Bayesian Kernel Methods

Unit 1: Bayes Rule, Approximate Inference, Hyperparameters

Unit 2: Gaussian Processes, Covariance Function, Kernel

Unit 3: GP: Regression

Unit 4: GP: Classification

Unit 5: Implementation: Laplace Approximation, Low Rank Methods

Unit 6: Implementation: Low Rank Methods, Bayes Committee Machine

Unit 7: Relevance Vector Machine: Priors on Coefficients

Unit 8: Relevance Vector Machine: Efficient Optimization and Extensions

http://mlg.anu.edu.au/∼smola/summer2002/

Regression with Gaussian Noise

THE AUSTRALIAN NATIONAL UNIVERSITY

Likelihood

For fixed s, we have additive normal noise in the observations. This means that $p(y_i|f(x_i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - [K\alpha]_i)^2\right)$.

Prior

Furthermore we have $\alpha \sim \mathcal{N}(0, S)$, where $S = \mathrm{diag}(s_1^2, \ldots, s_m^2)$.

Posterior

Since both prior and likelihood are normal, also $p(\alpha|X, Y, s)$ is normal. In particular, we get

$$\begin{aligned} -\log p(\alpha|X,Y,s) &= \frac{1}{2}(\mathbf{y} - K\alpha)^\top \sigma^{-2}(\mathbf{y} - K\alpha) + \frac{1}{2}\alpha^\top S^{-1}\alpha + \text{const.} \\ &= \frac{1}{2}\alpha^\top \underbrace{(K^\top \sigma^{-2} K + S^{-1})}_{:=\Sigma^{-1}}\alpha - \underbrace{\mathbf{y}^\top \sigma^{-2} K \Sigma}_{:=\mu^\top}\Sigma^{-1}\alpha + \text{const.} \end{aligned}$$

In other words,

$$\alpha \sim \mathcal{N}(\mu, \Sigma) \text{ where } \Sigma = (K^\top \sigma^{-2} K + S^{-1})^{-1} \text{ and } \mu = \sigma^{-2}\Sigma K^\top \mathbf{y}$$

Adjusting S

THE AUSTRALIAN NATIONAL UNIVERSITY

Effective Likelihood

By integrating out α we can contract the posterior into $p(Y|X,s)p(s)$, where

$$p(Y|X,s) = \int p(Y|X,\alpha)p(\alpha|s)d\alpha.$$

Since we have only normal distributions ($y = K\alpha + \xi$), this leads to

$$\mathbf{y} \sim \mathcal{N}(0, (\sigma^2\mathbf{1} + KSK^\top))$$

MAP2 Approximation

Maximize $p(Y|X,s)p(s)$ with respect to s, σ^2:

$$\underset{s,\sigma^2}{\text{maximize}}\ (2\pi)^{\frac{m}{2}}|\sigma^2\mathbf{1} + KSK|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}^\top(\sigma^2\mathbf{1} + KSK)^{-1}\mathbf{y}\right) p(s)p(\sigma).$$

To find the optimal solution, we take derivatives with respect to s, σ^2 and minimize.

Adjusting S, Part II

THE AUSTRALIAN NATIONAL UNIVERSITY

Goal

We want to compute derivatives of

$$\log\det(\sigma^2\mathbf{1} + KSK) + \mathbf{y}^\top(\sigma^2\mathbf{1} + KSK)^{-1}\mathbf{y}$$

with respect to σ^2 and all s_i^2 and find fixed-point update equations.

Matrix Magic

- $\partial_t A^{-1} = A^{-1}(\partial_t A)A^{-1}$
- $\frac{d}{dA}\log\det A = A^*$
- $|A||C - B^\top A^{-1}B| = |C||A - BC^{-1}B^\top|$ (Schur complements)
- $(A + BCB^\top)^{-1} = A^{-1} - A^{-1}B(C + B^\top AB)^{-1}B^\top A^{-1}$ (Sherman-Morrison-Woodbury).

Update Equations

$$\sigma^2 \leftarrow \frac{\|\mathbf{y} - \Sigma\mu\|^2}{m - \sum_{i=1}^n \xi_i}, \quad s_i^2 \leftarrow \frac{\mu_i^2}{\xi_i}, \quad \xi_i := 1 - s_i^{-2}\Sigma_{ii}$$

The Relevance of S

Recall

$$\sigma^2 \leftarrow \frac{\|\mathbf{y} - \Sigma\mu\|^2}{m - \sum_{i=1}^n \xi_i}, \quad s_i^2 \leftarrow \frac{\mu_i}{\xi_i}, \quad \xi_i := 1 - s_i^{-1}\Sigma_{ii}$$

where $\Sigma = (K^\top \sigma^{-2} K + S^{-1})^{-1}$ and $\mu = \sigma^{-2}\Sigma K^\top \mathbf{y}$.

Sparsity

It turns out that many s_i rapidly converge to 0. These coefficients can be removed, which makes computing Σ less costly.

The sparsity comes from the effective prior (if we integrate out over the hyperprior).

Variance

The variables ξ_i essentially denote how much the liberty in α_i is exploited, that is, $m - \sum_{i=1}^n \xi_i$ denotes the number of *free* parameters.

From classical statistics we know that the residual error can be estimated as a multiple of the number of free parameters and the additive noise.

General Case

THE AUSTRALIAN NATIONAL UNIVERSITY

Non-Gaussian Likelihood

Minimization of the negative log-posterior cannot be carried out explicitly any more as in the case of Normal additive noise.

Laplace Approximation

A quadratic approximation at the minimum can be used to obtain approximate confidence intervals (we approximate three times: MAP, MAP2, Laplace Approximation).

Practical Solution

Newton method or Fisher Scoring (compute the expectation of the Hessian) leads to rapid convergence.

Classification

Completely analogous to GP Classification.

Non-Gaussian Likelihood Revisited

THE AUSTRALIAN NATIONAL UNIVERSITY

Idea

We managed to avoid the MAP estimation in regression with normal noise by using a Gaussian prior and a Gaussian additive noise model.

Can we use the RVM trick also for the likelihood?

Decomposing the Likelihood

Rewrite $p(y_i|f(x_i))$ as $\int p(y_i|f(x_i), t_i)p(t_i)dt_i$, where $p(y_i|f(x_i), t_i)$ is a Normal distribution with zero mean and Variance t_i^2.

Effective Likelihood

If we fix t (hyperprior for likelihood) and s (hyperprior for prior), we obtain

$$\mathbf{y} \sim \mathcal{N}(0, (T + KSK^\top))$$

where $T = \mathrm{diag}(t_1^2, \ldots, t_m^2)$.

Gaussianized Kernel Machine

THE AUSTRALIAN NATIONAL UNIVERSITY

Update Equations

After long and tedious algebra we obtain

$$\begin{aligned}
\Sigma &= (S^{-1} + K^\top T^{-1} K)^{-1} \\
\mu &= \Sigma K^\top T^{-1} \mathbf{y}
\end{aligned}$$

$$\begin{aligned}
s_i^2 &\leftarrow \frac{\mu_i^2}{\xi_i} \text{ where } \xi_i = 1 - s_i^{-2}\Sigma_{ii} \\
t_i^2 &= \frac{(y - [K\mu]_i)^2}{1 - t_i^{-2}[K\Sigma K^\top]_{ii}}
\end{aligned}$$

Consequence

Update equations are not much more expensive than in the Gaussian Regression case (we have to update the $[K\Sigma K^\top]_{ii}$ terms, though).

Exact integration over prior in exchange for the approximation when performing MAP2 over hyperprior.