown in Figure 2, the present model has the following prior distributions. All the hyperparameters $(\alpha, C, \mu, \alpha', \beta', \alpha_0, \beta_0)$ are known, shaded nodes are observed.

1. π has a $\mathrm{Dir}(\alpha)$ prior.
2. $\tau_{k,i}$ has a $\mathrm{Ga}(\alpha', \beta')$ prior.
3. $\mu_{k,i}$ has a $\mathcal{N}(\mu, (C\tau_{k,i})^{-1})$ prior.
4. λ_k has a $\mathrm{Ga}(\alpha_0, \beta_0)$ prior.

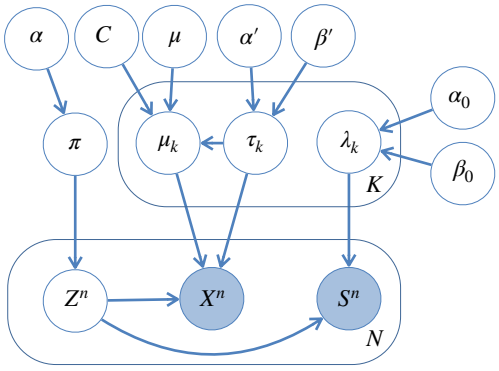

Figure 2: The K super epicentra model with priors.

Question 2.7.21: *Derive a VI algorithm that estimates the posterior distribution for this model.*

2.8 Sampling from a tree GM

Let T be a binary tree, with leaves $L(T)$, and consider the graphical model T, Θ where $T, \Theta = \{\theta_v : v \in V(T)\}$, the variable $\{X_v : v \in V(T)\}$ takes on values in $[K]$, and for each $v \in V(T)$ the CPD $\theta_v = p(X_v | x_{\mathrm{pa}(v)})$ is a categorical. The observable variables O is $\{X_l : l \in L(T)\}$. Consider the unique

distribution $q(x)$ that (i) is defined on the $X \in O$ such that $\sum_{l \in L(T)} X_l$ is odd and (ii) is proportional to $p(X|T, \Theta)$. The proportionality means that there is a constant C such that $q(X) = Cp(X|T, \Theta)$.

Building on the "standard" Dynamic Programming (DP) algorithm for computing the likelihood of a tree DGM, derive a polynomial time sampling algorithm (which excludes rejection sampling, since its running time is unbounded) that, for given T, Θ, samples an $x \in O$ from $q(x)$. Start out by designing a bottom-up (i.e., leaves-to-root) DP algorithm by first identifying recursions pertinent for obtaining the probability, p, of getting an odd-sum output below vertex v of T given that $X_v = k$. As easily notice the probability of an even-sum output given the same condition is $1 - p$. Then design a top-down (i.e., root-to-leaves) sampling algorithm that samples from q.

Question 2.8.22: *Derive these algorithms.*

Question 2.8.23: *Implement your bottom up DP algorithm for the probability of generating an odd sum output.*

Question 2.8.24: *Implement your sampling algorithm.*

Question 2.8.25: *Apply your algorithm to the graphical model and data provided separately.*

2.9 Failing components VI

This problem is concerned with components connected to each other as vertices in a directed graph. However, to simplify, we let the components be elements in a $R \times C$-grid. Let the neighbours of a row index r, be the set $N(r) = \{r' : |r' - r| \leq 1 \text{ or } \{r, r'\} = \{1, R\}\}$, and notice that every such index has three neighbours.

In the first column, i.e., $c = 1$, exactly one uniformly sampled element fails and, if $c < C$, failure of element (r, c) results with equal probability in failure of one of the grid elements in $\{(r', c+1) : r' \in N(r)\}$, i.e., exactly one of them fails, and no other element of that column fails, i.e., we have a Markov property. If element (r, c) has not failed, its value, $X_{r,c}$, is sampled from a Bernoulli distribution with parameter $\theta_{r,c}$ that has a Beta prior with hyper-parameters a, b. Moreover, If element (r, c) has failed, $X_{r,c}$ is sampled from a Bernoulli distribution with parameter θ_f that has a Beta prior with hyper-parameters a_f, b_f.

Notice that the Beta distribution is a conjugate prior to the Bernoulli distribution. Introduce hidden variables $Z = \{Z_c : c \in [C]\}$ such that $Z_c = r$ indicates that the failure in column c is in row r, and no other row. Consider a variational distribution q, over the variables Z and $\Theta = \theta_f \cup \{\theta_{r,c} : r \in [R], c \in [C]\}$, such that $q(Z, \Theta) = q(Z)q(\Theta)$.

Question 2.9.26: *Derive a VI algorithm that estimates the posterior distribution for this model*