STAT 518
Intro Student Presentation

Wen Wei Loh

April 11, 2013

Title of paper

BAYESIAN STATISTICS 6, pp. 475–501
J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (Eds.)

Regression and Classification Using Gaussian Process Priors

RADFORD M. NEAL
University of Toronto, Canada

SUMMARY

Gaussian processes are a natural way of specifying prior distributions over functions of one or more input variables. When such a function defines the mean response in a regression model with Gaussian errors, inference can be done using matrix computations, which are feasible for datasets of up to about a thousand cases. The covariance function of the Gaussian process can be given a hierarchical prior, which allows the model to discover high-level properties of the data, such as which inputs are relevant to predicting the response. Inference for these covariance hyperparameters can be done using Markov chain sampling. Classification models can be defined using Gaussian processes for underlying latent values, which can also be sampled within the Markov chain. Gaussian processes are in my view the simplest and most obvious way of defining flexible Bayesian regression and classification models, but despite some past usage, they appear to have been rather neglected as a general-purpose technique. This may be partly due to a confusion between the properties of the function being modeled and the properties of the best predictor for this unknown function.

- Radford M. Neal [1999]
- Bayesian Statistics, 6: 475-501, 1999

What the paper is about

- Regression and Classification
  - Flexible models not limited to simple parametric forms
  - Objective is to obtain a predictive distribution for the outcome of a future observation
- Gaussian Process Priors
  - Bayesian approach to obtain posterior distributions of model parameters
  - Integrate over model parameters

  $\Rightarrow$ predictive distribution depends only on the known observations
  - Gaussian Processes (GP) to define the covariance functions between outcomes

Linear regression example

- Observed data : $\left(x^{(1)}, t^{(1)}\right), \ldots, \left(x^{(n)}, t^{(n)}\right)$
- $t^{(i)} =$ outcome (target) for case i
- $x^{(i)} = \left[x_1^{(i)}, \ldots, x_p^{(i)}\right]$
  $=$ vector of p fixed inputs (predictors) for case i
- Fit a familiar model:

$$t^{(i)} = \alpha + \sum_{u=1}^{p} x_u^{(i)} \beta_u + \epsilon^{(i)}, \quad \epsilon^{(i)} \underset{iid}{\sim} N\left(0, \sigma_\epsilon^2\right)$$

- Put independent priors on unknown parameters α and β_u:

$$\alpha \sim N\left(0, \sigma_\alpha^2\right), \quad \beta_u \underset{iid}{\sim} N\left(0, \sigma_u^2\right)$$

Linear regression example

$\Rightarrow$ Prior joint multivariate Gaussian distribution for $t^{(i)}$:

$$\mathrm{E}\left[t^{(i)}\right] = \mathrm{E}\left[\alpha + \sum_{u=1}^{p} x_u^{(i)} \beta_u + \epsilon^{(i)}\right] = 0$$

$$\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] = \mathrm{E}\left[\left(\alpha + \sum_{u=1}^{p} x_u^{(i)} \beta_u + \epsilon^{(i)}\right)\left(\alpha + \sum_{u=1}^{p} x_u^{(j)} \beta_u + \epsilon^{(j)}\right)\right]$$

$$= \sigma_\alpha^2 + \sum_{u=1}^{p} x_u^{(i)} x_u^{(j)} \sigma_u^2 + \delta_{ij}\sigma_\epsilon^2$$

$\delta_{ij} = 1$ if $i = j$ and 0 otherwise i.e. Kronecker delta

$$C = \left\{\mathrm{Cov}\left[t^{(i)}, t^{(j)}\right] \;\; i, j \in [1, n]\right\}$$

Linear regression example

- Observed targets $t = \left[t^{(1)}, \ldots, t^{(n)}\right]^T \sim N_n(\mathbf{0}, C)$
- Given the inputs for a new case $x^{(n+1)}$, the predictive distribution is Gaussian:

$$\mathrm{E}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right] = k^T C^{-1} t$$
$$\mathrm{var}\left[t^{(n+1)} \mid t^{(1)}, \ldots, t^{(n)}\right] = V - k^T C^{-1} k$$

- $k = \left(\mathrm{Cov}\left[t^{(n+1)}, t^{(1)}\right], \ldots, \mathrm{Cov}\left[t^{(n+1)}, t^{(n)}\right]\right)^T$
- $V = \mathrm{Cov}\left[t^{(n+1)}, t^{(n+1)}\right] = \text{prior var}\left[t^{(n+1)}\right]$

Why use Gaussian processes?

