Bayesian Models

Matheus Rosso

Contents

1 Bayesian inference

The Bayesian inference seeks to estimate a probability distribution for a vector of parameters θ from a given dataset $\mathcal{D}$, $P(\theta|\mathcal{D})$. The Bayesian approach follows from applying the Bayes rule to calculate $P(\theta|\mathcal{D})$.

Here, the elements involved when using Bayes rule are: i) a **prior distribution** assumed for θ, $P(\theta)$; ii) a **probability distribution for the data**, $P(\mathcal{D}|\theta)$.

In these notes, the conditional setting is considered, so that $P(\mathcal{D}) = P(Y|X)$, where Y is a response variable and X is a vector of inputs which relates to Y through θ.

Then, applying Bayes rule:

$$P(\theta|\mathcal{D}) = \frac{P(\theta,\mathcal{D})}{P(\mathcal{D})} = \frac{P(\mathcal{D},\theta)}{\displaystyle\int P(\mathcal{D},\theta)d\theta} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\displaystyle\int P(\mathcal{D}|\theta)P(\theta)d\theta} \tag{1}$$

This expression is the main concern when Bayesian models are to be estimated, and it is defined as the **posterior distribution** of θ, since it considers a given dataset for modeling the density of the parameters.

The expression in (1) is changed slightly by the inclusion of a hyper-parameter α, which determines the prior distribution of θ, $P(\theta|\alpha)$. To this hyper-parameter a **hyper-prior** is defined, $P(\alpha)$. Consequently, (1) changes to:

$$P(\theta,\alpha|\mathcal{D}) = \frac{P(\theta,\alpha,\mathcal{D})}{P(\mathcal{D})} = \frac{P(\mathcal{D},\theta,\alpha)}{\displaystyle\int\int P(\mathcal{D},\theta,\alpha)d\theta d\alpha} = \frac{P(\mathcal{D}|\theta,\alpha)P(\theta,\alpha)}{\displaystyle\int\int P(\mathcal{D}|\theta,\alpha)P(\theta,\alpha)d\theta d\alpha}$$

$$P(\theta,\alpha|\mathcal{D}) = \frac{P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha)}{\displaystyle\int\int P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha)d\theta d\alpha} = \frac{P(\mathcal{D}|\theta)P(\theta|\alpha)P(\alpha)}{\displaystyle\int\int P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha)d\theta d\alpha} \tag{2}$$

<u>Note:</u> $P(\mathcal{D}|\theta,\alpha)$ apparently can be replaced by $P(\mathcal{D}|\theta)$ in (2), since $\mathcal{D}$ and α should be independent once the data density is controlled for θ.

Sometimes, the expression in (2) is summarized by the following notation:

$$P(\theta,\alpha|\mathcal{D}) \propto P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha) \tag{3}$$

Since the main focus is pointed to θ, and not to α, without any further assumption $P(\theta|\mathcal{D})$ can be obtained by marginalizing (2) over the hyper-parameter α:

$$P(\theta|\mathcal{D}) = \int P(\theta,\alpha|\mathcal{D})d\alpha \tag{4}$$

Finally, once determined (4), empirical motivations lead to the definition of the **predictive density** $P(y|x,\mathcal{D})$:

$$P(y|x,\mathcal{D}) = \int P(y|x,\theta).P(\theta|\mathcal{D})d\theta \tag{5}$$

Where (y,x) refers to a new data point and $P(y|x,\theta)$ is the probability distribution that constitutes $P(\mathcal{D}|\theta) = P(Y|X,\theta)$, which aggregates $P(y_i|x_i,\theta)$ overall observations $i \in \{1,2,...,N\}$.

It is convenient to briefly comment over the intuition behind the Bayesian treatment for the posterior distribution estimation, $P(\theta|\mathcal{D})$. Additional to statistical structure for the distribution of θ, given by the prior and the hyper-prior distributions, $P(\theta|\alpha)$ and $P(\alpha)$, respectively, Bayesian inference learns from the available data, which is also referenced by a statistical distribution $P(\mathcal{D}|\theta)$. Therefore, the Bayesian approach starts from assumed prior distributions for the parameters and evolves towards to posterior distributions that are based on the data.

2 Variational Bayesian methods

The evaluation of (2) is not feasible due to the intractability of assessing $P(\mathcal{D})$. Notwithstanding, it is possible to approximate (2) through variational methods based on the *calculus of variations*. This set of techniques differentiates functions $L(.)$ by another functions $Q(.)$ in the attempt of optimizing the value of $L(.)$.

