Multi-class Gaussian process classification of single-trial MEG based on frequency specific latent features extracted with linear binary classifiers

Pasi Jylänki, Jaakko Riihimäki, Aki Vehtari
Deptartment of Biomedical Engineering and Computational Science,
Aalto University, Finland
{pasi.jylanki, jaakko.riihimaki, aki.vehtari}@aalto.fi
http://www.becs.tkk.fi

The approach is based on calculating power features from the filtered MEG signals and doing a supervised linear dimensionality reduction for the gradiometer channel space. The dimensionality reduction is done with binary classifiers separately for each class and frequency band. The resulting lower dimensional features are classified using a multi-class Gaussian process classifier [2].

The Power features were extracted by calculating the mean squared amplitude from all the 204 planar gradiometer channels for each of the five prefiltered frequency bands. Logarithms of these power features were normalized to zero mean and unit variance separately for the both measurement days to give a 204-dimensional feature vector $\mathbf{x}_{i,k}$ for all the labeled observations $i = 1, ..., n$ and frequency bands $k = 1, .., K$, where $K = 5$.

Dimensionality reduction

Linear one-versus-rest logistic classifiers were used to reduce the 204-dimensional feature space into a one dimensional latent space for each of the five classes and five frequency bands separately. For a frequency band k and an input vector $\mathbf{x}_{i,k}$, the probability of class c is modeled as

$$p(y_{i,c} = 1 | \, \mathbf{w}_{k,c}, v_{k,c}, x_{i,k}) = (1 + \exp(-z_{i,k,c}))^{-1}, \tag{1}$$

where $\mathbf{w}_{k,c}$ are the coefficients of the linear predictor and $v_{k,c}$ a bias term, $z_{i,k,c} = \mathbf{x}_{i,k}^{\mathrm{T}} \mathbf{w}_{k,c} + v_{k,c}$ the latent value we are trying to estimate, and $y_{i,c} \in \{-1, 1\}$ a class label which is 1 for all the observations in the class c and -1 otherwise (see, e.g., [1]). To model the possible linear shifts in the power features between the different measurement days, a dummy variable $x_{i,0} \in \{-1, 1\}$ indicating the recording day, was included in $\mathbf{x}_{i,k}$ as an additional predictor. A Gaussian prior $p(\mathbf{w}_{k,c}) = \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$ with a variance parameter σ_w^2 was assumed for the linear coefficients, and also a Gaussian prior $v_{k,c} \sim \mathcal{N}(0, \sigma_v^2)$ was set for the bias term.

Combining the likelihood of all the labeled observations $\mathbf{y}_c = \{y_{1,c}, ..., y_{n,c}\}$ from the both measurement days with the priors results in a conditional posterior distribution

$$p(\mathbf{w}_{k,c}, v_{k,c} | \, \mathcal{D}_{k,c}, \sigma_w^2, \sigma_v^2) \propto \left(\prod_{i=1}^{n} (1 + \exp(-y_i z_{i,k}))^{-1} \right) p(\mathbf{w}_{k,c}) p(v_{k,c}), \quad (2)$$

where $\mathcal{D}_{k,c} = \{\mathbf{y}_c, \mathbf{X}_k\}$, $\mathbf{X}_k = [\mathbf{x}_{1,k}, ..., \mathbf{x}_{n,k}]^{\mathrm{T}}$. Since the posterior distribution (2) is analytically intractable an approximative inference method is required. The Laplace approximation was chosen because it is computationally convenient for the logistic model (see, e.g, [1, 2]). In the Laplace approximation a multivariate Gaussian approximation

$$q(\mathbf{w}_{k,c}, v_{k,c}) = \mathcal{N}(\boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c})$$

is formed by doing a second order Taylor expansion for

$$\log p(\mathbf{w}_{k,c}, v_{k,c} | \, \mathcal{D}_{k,c}, \sigma_w^2, \sigma_v^2)$$

around the posterior mode. Point estimates for the parameters σ_w^2 and σ_v^2 were determined by optimizing the approximative log marginal posterior distribution $\log q(\sigma_w^2, \sigma_v^2 | \, \mathcal{D}_{k,c})$ obtained by approximating the log marginal likelihood, $\log p(\mathbf{y}_c \, | \, \mathbf{X}_k, \sigma_w^2, \sigma_v^2)$, with the Laplace's method as described in [2]. Relatively flat half-Student-t priors with scale 10 and degrees of freedom $\nu = 10$ were assigned for the variance parameters to prevent them from becoming very large.