- $\text{Cov}\left[t^{(i)}, t^{(j)}\right]$ is the key term in the predictive distribution

  e.g. linear combination of prior hyperparameters $\sigma_\alpha^2, \sigma_u^2, \sigma_\epsilon^2$ for linear regression
- Gaussian process procedure can handle more interesting and flexible models, simply by using a different covariance function

  e.g. regression model based on a class of smooth functions may be obtained with a covariance function:

$$\text{Cov}\left[t^{(i)}, t^{(j)}\right] = \eta^2 \exp\left(-\sum_{u=1}^{p} \rho_u^2 \left(x_u^{(i)} - x_u^{(j)}\right)^2\right) + \delta_{ij}\sigma_\epsilon^2$$

Covariance functions

- May be constructed with various parts representing prior variance/covariances e.g. constant, linear, exponential.
- Valid covariance function must always result in a positive definite covariance matrix for the targets.
- Various hyperparameters may be used to control:
  - amount of noise in the model
  - strength of association between each of the predictors with the target
  - sizes of the different additive components of the model.
- Posterior inference for these hyperparameters will reveal high-level structure in the data, and make model selection easier and more intuitive.

The title is "Regression and *classification* . . . "

- Suppose the targets $t^{(i)}$ are from the set $\{0, \ldots, K-1\}$.
- The distribution of $t^{(i)}$ would then be in terms of unobserved real-valued "latent" variables $y_1^{(i)}, \ldots, y_{K-1}^{(i)}$ for each case i

  e.g. Class probabilities for K classes:

$$\Pr\left(t^{(i)} = k\right) = \frac{\exp\left(y_k^{(i)}\right)}{\sum_{h=0}^{K-1} \exp\left(y_h^{(i)}\right)}$$

- The K latent values can then be given independent and identical Gaussian process priors.

Bayesian approach

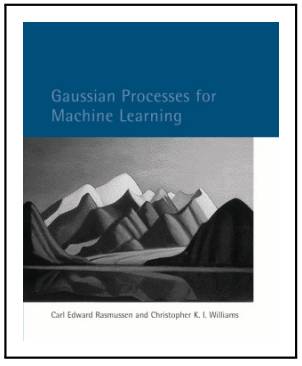
- "Latent" values for the targets have to be integrated over

$\Rightarrow$ Markov chain Monte Carlo methods e.g. Gibbs sampling

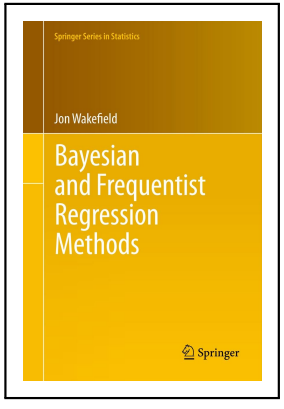
- Integrate over the posterior distribution for the hyperparameters of the covariance function e.g. $\eta^2, \rho_{\boldsymbol{u}}^2, \sigma_\epsilon^2$

$\Rightarrow$ Markov chain sampling e.g. hybrid Monte Carlo

Other literature



- Gaussian processes have recently received more attention with the introduction of kernel machines in machine learning
- [Rasmussen and Williams, 2006]



- Various nonparametric and flexible regression and classification methods
- Chapters 10-12 of [Wakefield, 2013].

Conclusion

- Why I chose this paper
  - Gaussian Process models are a form of flexible regression: letting the data speak for itself
  - Use of covariance functions could be a "natural" way of thinking about the underlying physical processes
- What's Next
  - Construct various covariance functions and the corresponding regression models (some graphs to plot)
  - Reproduce the simulation example (classification problem)
  - Implement the Bayesian procedure in `R`

Components of a covariance function

- Constant

→ same for any pair of cases, regardless of inputs x e.g. σ_α^2

- Linear

→ has form: $\sum_{u=1}^{p} x_u^{(i)} x_u^{(j)} \sigma_u^2$

- Exponential

→ has form: $\eta^2 \exp\left(-\sum_{u=1}^{p} \rho_u^R |x_u^{(i)} - x_u^{(j)}|^R\right)$,

→ $R \in [0, 2]$ for the covariance matrix to be positive definite

- Noise/Jitter

→ has form $\delta_{ij}\sigma_\epsilon^2$ or $\delta_{ij}J^2$

→ $\delta_{ij} = 1$ if $i = j$ and 0 otherwise

References

R M Neal. Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics*, 6: 475–501, 1999.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006. ISBN 978-0-262-18253-9.

Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, 2013.