Bayesian Framework for Gaussian Process Variable Selection

Module: Machine Learning for Finance

Shirley He

June 2020

1 Introduction

The typical method for variable selection when implementing Gaussian process models is automatic relevance determination (ARD). In this project I follow the paper by Paananen et al [1], to implement two variable selection methods that are proposed for Gaussian process models via sensitivity analysis of the posterior predictive distribution. I implement and reproduce their empirical results on synthetic data to show an improvement in variable selection compared to automatic relevant determination. I also explore other alternatives for ARD in the literature.

2 Background

2.1 Gaussian Processes and Automatic Relevance Determination

A Gaussian process is a collection $\{f(x), x \in \mathcal{X}\}$ (let $\mathcal{X}$ be a set in $\mathbb{R}^d$), such that for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the random vector $(f(x_1), \ldots, f(x_n))$ has a joint multivariate Gaussian distribution. Therefore, we can characterize the GP by its mean function $m(x) = \mathbb{E}[f(x)]$ and its covariance function (1). The covariance function is called the 'kernel' function.

$$k(x, x') = \operatorname{cov}(f(x), f(x')) = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{1}$$

The squared exponential covariance kernel in (2) is called the automatic relevance determination kernel because it can be used to discover relevant dimensions of the inputs and perform implicit feature selection, while optimizing the hyperparameters of the kernel, σ_f^2 and $(\ell_1, \ldots, \ell_d)$.

$$k_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{(x_i - x_i')^2}{l_i^2}\right) \tag{2}$$

The hyperparameters σ_f^2 and $(\ell_1, \ldots, \ell_d)$ determine the overall variability and lengthscale (number of turning points in the function, sometimes referred to as 'bandwidth') which allow us to scale each dimension of the inputs separately. Their choice is critical because they influence how the GP model can fit the observed data. They can be fixed or estimated.

For a given model, the parameters of the model θ, are typically selected by maximising the likelihood function $L(\theta) = p(y \mid \theta)$. The main idea is to maximize the probability of the sample data y. In the case of the Gaussian process regression, $\theta = (\theta_{\mathcal{K}}, \sigma_\varepsilon)$, where θ_x are the parameters of the kernel function, and σ_ε is the standard deviation of the noise. Letting $z = f(x)$ be the GP, we have:

$$p(y \mid \theta) = \int p(y \mid \theta, z) p(z \mid \theta) \mathrm{d}z \tag{3}$$

Then by maximizing the log marginal likelihood, where we integrate out the latent values of z, would solve:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta) \tag{4}$$

where $\ell(\theta) = \ln p(y \mid \theta)$.

The typical issues associated with comparing posterior probabilities using a Bayesian framework for model comparison are related to the marginal likelihood. Computing marginal likelihoods when θ is large requires computing a multi-dimensional integral. Additionally, the sensitivity of the marginal likelihood to our choice of priors on the model parameters creates a poor framework for model comparison. It is not optimal for the marginal likelihood to change significantly even if changes to the priors that we give to the parameters don't affect the posterior very much. A better method to compare models is look at the predictive accuracy of out-of-sample data.

Paananen et al express two problems associated with variable selection method for Gaussian process models via ARD. Firstly, they explain that the length-scale parameters alone are not well defined, but only the ratios of σ_f^2 and l_i [2], which increases variance of the relevance measure. Secondly, they describe how ARD systematically overestimates the predictive relevance of nonlinear variables relative to linear variables of equal relevance in the squared error sense [3].

Based on this information, they propose two alternative methods to improve the selection of optimal input variables in terms of predictive performance by utilizing the predictions of a full model in the vicinity of the training points, and thereby rank the variables based on their predictive relevance. In this report, I compare the results of a GP model with the ARD covariance function in equation (2) with the two alternative methods for variable selection for Gaussian process models, which are explained below.

2.2 Kullback-Leibler Divergence as a Measure of Predictive (KL Method) Performance

Paananen et al introduce the KL method based on Kullback-Leibler divergence (KLD), which is a known measure for comparing the difference between probability distributions. They claim that KLD is a favourable measure for predictive relevance because it takes into account changes in both the predictive mean and uncertainty.

