10-801: Advanced Topics in Graphcal Models 10-801, Spring 2007

Bayesian Feature Selection

Lecturer: Eric P. Xing *Scribes: Fan Guo*

1 Bayesian Feature Selection

We use feature selection in linear regression as the example in the following discussion. The standard linear regression formula is $y_n = \vec{\beta}^T \vec{x}_n + \epsilon = \sum_{k=1}^{K} \beta_k x_{nk} + \mathcal{N}(0, \sigma^2)$, where ϵ represents a zero-mean Gaussian noise with variance σ^2. We want to find a sparse $\vec{\beta}$ to reduce the dimension of inputs. Although this sparse formulation does not make much difference in the computational at first glance, dimension reduction may lead to very different generalization error of the model and reduce the work load of data collection.

1.1 The Hierarchial Bayesian Model for Bayesian Feature Selection

The feature selection could be performed in a principled fansion using the hierarchical Bayesian model whose graphical structure is depicted in Fig. 1. Both of the parameters $\vec{\beta}$ and σ^2 are treated as random variables. Each component of $\vec{\beta}$ is sampled from a mixture of Gaussian distribution:

$$\beta_k | \gamma_k = 0 \sim \mathcal{N}(0, \tau^2), \tag{1}$$

$$\beta_k | \gamma_k = 1 \sim \mathcal{N}(0, c\tau^2). \tag{2}$$

γ_k is an indicator variable for β_k; $\gamma_k = 0$ represents the case that the kth dimension of inputs is not selected. Accordingly, τ^2 is supposed to be small such that the conditional distribution of β_k given $\gamma_k = 0$ is shrinked to 0, c should be much greater than 1 which reflects greater degrees of uncertainty for the parameter when its corresponding feature is present. Although we can again set up prior distributions for c, τ^2, however, for the discussion in the sequel, we assume c and τ^2 are known constants estimated from the empirical Bayes approach, and let γ_k simply follows an independent Bernoulli prior distribution.

$$\gamma_k \sim \text{Bernoulli}(p_k). \tag{3}$$

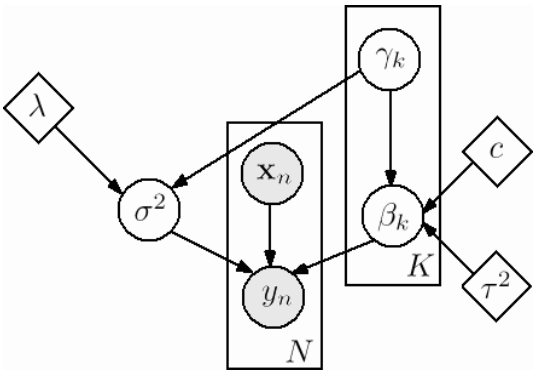

Figure 1: Graphical structure for Bayesian feature selection in linear regression.

For the variable σ^2, it is natural to choose a conjugate prior distribution. A convenient parameterization is the inverse gamma distribution which is a two-parameter family of continuous probability distribution with the density function

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)}\sigma^{2(-a-1)} \exp\left(-\frac{b}{\sigma^2}\right), \tag{4}$$

where a is the shape parameter and b is the scale parameter. We assume a and b is determined by $\vec{\gamma}$ as well as some known constants λ, and write the prior on σ^2 as:

$$\sigma^2|\vec{\gamma} \sim \text{Inv-Gamma}(\sigma^2; a_\lambda(\vec{\gamma}), b_\lambda(\vec{\gamma})), \tag{5}$$

and we may drop the arguments of a and b in future discussion. Now that we have defined all the prior distribution, the model for generating the data is

$$y_n|\vec{x}_n, \vec{\beta}, \sigma^2 \sim \mathcal{N}(\vec{\beta}^T\vec{x}_n, \sigma^2), \tag{6}$$

where we assume $\vec{x}_n$ is fixed and known. We will also adapt a matrix representation of the features in the following: X is N by K matrix of all the input features.

1.2 Inference

The goal of the inference is to compute

$$p(\vec{\gamma}|X, Y) = \int\int p(\vec{\gamma}, \vec{\beta}, \sigma^2|X, Y)d\vec{\beta}d(\sigma^2). \tag{7}$$

The integral in intractable and approximate inference could be carried via the standard Gibbs sampling scheme. In the following we briefly discuss the sampling formulae.