In the context of Bayesian inference, the objective function follows from the **Kullback-Leibler divergence**, $d_{KL}(Q||P)$. This statistic is defined by:

$$d_{KL}(Q||P) = \int\int Q(\theta,\alpha).\log\Big(\frac{Q(\theta,\alpha)}{P(\theta,\alpha|\mathcal{D})}\Big)d\theta d\alpha \tag{6}$$

Where:

$$Q(\theta,\alpha) \approx P(\theta,\alpha|\mathcal{D}) \tag{7}$$

Manipulating expression (6):

$$d_{KL}(Q||P) = \int\int Q(\theta,\alpha)\log\Big(\frac{Q(\theta,\alpha)}{P(\theta,\alpha,\mathcal{D})/P(\mathcal{D})}\Big)d\theta d\alpha$$

$$d_{KL}(Q||P) = \int\int Q(\theta,\alpha)\log\Big(\frac{Q(\theta,\alpha)}{P(\theta,\alpha,\mathcal{D})}\Big)d\theta d\alpha + \log(P(\mathcal{D}))$$

$$\log(P(\mathcal{D})) = d_{KL}(Q||P) - \int\int Q(\theta,\alpha)\log\Big(\frac{Q(\theta,\alpha)}{P(\theta,\alpha,\mathcal{D})}\Big)d\theta d\alpha$$

$$\log(P(\mathcal{D})) = d_{KL}(Q||P) + \int\int Q(\theta,\alpha)\log\Big(\frac{P(\theta,\alpha,\mathcal{D})}{Q(\theta,\alpha)}\Big)d\theta d\alpha$$

$$\log(P(\mathcal{D})) = d_{KL}(Q||P) + L(Q) \tag{8}$$

$L(Q)$ is named as **variational lower bound**, since it provides a lower bound for the *model evidence*, $\log(P(\mathcal{D}))$:

$$\log(P(\mathcal{D})) \geq L(Q) \tag{9}$$

If the objective is to find $Q(\theta,\alpha)$ that best approximates $P(\theta,\alpha|\mathcal{D})$, then $Q(.)$ should minimize the Kullback-Leibler divergence, $d_{KL}(Q||P)$. Therefore, the variational problem turns out to be finding a function $Q(.)$ that maximizes the lower bound $L(Q)$.

It is convenient to restrict the family of possible solutions to the maximization of $L(Q)$ to functions $Q(.)$ such that:

$$Q(Z) = \prod_{j=1}^{M} q_j(Z_j) = q_\theta(\theta)q_\alpha(\alpha) \tag{10}$$

Where $Z = (\theta,\alpha)$ and M is the total number of parameters in Z.

Given (10), the following decomposition of $L(Q)$ holds:

$$L(Q) = \int q_j \log(\tilde{P}(Z_j,\mathcal{D}))dZ_j - \int q_j \log(q_j)dZ_j + const \tag{11}$$

Where $\tilde{P}(Z,\mathcal{D})$ is such that:

$$\log(\tilde{P}(Z_j,\mathcal{D})) = E_{i\neq j}(\log(P(Z,\mathcal{D}))) + const \tag{12}$$

From (11):

$$L(Q) = \int q_j \log\Big(\frac{\tilde{P}(Z_j,\mathcal{D})}{q_j}\Big)dZ_j + const = -\int q_j \log\Big(\frac{q_j}{\tilde{P}(Z_j,\mathcal{D})}\Big)dZ_j + const$$

$$L(Q) = -d_{KL}(q_j||\tilde{P}(Z_j,\mathcal{D})) + const \tag{13}$$

Therefore, q_j must minimize $d_{KL}(q_j||\tilde{P}(Z_j,\mathcal{D}))$, which implies that:

$$q_j = \tilde{P}(Z,\mathcal{D})$$

Given (12):

$$\log(q_j(Z_j)) = E_{i\neq j}(\log(P(Z,\mathcal{D})) + const = E_{i\neq j}(\log(P(\mathcal{D}|Z)P(Z))) + const$$

$$\log(q_j(Z_j)) = E_{i\neq j}(\log(P(\mathcal{D}|\theta,\alpha)P(\theta,\alpha)))) + const$$

$$\log(q_j(Z_j)) = E_{i\neq j}(\log(P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha))) + const \tag{14}$$

Aggregating overall j such that $Z_j \in \theta$, and overall j' such that $Z_{j'} \in \alpha$:

$$\log(q_\theta(\theta)) = E_\alpha(\log(P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha))) + const \tag{15}$$

$$\log(q_\alpha(\alpha)) = E_\theta(\log(P(\mathcal{D}|\theta,\alpha)P(\theta|\alpha)P(\alpha))) + const \tag{16}$$

The expressions (15) and (16) provide variational approximations for the posterior probability $P(\theta,\alpha|\mathcal{D})$. Thus, the assumption of (10) makes unnecessary the marginalization for obtaining $Q(\theta)$ from $Q(\theta,\alpha)$, such that in (4).