They find that by relating to the total-variation distance of Pinsker's inequality, it is reasonable to utilize the Kullback- Leibler divergence from density p to q, $D_KL(p||q)$, as a measure of distance in the form [4], and that using the square root also allows linear approximation of infinitesimal changes in the predictive distribution via perturbations in the input variables.

$$d(p\|q) = \sqrt{2\mathcal{D}_{\mathrm{KL}}(p\|q)} \tag{5}$$

They compare predictive distributions at training points and points that are moved with respect to one variable. This means that at a training point $i, p\left(y_* \mid \mathbf{x}^{(i)}, \mathbf{y}\right)$, and a point that is perturbed by amount Δ with respect to variable $j, p\left(y_* \mid \mathbf{x}^{(i)} + \Delta_j, \mathbf{y}\right)$, they apply $d(p\|q) = \sqrt{2\mathcal{D}_{\mathrm{KL}}(p\|q)}$ to the posterior predictive distribution, so the measure of predictive relevance is:

$$r(i, j, \Delta) = \frac{d\left(p\left(y_* \mid \mathbf{x}^{(i)}, \mathbf{y}\right) \| p\left(y_* \mid \mathbf{x}^{(i)} + \Delta_j, \mathbf{y}\right)\right)}{\Delta} \tag{6}$$

where Δj is a vector of zeroes with Δ on the $j'th$ entry.
Then averaging this measure over all of the training points i yields a relevance estimate for the $j'th$ variable

$$\mathrm{KL}_j = \frac{1}{n}\sum_{i=1}^{n} r(i, j, \Delta) \tag{7}$$

From the averages, they rank input variables by relevance and desired number of them can be selected.

2.3 Variance of the Posterior Latent Mean

VAR is the second method introducded by Paananen et al, which ranks input variables based on the variability of the GP latent mean. Compared to the KL method, the VAR method only considers the latent mean, but examines it throughout the conditional distribution of each variable at the training point and not just the immediate vicinity of the point. This is advantageous because modelling the distribution of the inputs allows us to examine the GP posterior for out-of-sample behavior in a larger area of the input space than just at the training data location, and because computing the predictive mean of the GP is computationally chepaer than predicting the marginal variance.

In order to estimate the variance of the mean of the latent function, they approximate the distribution of the input variables. They use the sample mean μ and sample covariance Σ from n training points $X = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$. Then the conditional distribution of variable j at training point $i, p\left(x_j \mid \mathbf{x}_{-j}^{(i)}\right)$ is a multivariate normal:

$$\begin{aligned} x_j \mid \mathbf{x}_{-j} \sim \mathcal{N}\left(m_j, s_j^2\right) \\ m_j = \mu_j + \boldsymbol{\sigma}_{j,-j} \boldsymbol{\Sigma}_{-j,-j}^{-1} \left(\mathbf{x}_{-j} - \boldsymbol{\mu}_{-j}\right) \\ s_j^2 = \sigma_{j,j} - \boldsymbol{\sigma}_{j,-j} \boldsymbol{\Sigma}_{-j,-j}^{-1} \boldsymbol{\sigma}_{-j,j} \end{aligned} \tag{8}$$

Note: the subscript j refers to selecting the row or column j from μ or Σ, whereas the subscript $-j$ refers to excluding them.

The variance of the posterior mean along the $j'th$ dimension is then given by integrating over the conditional distribution $p\left(x_j \mid \mathbf{x}_{-j}^{(i)}\right)$

$$\begin{aligned} \operatorname{Var}\left[\bar{f}_j^{(i)}\right] = & \int \left(\bar{f}_j^{(i)}\right)^2 (x_j) \mathcal{N}\left(x_j \mid m_j, s_j^2\right) \mathrm{d}x_j \\ & - \left(\int \bar{f}_j^{(i)} (x_j) \mathcal{N}\left(x_j \mid m_j, s_j^2\right) \mathrm{d}x_j\right)^2 \end{aligned} \tag{9}$$

They approximate the variance with the Guass-Hermite quadrature by using a change of variables $k = (x_j - m_j) / \left(\sqrt{2} s_j\right)$.

