

Assignment 2

Statistical Methods in Applied Computer Science

DD2447

Jim Holmström 890503-7571
jimho@kth.se

November 24, 2012

Exercise A1 Compute $p(D|T \in \text{Polytree})$ with Bernoulli CPD's
Show how to compute $p(D|T)$ where T is a DGM which is a polytree and $D = \{x_1, \dots, x_N\}$ (and x_i is an assignment of values to all variables of T). Assume that all variables are binary and all CPD's Bernoulli.

Solution.

Theorem 0.1. Chain rule

$$p(\bigcap_k x_k) = \prod_k p(x_k | \bigcap_{j:j < k-1} x_j) \quad (1)$$

Proof. We can extend the rule

$$p(x_i, x_j) = p(x_i | x_j) p(x_j) \quad (2)$$

since x_i, x_j is just events we might as well have $x_j = \bigcap_{j'} x_{j'}$ which makes the expression look like this

$$p(x_i \cap (\bigcap_{j'} x_{j'})) = p(x_i | \bigcap_{j'} x_{j'}) p(\bigcap_{j'} x_{j'}) \quad (3)$$

by tail-recursion on $p(\bigcap_{j'} x_{j'}) = p(x_a \cap (\bigcap_{j' \neq a} x_{j'}))$ we will exhaust all j' which leaves us with base-case $p(x_a)$. Writing out the entire trace of the recursion will give us the wanted expression for $p(\bigcap_k x_k)$. \square

The goal is to break down $p(D|T)$ into CPD's for each node

$$p(D|T) = \phi(\{p(x_t|\cdot)\}_t) \quad (4)$$

which gives us an explicit expression for the $p(D|T)$.

polytree \Rightarrow no directed cycles \Rightarrow topological sortable chainrule on topologically sorted nodes show that we will always have all the parents in $p(\text{xt}|\cdot)$ since it's topologically sorted the CI assumption removes all non parents \square

Exercise A2 Marginalize over non-observed variables

Assume instead that each x_i is an assignment to a subset of the variables say O . Show how to marginalize over $V \setminus O$ (i.e., the non-observed variables).

Solution.

Exercise 11.3 EM for the mixtures of Bernoullis

- Show that the M step for ML estimation of a mixture of Bernoullis is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}} \quad (5)$$

- Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(\alpha, \beta)$ prior is given by

$$\mu_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{(\sum_i r_{ik}) + \alpha + \beta - 2} \quad (6)$$

Solution. The M step is to optimize the auxiliary function Q with respect to π, θ' . Q is the expected posterior log-likelihood¹ with respect to the last parameter θ and the observed data D .² The expression for this is

$$Q(\theta', \theta) = E [\log \mathcal{L}_{MAP}(\theta') | D, \theta] = E [\log (\mathcal{L}(\theta') p(\theta')) | D, \theta] = \quad (7)$$

$$= E [\ell(\theta') + \log p(\theta') | D, \theta] = E [\ell(\theta') | D, \theta] + \log p(\theta') \quad (8)$$

and to derive this expression we introduce the latent variable z_i which corresponds to the hidden or missing variables which basically is the *r.v.* for how x_i belongs to the class k .³ We start with MAP and then derive ML by setting a uniform prior.

$$Q(\theta', \theta) = E \left[\sum_i \log p(x_i, z_i | \theta') \right] + p(\theta') = \quad (9)$$

¹In the book it's actually just log-likelihood, but this will work in the same way instead of using $Q(\theta', \theta) + \log p(\theta')$ without derivation, we will use posterior-Q as Q instead.

²It can be shown that Q is so that the new parameters is always better or as good as the last one, but exclude the proof for this since it's not needed by the exercise.

³Responsibility $r_{ik} \triangleq p(z_i = k | x_i, \theta)$

since E is linear and we can factor on the different classes and the factors become given with the parameters (π_k, θ'_k) of class k :

$$= \sum_i \left[E \log \left(\prod_k (\pi_k p(x_i | \theta'_k))^{\mathbb{I}(z_i=k)} \right) \right] + \log \prod_k p(\theta'_k) = \quad (10)$$

then log the inner parts and let E operate on the expression

$$= \sum_i \sum_k \left[E [\mathbb{I}(z_i = k)] \log [p(x_i | \theta'_k)] \right] + \sum_k \log p(\theta'_k) = \quad (11)$$

then we have that the expected $E [\mathbb{I}(z_i = k)]$ will be $p(z_i = k | x_i, \theta) = r_{ik}$ which is the expected class-belonging of x_i given the previous parameters θ . The log product rule gives us

$$= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k \{r_{ik} \log p(x_i | \theta'_k)\} + \sum_k \log p(\theta'_k) \quad (12)$$

Now since we have that $\sum_{i,k} r_{ik} \log \pi_k \perp \sum_{i,k} \{r_{ik} \log p(x_i | \theta'_k)\} + \sum_k \log p(\theta'_k)$ we can optimize the parameters π_k and θ'_k separately and only the θ'_k -term is asked for in the exercise which we denote $\ell(\theta'_k)$.