3 Bayesian logistic regression

Given $y_i \in \{0,1\}$, a vector of inputs $(1xp)$, x_i, and coefficients θ $(px1)$, then the data conditional density is given by:

$$P(y_i|x_i,\theta) = \Big(\frac{1}{1+\exp(x_i\theta)}\Big)^{y_i}\Big(1-\frac{1}{1+\exp(x_i\theta)}\Big)^{1-y_i} = \sigma(x_i\theta)^{y_i}(1-\sigma(x_\theta))^{1-y_i} \tag{17}$$

Instead of modeling (17) by maximum likelihood, so that $\hat{\theta}$ can be obtained with associated variance and additional inference entities (**frequentist approach**), the **Bayesian approach** uses the data conditional density together with a prior distribution over θ to produce a posterior distribution for the coefficients, as discussed in section 1.

The *prior distribution* of θ may be given by:

$$P(\theta|\alpha) = \mathcal{N}(0,\alpha^{-1}I_p) \tag{18}$$

Where I_p is the (pxp) identity matrix and α is a *hyper-parameter* whose distribution is named *hyper-prior* following:

$$P(\alpha) = Gam(a_0,b_0) \tag{19}$$

See section 4 for extensions of (18) and (19).

The *posterior distribution* of θ and α is defined under the Bayes rule, as developed in equations (1) and (2) from section 1:

$$P(\theta, \alpha|\mathcal{D}) = \frac{P(Y|X, \theta).P(\theta|\alpha).P(\alpha)}{\int\int P(Y|X, \theta, \alpha).P(\theta|\alpha).P(\alpha)d\theta d\alpha} \tag{20}$$

Where $\mathcal{D} = Y|X$ refers to the data available and $P(Y|X, \theta) = \prod_i P(y_i|x_i, \theta)$.

Section 2 brought the necessity of approximating (20) through variational methods, which maximizes a lower bound $L(Q)$ to define the variational posterior $Q(\theta, \alpha)$. Given the non-conjugacy between $P(Y|X, \theta)$ and $P(\theta|\alpha)$, since the first is not from the exponential family, an additional approximation is implemented in the context of Bayesian logistic regression model.

The variational approach that leads to (14) is global, since no local restriction has been imposed. In order to construct an approximation to $P(Y|X, \theta)$ that conjugates with $P(\theta|\alpha)$, a *local variational method* can be implemented. Instead of considering $P(y_i|x_i, \theta)$ that uses the sigmoid function $\sigma(.)$, the following local approximation is adopted:

$$\sigma(z) \geq \sigma(\xi). \exp\Big(\frac{(z - \xi)}{2} + \lambda(\xi)(z^2 - \xi^2)\Big) \tag{21}$$

Where ξ is the *variational parameter* and $\lambda(\xi)$ is given by:

$$\lambda(\xi) = -\frac{1}{2\xi}\Big(\sigma(\xi) - \frac{1}{2}\Big) \tag{22}$$

Given the expression (21), the following approximation to $P(Y|X, \theta)$ holds:

$$P(Y|X, \theta) = \prod_{i=1}^{N} P(y_i|x_i, \theta) \geq \prod_{i=1}^{N} \sigma(\xi_i) \exp\Big(y_i x_i \theta - \frac{(x_i\theta + \xi_i)}{2} + \lambda(\xi_i)((x_i\theta^2) - \xi_i^2)\Big) \tag{23}$$

The expression in the right side of the inequality is defined as $h(\theta, \xi)$.

In order to get the variational solution for approximating $P(\theta|\mathcal{D})$, it is necessary to impose another restriction to the global lower boundary, which is obtained from the developments that lead to expression (8):

$$L(Q) = \int\int Q(\theta, \alpha) \log\Big(\frac{P(Y|X, \theta)P(\theta|\alpha)P(\alpha)}{Q(\theta, \alpha)}\Big) d\theta d\alpha \tag{24}$$

Replacing $P(Y|X, \theta)$ by $h(\theta, \xi)$ in the lower boundary expression (24) implies in the following new lower boundary:

$$\log(P(Y|X)) \geq L(Q) \geq L(Q, \xi) = \int\int Q(\theta, \alpha) \log\Big(\frac{h(\theta, \xi)P(\theta|\alpha)P(\alpha)}{Q(\theta, \alpha)}\Big) d\theta d\alpha \tag{25}$$

