Sparse Bayesian Classification with the Relevance Vector Machine

In a pattern recognition setting, the Relevance Vector Machine (RVM) [3] can be viewed as a simple logistic regression model, with a Bayesian Automatic Relevance Determination (ARD) prior [1] over the weights associated with each feature in order to achieve a parsimonious model. Let $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ represent an alphabet of symbols representing the bases adenine, cytosine, guanine and thymine respectively. The RVM constructs a decision rule from a set of labelled training data,

$$\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{\ell}, \qquad \boldsymbol{x}_i \in \mathcal{A}^{n_i}, \qquad t_i \in \{0, \ +1\},$$

where the input patterns, $\boldsymbol{x}_i$, consist of strings drawn from $\mathcal{A}$ of varying lengths, describing the upstream promoter regions of a set of co-regulated genes. The target patterns, t_i, indicate whether the corresponding gene is up-regulated (class $\mathcal{C}_1$, $y_i = +1$) or down-regulated (class $\mathcal{C}_2$, $y_i = 0$) under a given treatment. The RVM constructs a logistic regression model based on a set of sequence features derived from the input patterns, i.e.

$$p(\mathcal{C}_1|\boldsymbol{x}) \approx \sigma\left\{y(\boldsymbol{x}; \boldsymbol{w})\right\} \qquad \text{where} \qquad y(\boldsymbol{x}; \boldsymbol{w}) = \sum_{i=1}^{N} w_i \varphi_i(\boldsymbol{x}) + w_0, \tag{1}$$

and $\sigma\left\{y\right\} = (1 + \exp\{y\})^{-1}$ is the logistic inverse link function. In this study a feature, $\varphi_i(\boldsymbol{x})$, represents the number of times an arbitrary substring, $s_i \in \mathcal{A}^d$, ocurrs in a promoter sequence $\boldsymbol{x}$. A sufficiently large set of features is used such that it is reasonable to expect that some of these features will represent oligonucleotides forming a relevant promoter protein binding site and so provide discriminatory information for the pattern recognition task at hand. Assuming a Bernoulii distribution for $P(t|\boldsymbol{x})$, the *likelihood* of the training data, $\mathcal{D}$, can be written as

$$P(\mathcal{D}|\boldsymbol{w}) = \prod_{i=1}^{\ell} \sigma\left\{y(\boldsymbol{x}_i; \boldsymbol{w})\right\}^{t_i} \left[1 - \sigma\left\{y(\boldsymbol{x}_i; \boldsymbol{w})\right\}\right]^{1-t_i} \tag{2}$$

To form a Bayesian training criterion, we must also impose a prior distribution over the vector of model parameters or *weights*, $p(\boldsymbol{w})$. The RVM adopts a separable Gaussian prior, with a distinct hyper-parameter, α_i, for each weight,

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1}). \tag{3}$$

The optimal parameters of the model are then given by the minimiser of the penalised negative log-likelihood,

$$\log\left\{P(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})\right\} = \sum_{i=1}^{\ell} \left[t_i \log y_i + (1 - t_i)\log(1 - y_i)\right] - \frac{1}{2}\boldsymbol{w}^T \boldsymbol{A}\boldsymbol{w}. \tag{4}$$

where $y_i = \sigma\left\{y(\boldsymbol{x}_i; \boldsymbol{w})\right\}$ and $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha})$ is a diagonal matrix with non-zero elements given by the vector of hyper-parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. This is achieved via the efficient Iteratively Re-Weighted Least Squares (IRWLS) algorithm [2]. Next, Laplace's method is used to obtain a Gaussian approximation to the posterior density of the weights,

$$p(\boldsymbol{w}|\mathcal{D}, \boldsymbol{\alpha}) \approx \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5}$$

where the posterior mean and covariance are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{t}, \qquad \text{and} \qquad \boldsymbol{\Sigma} = \left[\boldsymbol{\Phi}^T\boldsymbol{B}\boldsymbol{\Phi} + \boldsymbol{A}\right]^{-1}$$

respectively, $\boldsymbol{\Phi}$ is an $\ell \times N$ matrix of features for each promoter in the training set and $\boldsymbol{B}$ is a diagonal matrix with non-zero elements $b_{ii} = y_i(1 - y_i)$. The hyper-parameters are then updated in order to maximise their marginal likelihood, $p(\mathcal{D}|\boldsymbol{\alpha})$, according to the efficient update formula

$$\alpha_i^{\mathrm{new}} = \frac{\gamma_i}{\mu_i^2} \qquad \text{where} \qquad \gamma_i = 1 - \alpha_i\Sigma_{ii}. \tag{6}$$

This process is repeated until an appropriate convergence criterion is met (see [3] for details). The maximisation of the marginal likelihood, or *evidence*, for the hyper-parameters, $\boldsymbol{\alpha}$, leads to the hyper-parameters associated with uninformative features becoming very large. This in turn forces the value of the associated weight essentially to zero, allowing redundant features to be easily identified and pruned from the model. Given a sufficiently rich set of sequence features, it seems reasonable to suggest that the features retained by the RVM *may* represent (parts of) transcription factor binding sites as they provide discriminatory information distinguishing between up- and down-regulated genes.

References

[1] D. J. C. MacKay. Bayesian methods for backpropagation networks. In *Models of Neural Networks III*, chapter 6, pages 211–254. Springer, 1994.

[2] I. T. Nabney. Efficient training of RBF networks for classification. In *Proceedings of the Ninth Int. Conf. on Artificial Neural Networks*, volume 1, pages 210–215, September 7–10 1999.

[3] M. E. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.