DD2434 Machine Learning, Advanced Course
Assignment 3, 2020 (V1.1)

Jens Lagergren

Deadline, see Canvas

Read this before starting

There are some commonalities between the problems and they cover different aspects of the course and very in difficulty, 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.

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,

- **C** Two correctly solved problems of 3.1–3.6
- **B** Four correctly solved problems of 3.1–3.6
- **A** Five correctly solved problems of 3.1–3.6

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E, which makes it meaningless to hand in late solutions for this assignment.

Good Luck!

3.1 Easier EM for Advertisements

We have data concerning whether or not each of N readers, $r_1, \cdots, r_N$, have clicked on each of M advertisements, $a_1, \cdots, a_M$, in each of L journal editions, $e_1, \cdots, e_L$. The indices of the advertisement a_m that occur in an edition e_l is given by the set $A(l)$. We suspect that the readers can be grouped into C groups, the editions can be grouped into D groups, the advertisements can be grouped into F sentiment groups, and, the probability that a reader clicks on an advertisement in an edition read by the reader depends only on the readers group Z^r_n, the sentiment group of the edition Z^e_l, and the advertisements group Z^a_m, i.e., it is

$$\psi_{c,d,f} | Z^r_n = c, Z^e_l = d, Z^a_m = f.$$

Moreover, these latent class variables have categorical distributions $(\theta^r, \theta^a, \theta^e)$. Finally each reader reads all editions. Design an EM-algorithm for obtaining MLE of all the models parameters

$$\Omega = (\theta^r, \theta^a, \theta^e, \{\psi_{c,d,f} : c \in [C], d \in [D], f \in [F]\}).$$

3.2 Hard EM for Advertisements

This model is obtained by, to the model in 3.1 , for each reader r_n and edition e_l, adding a latent variable Z^{re}_{nl}, which is 1 when reader r_n has read edition e_l and 0 otherwise. A reader can only click an advertisement in an edition that the reader has read. Moreover, each Z^{re}_{nl} follows a Bernoulli distribution with parameter θ. So, now

$$\Omega = (\theta, \theta^r, \theta^a, \theta^e, \{\psi_{c,d,f} : c \in [C], d \in [D], f \in [F]\}).$$

Hint, make sure that you solved 3.1 properly, and work with vectors

$$X_{n,l} = X_{n,l,1}, \ldots, X_{n,l,M},$$

where $X_{n,l,m} = 0$ if $m \notin A(l)$. It may be a good idea to first resolve how to maximize θ.

3.3 Complicated likelihood for leaky units on a tree

Consider the following model. A binary tree T has random variables associated with its vertices. A vertex u has an observable variable X_u and a latent class variable Z_u. Each class $c \in [C]$ has a normal distribution $N(\mu_k, \sigma^2)$. If $Z_u = c$ and for the three neighbors of u, let us call them v_1, v_2, and v_3, the latent class variable satisfies $Z_{v_i} = c_i$, then

$$p(X_u) = N(X_u | (1-\alpha)\mu_c + \sum_{i \in [3]} \frac{1}{3} \alpha \mu_{c_i}, \sigma^2).$$

The class variables are iid, each follows the categorical distribution π. Provide a linear time algorithm that computes $P(X|T, M, \sigma, \alpha, \pi)$ when given a tree T (with vertices $V(T)$), observable variables for its vertices $X = \{X_v : v \in V(T)\}$, and parameters $M = \{\mu_c : c \in [C]\}, \sigma, \alpha$.

3.4 Easier VI for Covid-19

We have a workplace with K workers, $w_1, \cdots, w_K$, where we monitor Covid-19. Any day d each worker w_k is either non-infected, infected, or has antibodies, i.e., there is a latent variable Z^k_d with a value in $\{n, i, a\}$, with the obvious interpretation. A non-infected individual becomes with probability ι infected the day after the individual has had contact with an infected individual (and though only one such contact may occur with any single infected individual during a day, an uninfected may have contact with several infected during a day). An individual that becomes infected on day d is aware of the infection, and will on day $d + 9$ get antibodies with probability α. Otherwise, the individual remains/returns to the non-infected state. An infected individual stays at home with probability σ and is otherwise present at the workplace. We have access to a contact graph G_d and an absence table A_d for each day $d \in [D]$, $A_d^k = 1$ if worker k is home on day d and otherwise 0. Consider $G = G_d$ as given so the joint is

$$p(A, Z, \Omega | G),$$

where $\Omega = (\iota, \alpha, \sigma)$. There are beta priors on Bernoulli parameters ι, α, and σ. No other reasons than Covid-19 makes any worker stay at home. On day one w_1 is infected and all other workers are non-infected. Let $Z^k = Z_1^k, ..Z_D^k$ and $Z = Z^1, ...Z^K$. Design a VI algorithm for approximating the posterior probability over Z and use the VI distribution

$$q(Z) = \prod_{d,k} q(Z_d^k).$$

Hint: extend the latent variable so that you also can keep track of how long an individual has been infected.

3.5 Hard VI for Covid-19

For the above model, i.e., in 3.4 , design a VI algorithm for approximating the posterior probability over Z, but in this case, use the VI distribution

$$q(Z) = \prod_{k} q(Z^k).$$

Hint notice that Z^k has a Markov property given Z^{-k}, i.e., given Z^{-k} and Z_d^k the variables $Z_{1:d-1}^k$ are independent of $Z_{d+1:D}^k$.

3.6 Spectral Graph Analysis

In this problem, you should solve each of the follwing three subproblems.

- Let $G = (V, E)$ be an undirected d-regular graph, let A be the adjacency matrix of G, and let $L = I - \frac{1}{d}A$ be the normalized Laplacian of G. Prove that for any vector $\mathbf{x} \in \mathbb{R}^{|V|}$ it is

$$\mathbf{x}^T L \mathbf{x} = \frac{1}{d} \sum_{(u,v) \in E} (x_u - x_v)^2. \tag{1}$$

- Show that the normalized Laplacian is a positive semidefinite matrix.
- Assume that we find a non-trivial vector $\mathbf{x}_*$ that minimizes the expression $\mathbf{x}^T L \mathbf{x}$. First explain what non-trivial means. Second explain how $\mathbf{x}_*$ can be used as an embedding of the vertices of the graph into the real line. Use Equation (1) to justify the claim that $\mathbf{x}_*$ provides a meaningful embedding.