Since $P(\theta, \alpha, \mathcal{D})$ was approximated through $h(\theta, \xi)P(\theta|\alpha)P(\alpha)$, $q_\theta(\theta)$ and $q_\alpha(\alpha)$ follow from (15) and (16) using $h(\theta, \xi)$ instead of $P(Y|X, \theta)$:

$$\log(q_\theta(\theta)) = E_\alpha(\log(h(\theta, \xi)P(\theta|\alpha)P(\alpha))) + const \tag{26}$$

$$\log(q_\alpha(\alpha)) = E_\theta(\log(h(\theta, \xi)P(\theta|\alpha)P(\alpha))) + const \tag{27}$$

To solve equations (26) and (27), $h(\theta, \xi)$ uses (23), while $P(\theta|\alpha)$ and $P(\alpha)$ make use of the assumptions (18) and (19).

The solutions point to $q_\theta(\theta)$ normally distributed, $N(\mu_N, \Sigma_N)$, with parameters:

$$\mu_N = \Sigma_N \sum_{i=1}^{N} (y_n - (1/2))x_i \tag{28}$$

$$\Sigma_N^{-1} = E_\alpha(\alpha).I - 2\sum_{i=1}^{N} \lambda(\xi_i)x_i x_i^T \tag{29}$$

For $q_\alpha(\alpha)$, it is found that it follows a Gamma distribution $\Gamma(a_N, b_N)$:

$$a_N = a_0 + \frac{p}{2} \tag{30}$$

$$b_N = b_0 + \frac{1}{2}E_\theta(\theta^T\theta) \tag{31}$$

Since (29) depends on ξ, it is necessary to define the variational parameters by maximizing the lower boundary $L(Q, \xi)$ on ξ. From (25):

$$\int\int Q(\theta, \alpha)\log\Big(\frac{h(\theta, \xi)P(\theta|\alpha)P(\alpha)}{Q(\theta, \alpha)}\Big)d\theta d\alpha = \int q_\theta(\theta)\log(h(\theta, \xi))d\theta + const \tag{32}$$

It holds, then, the following solution to ξ_i:

$$\xi_{i,new}^2 = x_i^T(\Sigma_N + \mu_N\mu_N^T)x_i \tag{33}$$

Finally, to define (28), (29), (30), (31) and (33), it is applied the **EM algorithm**, which begins by assigning an initial guess for $\xi_i^{(0)}$, then implying values for $\mu_N^{(1)}$, Σ_N^1, $a_N^{(1)}$ and $b_N^{(1)}$. Making use of these, ξ_i^0 is updated to $\xi_i^{(1)}$, allowing new definitions for $\mu_N^{(2)}$, $\Sigma_N^{(2)}$, $a_N^{(2)}$ and $b_N^{(2)}$. This process continues until some convergence criterion is satisfied.

Given the approximation $q_\theta(\theta)$ for $P(\theta|\mathcal{D})$, the *predictive density* can also be approximated:

$$P(y=1|x,\mathcal{D}) = \int P(y=1,\theta|x,\mathcal{D})d\theta = \int P(y=1|x,\theta).P(\theta|\mathcal{D})d\theta \approx \int P(y=1|x,\theta).q_\theta(\theta)d\theta \tag{34}$$

Again, the non-conjugacy between $P(y=1|x,\theta)$ and $q_\theta(\theta)$ requires new approximation. However, some alternative methods are available, such as:

1. **Monte Carlo simulations**: given the posterior $P(\theta|\mathcal{D})$, or its approximation $Q(\theta)$, M different samples $\tilde{\theta}$ are extracted from it and, thus, different probabilities $P(y=1|x,\tilde{\theta}_m)$ are calculated. Then, the average of such values consists on the estimation of $P(y=1|x,\mathcal{D})$ for the new data point:

$$P(y=1|x,\mathcal{D}) = (1/M)\sum_{m=1}^{M} P(y=1|x,\tilde{\theta}_m)$$

2. **Maximum a posteriori (MAP)**: the posterior $P(\theta|\mathcal{D})$, or its approximation $Q(\theta)$, is maximized over θ. The solution $\hat{\theta}^{max}$ implies in the following estimation of $P(y=1|x,\mathcal{D})$:

$$P(y=1|x,\mathcal{D}) = P(y=1|x,\hat{\theta}^{max})$$

Note: the density $P(\mathcal{D})$ follows from (17):

$$P(\mathcal{D}) = P(Y|X) = \prod_{i=1}^{N} P(y_i|x_i,\theta)$$

Also, can be derived from:

$$P(\mathcal{D}) = P(Y|X) = \int\int P(Y|X,\theta).P(\theta|\alpha).P(\alpha)d\theta d\alpha$$

