Notes on Bayesian Rule Lists

Richard Hydomako

Toronto Prob. Programming Meetup

Nov. 17, 2016

Bayesian Rule Lists overview

- Designed to be highly interpretable, accurate, and fast to train
- Sparse list of *if* - *then* clauses, which make it easy to track how a predicted value came about from the input vector
- Example from the paper related to the Titanic dataset:

if male **and** adult **then** survival probability 21% (19% – 23%)
else if 3rd class **then** survival probability 44% (38% – 51%)
else if 1st class **then** survival probability 96% (92% – 99%)
else survival probability 88% (82% – 94%)

References

Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model
https://arxiv.org/abs/1511.01644

Scalable Bayesian Rule Lists
https://arxiv.org/abs/1602.08610

Initial definitions

Training data:

- Consider a set of training data $\{(x_i, y_i)\}, i = 1 \ldots n$
- where $x_i \in \mathcal{R}^d$, are the d-dimensional features,
- and the labels are $y_i \in 1, \ldots, L$
- $\mathbf{x} = (x_1, \ldots, x_n)$
- $\mathbf{y} = (y_1, \ldots, y_n)$

Association rules

- An association rule relates an antecedent, a, to a consequent, b ($a \rightarrow b$)
- The antecedent is a true/false statement about a feature vector x_i, for example, ($x_{i,1} = \text{adult AND } x_{i,2} = \text{1st class}$)
- The number of conditions in the antecedent is the antecedent-cardinality
- The consequent is typically the prediction label y_i

Bayesian association rule

- A Bayesian association rule has a distribution over the predicted labels, for example, a multinomial distribution:

$$a \to y \sim \text{Multinomial}(\boldsymbol{\theta})$$

- Similarly, we can choose a prior distribution for the consequent parameter:

$$\boldsymbol{\theta} \mid \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

Bayesian association rule

- The posterior consequent distribution is then:

$$\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha} + N)$$

- where $N = (N_{\cdot,1}, \ldots, N_{\cdot,L})$ and $N_{\cdot,i}$ is the number of observations with prediction labels $i = l$

Bayesian Rule List

- A Bayesian decision list is an ordered list of antecedents $d = (a_1, \ldots, a_m)$
- We say that an antecedent (or rule) 'captures' an observation x_i if the observation satisfies the condition a_j but not $a_i, \ldots, a_{j-1}$
- Let $N_{j,l}$ be the number of observations captured by the jth rule with prediction label l, and $N_{0,l}$ is the number of observations that don't satisfy any of the rules and have prediction labels l
- Then $\mathbf{N}_j = (N_{j,1}, \ldots, N_{j,L})$ and $\mathbf{N} = (\mathbf{N}_0, \ldots, \mathbf{N}_m)$
- A fully specified Bayesian Rule List is then $D = (d, \boldsymbol{\alpha}, \mathbf{N})$

Bayesian Rule List

The unrolled model looks like this:

$$
\begin{array}{lll}
\textbf{IF} & a_1 \textbf{ THEN} & y \sim \text{Multi}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{N_1}) \\
\textbf{ELSE IF} & a_2 \textbf{ THEN} & y \sim \text{Multi}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2 \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{N_2}) \\
\vdots & & \\
\textbf{ELSE IF} & a_m \textbf{ THEN} & y \sim \text{Multi}(\boldsymbol{\theta}_m), \theta_m \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{N_m}) \\
\textbf{ELSE} & & y \sim \text{Multi}(\boldsymbol{\theta}_0), \boldsymbol{\theta}_0 \sim \text{Dirichlet}(\alpha + \mathbf{N_0})
\end{array}
$$

Bayesian Rule List

In the common case for binary prediction labels:

$$
\begin{array}{lll}
\textbf{IF} & a_1 \textbf{ THEN} & y \sim \text{Binomial}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N_1}) \\
\textbf{ELSE IF} & a_2 \textbf{ THEN} & y \sim \text{Binomial}(\boldsymbol{\theta}_2), \boldsymbol{\theta}_2 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N_2}) \\
\vdots & & \\
\textbf{ELSE IF} & a_m \textbf{ THEN} & y \sim \text{Binomial}(\boldsymbol{\theta}_m), \boldsymbol{\theta}_m \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N_m}) \\
\textbf{ELSE} & & y \sim \text{Binomial}(\boldsymbol{\theta_0}), \boldsymbol{\theta}_0 \sim \text{Beta}(\boldsymbol{\alpha} + \mathbf{N_0})
\end{array}
$$