The model parameters in this case is denoted by μ_k which is a vector and in index-notation μ_{kj} .⁴

We start with the MAP

$$\hat{\mu}'_k = \underset{\mu'_k}{argmax} \left\{ \sum_i \{r_{ik} \log p(x_i | \mu'_k)\} + \log p(\mu'_k) \right\} \quad (13)$$

which we find by $\frac{\partial \ell(\mu'_k)}{\partial \mu'_k} = 0$, where $p(x_i | \mu'_k) = \prod_j p(x_{ij} | \mu'_{kj})$ is the multivariate Bernoulli⁵ distribution for class k . The priori in the same way is $p(\mu'_k) = \prod_j p(\mu'_{kj})$. In our case for MAP this is $\beta(\alpha, \beta) = \frac{\mu'_{kj}{}^{\alpha-1} (1-\mu'_{kj})^{\beta-1}}{B(\alpha, \beta)}$.

⁴Not using Einstein notation for tensor product.

⁵Often called multinoulli.

$$\frac{\partial(\ell(\mu'_k))}{\partial\mu'_{kj}} = \frac{\partial\sum_i \left\{ r_{ik} \log \prod_{j'} p(x_{ij'} | \mu'_{kj'}) \right\} + \log \prod_{j'} p(\mu'_{kj'})}{\partial\mu'_{kj}} = (14)$$

$$= \frac{\partial\sum_{i,j'} \left\{ r_{ik} \log p(x_{ij'} | \mu'_{kj'}) \right\} + \sum_{j'} \log p(\mu'_{kj'})}{\partial\mu'_{kj}} = (15)$$

$$= \left\{ \frac{\partial \sum_{j \neq j'} (\cdot)}{\partial\mu'_{kj}} = 0 \right\} = \frac{\partial\sum_i \left\{ r_{ik} \log \left(\mu'_{kj} x_{ij} (1 - \mu'_{kj})^{(1-x_{ij})} \right) \right\} + \log \frac{\mu'_{kj}{}^{\alpha-1} (1-\mu'_{kj})^{\beta-1}}{B(\alpha,\beta)}}{\partial\mu'_{kj}} = (16)$$

$$= \frac{\partial\sum_i r_{ik} \left(x_{ij} \log \mu'_{kj} + (1 - x_{ij}) \log(1 - \mu'_{kj}) \right)}{\partial\mu'_{kj}} + (17)$$

$$+ \frac{\partial(\alpha - 1) \log \mu'_{kj} + (\beta - 1) \log(1 - \mu'_{kj}) - \log B}{\partial\mu'_{kj}} = (18)$$

$$= \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{\mu'_{kj}} - \frac{(\sum_i r_{ik} (1 - x_{ij})) + \beta - 1}{1 - \mu'_{kj}} = (19)$$

$$= \frac{(1 - \mu'_{kj}) (\sum_i r_{ik} x_{ij}) - \mu'_{kj} (\sum_i r_{ik} (1 - x_{ij}))}{\mu'_{kj} (1 - \mu'_{kj})} + (20)$$

$$+ \frac{(1 - \mu'_{kj}) (\alpha - 1) - \mu'_{kj} (\beta - 1)}{\mu'_{kj} (1 - \mu'_{kj})} = (21)$$

$$= \frac{(\sum_i r_{ik} x_{ij}) - \mu'_{kj} (\sum_i r_{ik}) + (\alpha - 1) - \mu'_{kj} (\alpha + \beta - 2)}{\mu'_{kj} (1 - \mu'_{kj})} (22)$$

and now find the zero of this expression

$$\frac{(\sum_i r_{ik} x_{ij}) - \mu'_{kj} (\sum_i r_{ik}) + (\alpha - 1) - \mu'_{kj} (\alpha + \beta - 2)}{\mu'_{kj} (1 - \mu'_{kj})} = 0 \quad (23)$$

$$(\sum_i r_{ik} x_{ij}) - \mu'_{kj} (\sum_i r_{ik}) + (\alpha - 1) - \mu'_{kj} (\alpha + \beta - 2) = 0 \quad (24)$$

$$\mu'_{kj} [(\sum_i r_{ik}) + \alpha + \beta - 2] = (\sum_i r_{ik} x_{ij}) + \alpha - 1 \quad (25)$$

$$\mu'_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{(\sum_i r_{ik}) + \alpha + \beta - 2} \quad (26)$$

which is the same as the equation (6). \square

By just looking at the expression for $\beta(\cdot)$ we see that

$$\beta(1,1) = U(0,1) \quad (27)$$

and knowing that ML is a MAP with a uniform priori we can just set $\alpha, \beta = 1$ in the derived expression to get the ML.

$$\mu'_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + 1 - 1}{(\sum_i r_{ik}) + 1 + 1 - 2} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}} \quad (28)$$

which is the same as the equation (5). □