4 Hierarchical Bayesian logistic regression

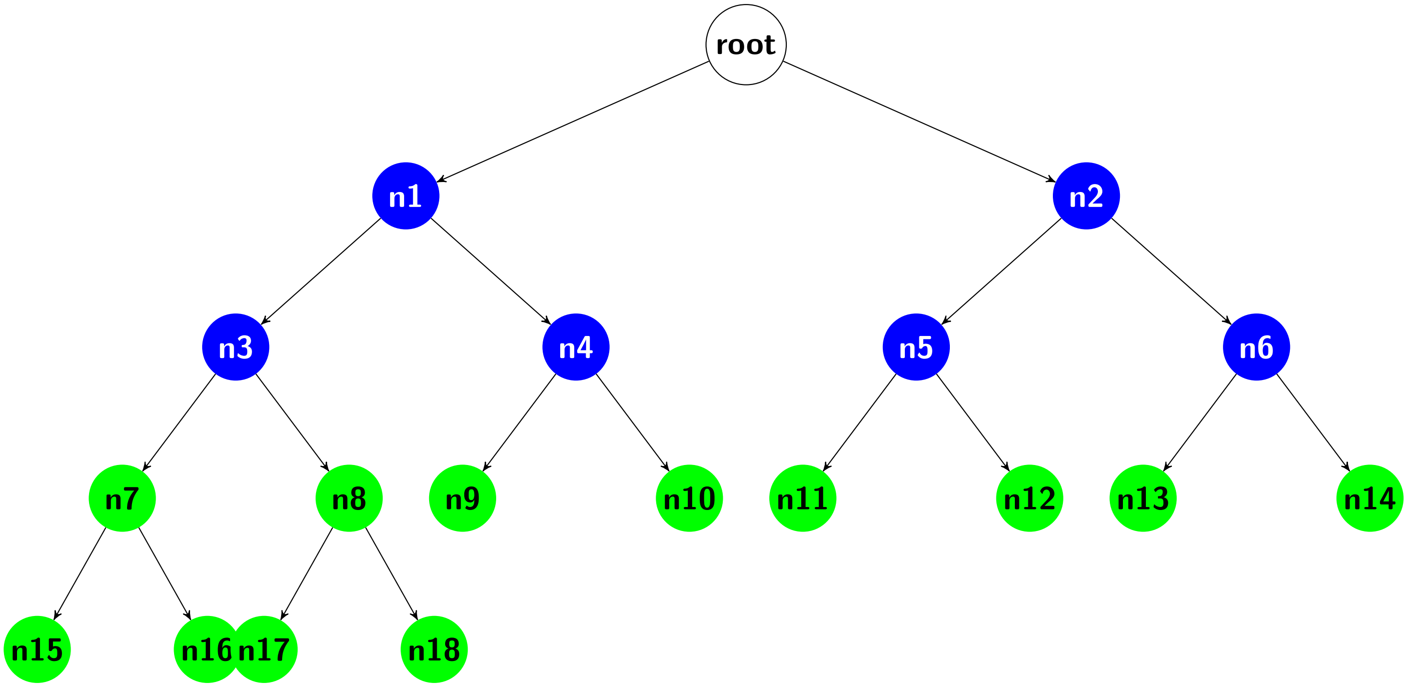
The **Hierarchical Bayesian Logistic Regression (HBLR)** model is suited for contexts when there are multiple classes for a categorical response variable Y, or when the binary response variable Y applies differently for groups of observations. In any case, there are different classes that may relate hierarchically to each other.

The Bayesian approach makes the ground of the HBLR model, because each group of classes consists on a specialization of a more general group, then implying that all groups with a same parent have their parameters following similar probability distributions, which constitute prior and posterior distributions as in BLR models.

The HBLR model has the following taxonomy. There are a set of nodes $\mathcal{N} = \{1, 2, ...\}$, from which a parental function is defined: $\pi : \mathcal{N} \rightarrow \mathcal{N}$, where $\pi(n)$ is the parent-node of node n. There is also a dataset $\mathcal{D} = \{(x_i, y_{i,t})\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^p$ is the input vector and $t \in T$ is the node for the class of observation i, where $T \subset \mathcal{N}$ is the subset of terminal nodes. Finally, C_n is the set of child-nodes of n. Each data point $(x_i, y_{i,t})$ necessarily falls into one of the nodes in T.