$p(\gamma_k|\beta_k)$: Since γ_k is a binary random variable, the computation is fairly easy:

$$p(\gamma_k|\vec{\gamma}_{-k}, \beta_k) = \frac{p(\gamma_k)p(\beta_k|\gamma_k)p(\sigma^2|\gamma_k, \vec{\gamma}_{-k})}{\sum_{\gamma_k\in\{0,1\}} p(\gamma_k)p(\beta_k|\gamma_k)p(\sigma^2|\gamma_k, \vec{\gamma}_{-k})}, \tag{8}$$

where $p(\gamma_k), p(\beta_k|\gamma_k), p(\sigma^2|\vec{\gamma})$ are defined in Sec.1.1.

$p(\vec{\beta}|\vec{\gamma}, \sigma^2, X, \vec{y})$: Blocked version of sampling can be applied to draw samples from the conditional distribution of $\vec{\beta}$. We first write out the conditional distribution $p(\vec{\beta}|\vec{\gamma})$ as a multivariate normal:

$$\vec{\beta}|\vec{\gamma} \sim \mathcal{N}(\vec{0}, R_\gamma), \tag{9}$$

where R_γ is a diagonal matrix and the kth diagonal element $R_{kk} = c^{\gamma_k}\tau^2$. This is a conjugate prior of the multivariate normal data distribution

$$\vec{y}|X, \vec{\beta}, \sigma^2 \sim \mathcal{N}(X\vec{\beta}, \sigma^2 I_N), \tag{10}$$

where I_N is the N-dimensional identity matrix.

So the posterior is still a multivariate normal distribution:

$$\begin{aligned} p(\vec{\beta}|\vec{\gamma}, \sigma^2, X, \vec{y}) &\propto p(\vec{\beta}|\vec{\gamma})p(\vec{y}|X, \vec{\beta}, \sigma^2) \\ &\propto \exp\left(-\frac{1}{2}\vec{\beta}^T R_\gamma^{-1}\vec{\beta} - \frac{1}{2\sigma^2}(\vec{y} - X\vec{\beta})^T(\vec{y} - X\vec{\beta})\right) \\ &\propto \exp\left(-\frac{1}{2}\vec{\beta}^T\left(R_\gamma^{-1} + \frac{X^T X}{\sigma^2}\right)\vec{\beta} + \vec{\beta}^T\frac{X\vec{y}}{\sigma^2}\right). \end{aligned} \tag{11}$$

Therefore we obtain the sampling formula for $\vec{\beta}$:

$$\vec{\beta}|\vec{\gamma},\sigma^2,X,\vec{y} \sim \mathcal{N}\left(\left(\sigma^2 R_\gamma^{-1} + X^T X\right)^{-1} X^T \vec{y}, \left(R_\gamma^{-1} + \frac{X^T X}{\sigma^2}\right)^{-1}\right). \tag{12}$$

$p(\sigma^2|\vec{\gamma},\vec{\beta},X,\vec{y})$: Since the data distribution $p(\vec{y}|X,\vec{\beta},\sigma^2)$ is conjugate of the inverse-Gamma prior on σ^2, the posterior is tractable and also in the inverse-Gamma family:

$$\begin{aligned} p(\sigma^2|\vec{\gamma},\vec{\beta},X,\vec{y}) &\propto p(\sigma^2|\vec{\gamma})p(\vec{y}|X,\vec{\beta},\sigma^2) \\ &\propto \sigma^{2(-a-1)} \exp\left(-\frac{b}{\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2}(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})\right) \\ &\propto \sigma^{2(-a-1)} \exp\left(-\frac{1}{\sigma^2}\left(b + \frac{1}{2(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}\right)\right), \end{aligned} \tag{13}$$

so we can get the sampling formula for σ^2:

$$\sigma^2|\vec{\gamma},\vec{\beta},X,\vec{y} \sim \text{Inv-Gamma}\left(\sigma^2; a_\lambda(\vec{\gamma}), b_\lambda(\vec{\gamma}) + \frac{1}{2(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta})}\right). \tag{14}$$