Mining antecedents

- Pre-mine the set of antecedents, $\mathcal{A}$ from the training data
- A tractible number of antecedents makes the overall inference faster and arguably more intpretable
- Here, they use an algorithm called **FP-growth** to find frequent itemsets with a specified minimum support and maximum cardinality

Algorithm 1 BRL Generative Model

procedure GENERATE RULE LIST
hyperparameters: $\boldsymbol{\alpha}, \lambda, \eta$
    Sample a decision list length $m \sim p(m \mid \lambda)$
    Sample the default rule parameter $\boldsymbol{\theta}_0 \sim \text{Dirichlet}(\alpha)$
    for decision list rule $j = 1, \ldots, m$ **do**
        Sample the cardinality of antecedent $a_j \in d$ as $c_j \sim p(c_j \mid c_{<j}, \mathcal{A}, \eta)$
        Sample a_j of cardinality c_j from $p(a_j \mid a_{<j}, c_j, \mathcal{A})$
        Sample rule consequent parameter $\boldsymbol{\theta}_j \sim \text{Dirichlet}(\boldsymbol{\alpha})$
    end for
    for observation $i = 1, \ldots, n$ **do**
        Find the antecedent a_j in d that is the first that applies to x_i
        If no antecedents in d apply, set $j = 0$
        Sample $y_i \sim \text{Multinomial}(\boldsymbol{\theta}_j)$
    end for
end procedure

Posterior distribution over antecedent lists

- Following from the generative model description, we can write down an equation for the posterior distribution:

$$p(d \mid \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \boldsymbol{\alpha}, \lambda, \eta) \propto \overbrace{p(\boldsymbol{y} \mid \boldsymbol{x}, d, \boldsymbol{\alpha})}^{\text{likelihood function}} \underbrace{p(d \mid \mathcal{A}, \lambda, \eta)}_{\text{hierarchical prior}}$$

Likelihood function

- The likelihood is the product of the multinomial PMFs for the observed labels for each rule:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, d, \boldsymbol{\theta}) = \prod_{j:\sum_l N_{j,l}>0} \text{Multinomial}(\mathbf{N}_j \mid \theta_j)$$

- with

$$\theta_j \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

- which gives the Dirichlet-multinomial distribution:

$$p(\boldsymbol{y} \mid \boldsymbol{x}, d, \boldsymbol{\alpha}) \propto \prod_{j=0}^{m} \frac{\prod_{l=1}^{L} \Gamma(N_{j,l} + \alpha_l)}{\Gamma(\sum_{l=1}^{L} N_{j,l} + \alpha_l)}$$

Prior

The full prior is given as:

$$p(d \mid \mathcal{A}, \lambda, \eta) = \overbrace{p(m \mid \mathcal{A}, \lambda)}^{\text{number of rules}} \prod_{j=1}^{m} \underbrace{p(c_j \mid c_{<j}, \mathcal{A}, \eta)}_{\text{antecedent cardinality}} \overbrace{p(a_j \mid a_{<j}, c_j, \mathcal{A})}^{\text{available antecedents}}$$

where the first term is given by a truncated Poisson:

$$p(m \mid \mathcal{A}, \lambda) = \frac{(\lambda^m / m!)}{\sum_{j=0}^{|\mathcal{A}|} (\lambda^j / j!)}, \quad m = 0, \ldots, |\mathcal{A}|$$

So λ can be interpreted as the prior belief of the length of the list required to model the data

Prior

The distribution of antecedent cardinalities is also given as a Possion truncated at zero and at the maximum cardinality C:

$$p(c_j \mid c_{<j}, \mathcal{A}, \eta) = \frac{(\eta^{c_j}/c_j!)}{\sum_{k \in R_{j-1}(c_{<j}, \mathcal{A})}(\eta^k/k!)}, \quad c_j \in R_{j-1}(c_{<j}, \mathcal{A})$$

where $R_j(c_1, \ldots, c_j, \mathcal{A})$ is the set of cardinalities available after having drawn antecedent j
After selecting the antecedent cardinality, c_j, an antecedent a_j is sampled from the available antecedents of cardinality c_j. This is done uniformly over the available antecedents, excluding any already selected:

$$p(a_j \mid a_{<j}, c_j, \mathcal{A}) \propto 1, \quad a_j \in \{a \in \mathcal{A} \setminus a_{<j} : |a| = c_j\}$$

MCMC

We can perform Metropolis-Hastings sampling on this posterior distribution. However, since we're performing inference over lists of antecedents, generating proposal rule lists unfolds somewhat differently that with the usual discrete/continuous distributions:

- Given an initial rule list, a proposal is a mutation of that rule list. Three list mutations are possible:
  1. insert a new rule,
  2. delete a rule,
  3. swap two existing rules
- We also have to ensure detailed balance — that is, we have to account for the transition to a proposed list, as well as the transition from the proposed list back to the current list

MCMC

Metropolis-Hastings acceptance condition:

$$r \geq \min\left(1, \frac{p(d^* \mid \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \boldsymbol{\alpha}, \lambda, \eta) Q(d^t \mid d^*)}{p(d^t \mid \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \boldsymbol{\alpha}, \lambda, \eta) Q(d^* \mid d^t)}\right)$$

where d^t is the 'current' rule list, and d^* is the proposal

$$\begin{aligned} Q(d^* \mid d^t) = & \; p(d^* \mid d^t, \text{swap}) p(\text{swap}) \\ & + p(d^* \mid d^t, \text{insert}) p(\text{insert}) \\ & + p(d^* \mid d^t, \text{delete}) p(\text{delete}) \end{aligned}$$

where

$$p(d^* \mid d^t, \text{swap}) = \frac{1}{(|d^t|)((|d^t| - 1)}$$

$$p(d^* \mid d^t, \text{insert}) = \frac{1}{|A| - |d^t|)(|d^t + 1)}$$

$$p(d^* \mid d^t, \text{delete}) = \frac{1}{|d^t|}$$

Point Estimates

For this kind of model, a point estimate is a single Bayesian Rule List. Moreover, for a given point estimate we can estimate a posterior predictive distribution (that is, the probability distribution of a predicted label l, conditioned on the given rule list):

$$\begin{aligned} p(\tilde{y} = l \mid \tilde{x}, d, \boldsymbol{x}, \boldsymbol{y}, \alpha) &= \int_{\theta} \theta_l \, p(\theta \mid \tilde{x}, \boldsymbol{x}, \boldsymbol{y}, \alpha) d\theta \\ &= \mathbb{E}[\theta_l \mid \tilde{x}, \boldsymbol{x}, \boldsymbol{y}, \alpha] \end{aligned}$$

Point Estimates

If we remember from earlier, the posterior consequent distribution is:

$$\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha} + N)$$

So, similarly, the posterior predictive consequent distribution is given as:

$$\boldsymbol{\theta} \mid \tilde{x}, \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{\alpha} + \mathbf{N}_{j(d,\tilde{x})}$$

or

$$p(\tilde{y} = l \mid \tilde{x}, d, \boldsymbol{x}, \boldsymbol{y}, \alpha) = \frac{\alpha_l + N_{j(d,\tilde{x})}}{\sum_{k=1}^{L} (\alpha_k + N_{j(d,\tilde{x}),k})}$$

where $j(d, \tilde{x})$ is the index of the antecedent in d that captures $\tilde{x}$

Point Estimates

The question is then: which d should be taken as an appropriate point estimate?

Since the posterior mean is a distribution over antecedent lists, the authors suggest taking an analog to the 'mean', where we subselect rule lists with lengths similar to the posterior average list length and widths similar to the average posterior cardinality $\bar{c} = \frac{1}{m}\sum_{j=1}^{m} c_j$. Within this subset, we can then take the rule with the highest posterior probability as the point estimate.

The authors note that the MAP estimate will likely give rule lists that underestimate the rule length (and therefore shorter lists), because of the nature or the prior distribution.

Full posterior predictive estimates

The full posterior distribution can also be used to make predictions:

$$p(\tilde{y} \mid \tilde{x}, \mathcal{A}, \boldsymbol{x}, \boldsymbol{y}, \alpha, \lambda, \eta) = \sum_{d \in \mathbf{D}} p(\tilde{y} \mid \tilde{x}, d, \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \alpha) p(d \mid \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \alpha, \lambda, \eta)$$

$$= \sum_{d \in \mathbf{D}} \frac{\alpha_l + N_{j(d,\tilde{x})}}{\sum_{k=1}^{L} (\alpha_k + N_{j(d,\tilde{x}),k})} p(d \mid \boldsymbol{x}, \boldsymbol{y}, \mathcal{A}, \alpha, \lambda, \eta)$$

- where $\mathbf{D}$ is the set of all ordered subsets of $\mathcal{A}$ — in practice we can use the MCMC trace samples to approximate the sum
- note that although using the full posterior should provide better prediction accuracy, it negates the benifits of employing a single short, interpretable rule-list