In figure 4.1, for instance, $n1 = \pi(n3) = \pi(n4)$ and $n7 = \pi(n15) = \pi(n16)$. Besides, $n15, n16, n17, n18 \in T$, and it may the case that the label for $n15$ and $n17$ is 0, while for $n16$ and $n18$ is 1.

Figure 4.1: Hierarchical model



For each node $n \in \mathcal{N}$, there is a set of parameters θ_n that follow a Normal prior distribution determined by a hyper-parameter α_n, which in turn follows a Gamma hyper-prior distribution. Below, four main specifications of the HBLR model are presented.

1. **Model 1 (M1):** parameters θ_n $(px1)$ follow a Normal distribution whose mean and variance matrices are dictated by the parameters of the parent node, $\pi(n)$.

   - Root node:

$$\theta_{root} \sim N(\theta_0, \Sigma_0) \tag{35}$$

$$\alpha_{root} \sim \Gamma(a_0, b_0) \tag{36}$$

   - Internal nodes:

$$\theta_n \sim N(\theta_{\pi(n)}, \Sigma_{\pi(n)}) \ \forall \ n \in \mathcal{N} \tag{37}$$

$$\Sigma_{\pi(n)} = \alpha_{\pi(n)}^{-1} I_p \tag{38}$$

$$\alpha_n \sim \Gamma(a_n, b_n) \ n \notin T \tag{39}$$

   - Terminal nodes:

$$y|x \sim Multinomial(p_1(x), p_2(x), ..., p_{|T|}(x)) \ \forall \ (x, y) \in \mathcal{D} \tag{40}$$

   - Probability of classes:

$$p_t(x) = \frac{\exp(\theta_t^T x)}{\sum_{t' \in T} \exp(\theta_{t'}^T x)} \tag{41}$$

2. **Model 2 (M2):** now, instead of one single α_n, there is one $\alpha_n^{(j)}$ for each input $x_j \in x$. Therefore, α_n is now a vector instead of a scalar: $\alpha_n = (\alpha_n^{(1)}, \alpha_n^{(2)}, ..., \alpha_n^{(p)})$. This approach produces an *Automatic Relevance Determination (ARD)*.

   - Root node: the same as (35) and (36).

   - Internal nodes:

$$\theta_n \sim N(\theta_{\pi(n)}, \Sigma_{\pi(n)}) \ \forall \ n \in \mathcal{N} \tag{42}$$

$$\Sigma_{\pi(n)}^{-1} = diag(\alpha_{\pi(n)}^{(1)}, \alpha_{\pi(n)}^{(2)}, ..., \alpha_{\pi(n)}^{(p)}) \tag{43}$$

$$\alpha_n^{(j)} \sim \Gamma(a_n^{(j)}, b_n^{(j)}) \;\; n \notin T \tag{44}$$

- Terminal nodes: the same as (40).
- Probability of classes: the same as (41).

3. **Model 3 (M3):** now, instead of variance matrices that are equal for all child of a same parent $\Sigma_{\pi(n)}$, each node has its own variance matrix Σ_n.

   - Root node: the same as (35) and (36).
   - Internal nodes:

$$\theta_n \sim N(\theta_{\pi(n)}, \Sigma_n) \;\; \forall \;\; n \in \mathcal{N} \tag{45}$$

$$\Sigma_n = \alpha_n^{-1} I_p \tag{46}$$

$$\alpha_n \sim \Gamma(a_n, b_n) \;\; \forall \;\; n \in \mathcal{N} \tag{47}$$

   - Terminal nodes: the same as (40).
   - Probability of classes: the same as (41).

4. **Model 4 (M4):** considers both extensions in M2 and M3, i.e., ARD and variance matrices node-specific.

   - Root node: the same as (35) and (36).
   - Internal nodes:

$$\theta_n \sim N(\theta_{\pi(n)}, \Sigma_n) \;\; \forall \;\; n \in \mathcal{N} \tag{48}$$

$$\Sigma_n^{-1} = diag(\alpha_n^{(1)}, \alpha_n^{(2)}, ..., \alpha_n^{(p)}) \tag{49}$$

$$\alpha_n^{(j)} \sim \Gamma(a_n^{(j)}, b_n^{(j)}) \;\; \forall \;\; n \in \mathcal{N} \tag{50}$$

- Terminal nodes: the same as (40).
- Probability of classes: the same as (41).