$$\begin{aligned} \operatorname{Var}\left[\bar{f}_j^{(i)}\right] = & \int \left(\bar{f}_j^{(i)}\right)^2 \left(\sqrt{2} s_j k + m_j\right) \frac{e^{-k^2}}{\sqrt{\pi}} \mathrm{d}k \\ & - \left(\int \bar{f}_j^{(i)} \left(\sqrt{2} s_j k + m_j\right) \frac{e^{-k^2}}{\sqrt{\pi}} \mathrm{d}k\right)^2 \\ \approx & \, \pi^{-1/2} \sum_{n_k=1}^{N_k} w_i \left(\bar{f}_j^{(i)}\right)^2 \left(\sqrt{2} s_j k_i + m_j\right) \\ & - \pi^{-1} \left(\sum_{n_k=1}^{N_k} w_i \bar{f}_j^{(i)} \left(\sqrt{2} s_j k_i + m_j\right)\right)^2 \end{aligned} \tag{10}$$

where N_k is the number of weights w_i and evaluation points k_i of the Gauss-Hermite quadrature approximation. The k_i are given by the roots of the physicists version of the Hermite polynomial $H_{N_k}(k)$ and the weights are:

$$w_i = \frac{2^{N_k - 1} N_k! \sqrt{\pi}}{N_k^2 \left[H_{N_{k-1}}(k_i)\right]^2} \tag{11}$$

This procedure is repeated with all of the conditional distributions for each of the training points. Then the average is used as an estimate of the predictive relevance of the variable j.

$$\mathrm{VAR}_j = \frac{1}{n} \sum^{n} \operatorname{Var}\left[\bar{f}_j^{(i)}\right] \tag{12}$$

3 Synthetic Data Experiments

By reproducing the experiments using synthetic data by Paananen et al, I compare the relevance estimates for input features using the three methods described above.

The target variable y is constructed as a sum of eight independent and additive variables which have varying degrees of nonlinearity. Specifically, y is generated based on the inputs by:

$$\begin{aligned} y &= f_1\left(x_1\right) + \ldots + f_8\left(x_8\right) + \varepsilon \\ \varepsilon &\sim \mathcal{N}\left(0, 0.3^2\right) \\ f_j\left(x_j\right) &= A_j \sin\left(\phi_j x_j\right) \end{aligned} \tag{13}$$

where ϕ_j are angular frequencies, equally spaced between $\pi/10$ and π, and A_j are the scaling coefficients for the 8 components, such that the variance of each $f_j\left(x_j\right)$ is one. The two separate mechanisms for generating the input data are: $x_j \sim \mathrm{U}(-1,1)$ or $x_j \sim \mathcal{N}\left(0, 0.4^2\right)$. The Latent Functions $f_j\left(x_j\right), j = 1, \ldots, 8$ with uniform inputs are shown in Figure 1. Input 1 is the most linear and input 8 is the most nonlinear. The varying degrees of nonlinearity in the function allow us to examine how the three methods identify the true relevances of a range of variables.

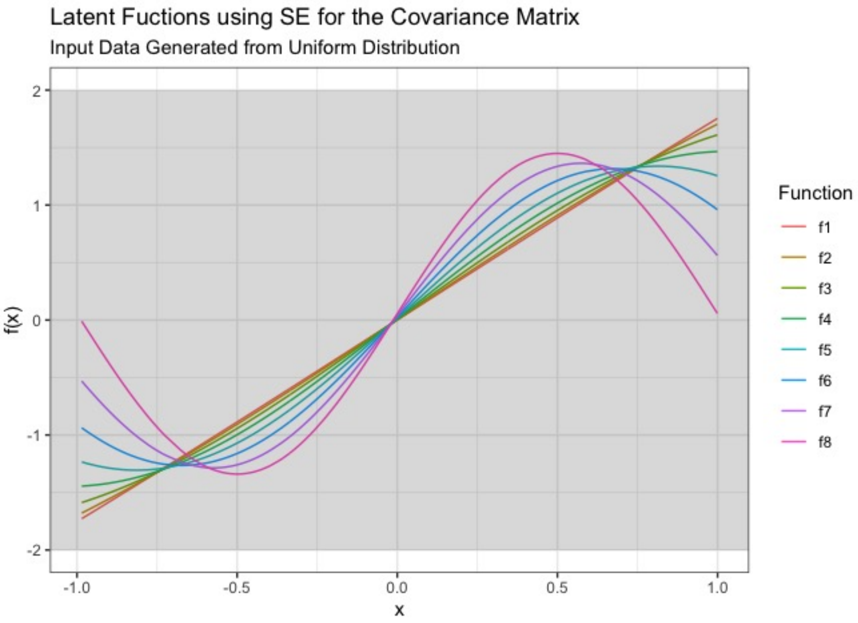

Figure 1: Latent Functions with Inputs from a Uniform Distribution

After averaging 200 runs and scaling so that the most relevant variable has a relevance of one, we plot the relevance estimates computed with ARD, KL, and VAR for the eight variables in Figure 2 and 3.