Assuming specification $M2$, then $\theta = (\theta_1, \theta_2, ..., \theta_{|\mathcal{N}|})$ and $\alpha = (\alpha_1, \alpha_2, ..., \alpha_{|\mathcal{N}|})$, where both θ_n and α_n are vectors $px1$. Therefore, the posterior probability of $P(\theta, \alpha|\mathcal{D})$ can be summarized by using expression (3):

$$P(\theta, \alpha|\mathcal{D}) \propto P(\mathcal{D}|\theta, \alpha)P(\theta|\alpha)P(\alpha) = P(\mathcal{D}|\theta)P(\theta|\alpha)P(\alpha)$$

$$P(\theta, \alpha|\mathcal{D}) \propto \Big(\prod_{(x,y_t)\in\mathcal{D}} p_t(x)\Big)\Big(\prod_{n\in\mathcal{N}} P(\theta_n|\alpha_n)\Big)\Big(\prod_{n\in\mathcal{N}} P(\alpha_n)\Big)$$

$$P(\theta, \alpha|\mathcal{D}) \propto \Big(\prod_{(x,y_t)\in\mathcal{D}} p_t(x)\Big)\Big(\prod_{n\in\mathcal{N}} P(\theta_n|\alpha_n)\Big)\Big(\prod_{n\in\mathcal{N}} \prod_{j=1}^{p} P(\alpha_n^{(j)})\Big)$$

$$P(\theta, \alpha|\mathcal{D}) \propto \Big(\prod_{(x,y_t)\in\mathcal{D}} \frac{\exp(\theta_t^T x)}{\sum_{t'\in T}\exp(\theta_{t'}^T x)}\Big)\Big(\prod_{n\in\mathcal{N}} N(\theta_{\pi(n)}, \Sigma_n)\Big)\Big(\prod_{n\in\mathcal{N}} \prod_{j=1}^{p} \Gamma(a_n^{(j)}, b_n^{(j)})\Big) \tag{51}$$

As discussed in sections 2 and 3, expression (51) must be approximated using a global and local variational method that leads to the following **variational posterior**:

$$Q(\theta, \alpha) = Q(\theta)Q(\alpha) = \prod_{n\in\mathcal{N}} q_\theta(\theta_n) \prod_{n\in\mathcal{N}} q_\alpha(\alpha_n) \propto \Big(\prod_{n\in\mathcal{N}} N(\mu_n, \Phi_n)\Big)\Big(\prod_{n\in\mathcal{N}} \prod_{j=1}^{p} \Gamma(\tau_n^{(j)}, v_n^{(j)})\Big) \tag{52}$$

Where the parameters μ_n, Φ_n, $\tau_n^{(j)}$ and $v_n^{(j)}$ follow:

$$\mu_n = \Phi_n\Big(I(n \in T) \sum_{(x,y_t)\in\mathcal{D}} \Big(I(t = n) - (1/2) - 2\lambda(\xi_{xy})\beta_x)x\Big) + diag(\tau_{\pi(n)}/v_{\pi(n)})\mu_{\pi(n)} + diag(\tau_n/v_n) \sum_{c\in C_n} \mu_c\Big) \tag{53}$$

$$\Phi_n^{-1} = I(n \in T) \sum_{(x,y_t)\in\mathcal{D}} \Big(2\lambda(\xi_{xy})xx^T\Big) + diag(\tau_{\pi(n)}/v_{\pi(n)}) + |C_n|diag(\tau_n/v_n) \tag{54}$$

$$\tau_n^{(j)} = a_n^{(j)} + \frac{|C_n|}{2} \tag{55}$$

$$v_n^{(j)} = b_n^{(j)} + \sum_{c\in C_n} \Big(\Phi_n^{(j,j)} + \Phi_c^{(j,j)} + (\mu_n^{(j)} - \mu_c^{(j)})^2\Big) \tag{56}$$

The file named “HBLR_demonstration.pdf” displays the derivation of parameters (53)-(56).

The variational parameters ξ_{xy} and β_x emerge as $h(\theta,\xi)$ in (23) is replaced by the *soft-max normalization*, which introduces this new parameter β_x (Bouchard, 2008). While parameters (53)-(56) are defined sequentially in step E of EM algorithm, the variational parameters are derived from step M:

$$\xi_{xy}^2 = x^T diag(\tau_n/\upsilon_n)x + (\beta_x - \mu_n^T x)^2 \tag{57}$$

$$\beta_x = \Big((1/2)(|T|/2-1) + \sum_{n\in T}\lambda(\xi_{xy})\mu_n^T x\Big)/\sum_{n\in T}\lambda(\xi_{xy}) \tag{58}$$

The *predictive density* in the context of HBLR model is given by:

$$P(y|x) = \int P(y,\theta|x)d\theta \approx \int P(y|x,\theta)q_\theta(\theta)d\theta \tag{59}$$

The non-conjugacy between $P(y|x,\theta)$ and $q_\theta(\theta)$ implies in the following approximation to $P(y,\theta|x)$:

$$P(y,\theta|x) \approx \tilde{q}(y,\theta) = \prod_{n\in T}\tilde{q}_\theta(\theta_n)\tilde{q}_y(y_n) = \prod_{n\in T} N(\tilde{\mu}_n,\tilde{\Phi}_n)Bernoulli(\tilde{p}_n) \tag{60}$$

Integrating (60) over θ shows that $P(y|x) \approx \int \tilde{q}(y,\theta)d\theta = \prod_{n\in T}\tilde{q}_y(y_n)$.

The set of parameters that configure the HBLR model is given by:

$$\theta_0, \Sigma_0, a_0, b_0, \text{ for root node, and } \theta_n, \Sigma_n, a_n, b_n \text{ (or, } a_n^{(j)}, b_n^{(j)}\text{), for } n \in \mathcal{N} \tag{61}$$

Since the mean of θ_n is defined by the parent node and Σ_n depends ultimately on a_n and b_n (or, $a_n^{(j)}$ and $b_n^{(j)}$), then the following prior parameters should be defined in a first place. Gopal et al (2012) suggest:

$$\theta_0 = 0, \Sigma_0 = I_p \tag{62}$$

$$(a_n^{(j)}, b_n^{(j)}) = \begin{cases} \Big(\sum_{c\in C_n} a_c^{(j)}, \sum_{c\in C_n} b_c^{(j)}\Big) & \text{if } n \in \mathcal{N}\backslash T \\ (1, \hat{I}(n)^{(j,j)-1}) & \text{if } n \in T \end{cases} \tag{63}$$

Where C_n is the set of children nodes for intermediate node n, and $\hat{I}(n)$ is the Fisher information matrix for terminal node n.

The use of the Fisher information matrix reflects that $Var(\hat{\theta}^{MLE}) = I^{-1}(\hat{\theta}^{MLE})$, and that $Var(\theta_n^{(j)}) = \alpha_n^{(j)-1}$, consequently, $E(Var(\theta_n^{(j)})) = E(\alpha_n^{(j)-1}) = b_n^{(j)}/a_n^{(j)}$. Since the Fisher information matrix is the hessian matrix of the sample log-likelihood, then for a terminal node $t \in T$:

$$\hat{I}(t) = \sum_{(x,y_t)\in\mathcal{D}} \hat{p}(x)(1-\hat{p}(x))xx^T \tag{64}$$

Where $\hat{p}(x) = \sum_{(x,y_t)\in\mathcal{D}} y_t/|\mathcal{D}|$, and y_t defined as:

$$y_t = \begin{cases} 1 & \text{if } n \in x \in t \\ 0 & \text{otherwise} \end{cases}$$

5 References

Bishop, Christopher M. *Pattern Recognition and Machine Learning*, 2006.

Bouchard, Guillaume. *Efficient Bounds for the Softmax Function and Applications to Approximate Inference in Hybrid Models*, 2008.

Drugowitsch, Jan. *Variational Bayesian Inference for Linear and Logistic Regression*, 2019.

Gopal, Siddharth; Bai, Bing; Yang, Yiming; Niculescu-Mizil, Alex. *Bayesian Models for Large-Scale Hierarchical Classification*, 2012.

Gopal, Siddharth; Bai, Bing; Yang, Yiming; Niculescu-Mizil, Alex. *Bayesian Models for Large-Scale Hierarchical Classification (Supplementary)*, 2012.

Jaakkola, Tommi S., Jordan, Michael I. *A Variational Approach to Bayesian Logistic Regression Models and their Extensions*, 1996.

5.1 Links

Conjugate prior: definition and examples.

Markov Chain Monte Carlo (MCMC): definition and applications.

Maximum a posteriori: definition and example.

5.2 Notes

A series of handwritten notes presents demonstrations of important results for Bayesian models.

1. “bayesian_notes.pdf”: provides notes based on Wikipedia pages for Bayesian inference and variational approximation.

2. “drugowitsch.pdf”: summarizes paper of Drugowitsch (2019) on Bayesian Logistic Regression (BLR) model.

3. “hblr.pdf”: presents and discusses the Hierarchical Bayesian Logistic Regression (HBLR) model based on Gopal et. al. (2012).

4. “jaakkola_jordan.pdf”: sums up main results from seminal paper of Jaakkola and Jordan (1996).

5. “bishop.pdf”: presents all notes and derivations from chapter 10 of Bishop (2006).