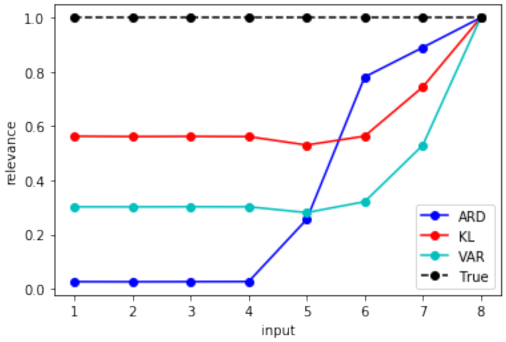

Figure 2: Relevance Estimates with Inputs from a Uniform Distribution

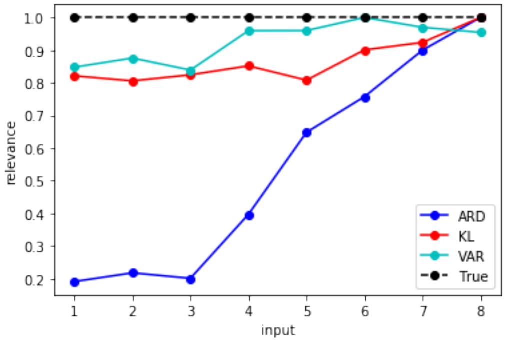

Figure 3: Relevance Estimates with Inputs from a Normal Distribution

The two methods, KL and VAR, are notably better than ARD in identifying the true relevances of the variables despite the varying degrees of nonlinearity. For uniform inputs, all three methods prefer nonlinear inputs over linear inputs. However, ARD assigns relevance values for the nonlinear inputs (1-4), very close to zero.

For Gaussian distributed inputs, the KL and VAR produce very similar results and assign a high relevance for all eight variables. However, the ARD method again assigns relevance values close to zero for three of the variables. Compared to the true relevance of all eight variables on the black line, it is clear that the methods proposed by Paananen et al are better than ARD in identifying true relevances of the variables, despite the varying degrees of nonlinearity.

4 Conclusion

Applying the KL and VAR variable section methods for Gaussian process modelling by Paananen et al allow us to utilize the predictions of a full model in a larger area of the training input area, rather than just at each training data point location. As a result, we can rank variables based on their predictive relevance. By comparing the commonly used automatic relevance determination method, which uses the inverse length-scale parameter of each input variable as a proxy for variable relevance, with the two variable selection methods for GP via the sensitivity analysis of the posterior predictive distribution, we have have reproduced their results on synthetic data.

An idea for further work would be to compare KL and VAR with alternative methods proposed for variable selection for GP, which also claim to address the issue of over-fitting when using automatic relevance determination. Qi et al [5] propose *predictive ARD* which approximates the integrals via Expectation Propogation. Compared to ARD, *predictive ARD* chooses the model with the best estimate of the predictive performance instead of choosing the one with the largest marginal likelihood. This argues against maximizing the evidence by tuning hyperparameters, and would therefore be an interesting comparison of results.

References

[1] Topi Paananen et al. *Variable selection for Gaussian processes via sensitivity analysis of the posterior predictive distribution.* 2017. arXiv: 1712.08048 [stat.ME].

[2] Hao Zhang. "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics". In: *Journal of the American Statistical Association* 99.465 (Mar. 2004), pp. 250–261. DOI: 10.1198/016214504000000241. URL: https://doi.org/10.1198/016214504000000241.

[3] Juho Piironen and Aki Vehtari. "Projection predictive model selection for Gaussian processes". In: *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, Sept. 2016. DOI: 10.1109/mlsp.2016.7738829. URL: https://doi.org/10.1109/mlsp.2016.7738829.

[4] Daniel Simpson et al. "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors". In: *Statistical Science* 32.1 (Feb. 2017), pp. 1–28. DOI: 10.1214/16-sts576. URL: https://doi.org/10.1214/16-sts576.

[5] Yuan (Alan) Qi et al. "Predictive automatic relevance determination by expectation propagation". In: *Twenty-first international conference on Machine learning - ICML '04*. ACM Press, 2004. DOI: 10.1145/1015330.1015418. URL: https://doi.org/10.1145/1015330.1015418.