From the posterior approximation $q(\mathbf{w}_{k,c}, v_{k,c})$, a Gaussian approximation is obtained for the latent values related to both the labeled and unlabeled input vectors for class c:

$$q(z_{i,k,c}) = \mathcal{N}(m_{i,k,c}, V_{i,k,c}), \quad (3)$$

where $m_{i,k,c} = \mathbf{x}_{i,k}^{\mathrm{T}} \boldsymbol{\mu}_{k,c}$, $V_{i,k} = \mathbf{x}_{i,k}^{\mathrm{T}} \boldsymbol{\Sigma}_{k,c} \mathbf{x}_{i,k}$, and one is appended to the feature vector $\mathbf{x}_{i,k}$ to account for the bias $v_{k,c}$. The expected values $m_{i,k,c}$

from all the C classes and K frequency bands as well as the dummy variable $x_{i,0}$ indicating the recording day were combined to form new 26-dimensional input vectors $\mathbf{m}_i = [m_{i,1,1}, m_{i,2,1}, ..., m_{i,K,C}, x_{i,0}]$ for a multi-class classifier.

Multi-class classification

Using the latent vectors $\mathbf{m}_i$ as new inputs, the type of the video stimulus was predicted using a nonlinear Gaussian process (GP) multi-class classifier with a squared exponential covariance function [2]. The softmax function was used to model the class probabilities according to

$$p(\mathbf{y}_i \,|\, \mathbf{f}_i) = \exp(f_{i,c}) \left(\sum_{j=1}^{C} \exp(\mathbf{y}_i^{\mathrm{T}} \mathbf{f}_i) \right)^{-1}, \tag{4}$$

where $\mathbf{f}_i = [f_{i,1}, ..., f_{i,C}]^{\mathrm{T}}$ is a vector of the latent function values related to data point i and $\mathbf{y}_i = [y_{i,1}, ..., y_{i,C}]^{\mathrm{T}}$ is the corresponding target vector which has entry one for the correct class for the observation i and zero entries otherwise. Following [2], independent zero-mean GP priors were placed for each class, that is, $p(\mathbf{f}_c \,|\, l_{\mathrm{se}}, \sigma_{\mathrm{se}}^2) = \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{f}_c$ collect all the latent function values related to class c. The covariance matrix $\mathbf{K}$ is defined by the squared exponential covariance function

$$[\mathbf{K}]_{i,j} = k_{\mathrm{se}}(\mathbf{m}_i, \mathbf{m}_j \,|\, \theta) = \sigma_{\mathrm{se}}^2 \exp\left(-\frac{1}{l_{\mathrm{se}}^2} \sum_{l=1}^{d} (\mathbf{m}_{i,l} - \mathbf{m}_{j,l})^2 \right), \tag{5}$$

where $d = 26$, σ_{se}^2 is a magnitude parameter which scales the overall variation of the unknown function, and l_{se} is a length-scale parameter which governs how fast the correlation decreases as the distance increases in the input space.

Combining the likelihood of the observations $\mathbf{y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ with the priors $p(\mathbf{f}_c \,|\, l_{\mathrm{se}})$ results in an analytically intractable posterior distribution for the latent function values $\mathbf{f} = \{\mathbf{f}_1, ..., \mathbf{f}_n\}$, and again the Laplace approximation is used for approximate inference as described in [2]. The Laplace approximation results in a Gaussian posterior approximation for $\mathbf{f}$, and to approximate the predictive distribution it can be analytically combined with the conditional GP prior $p(\mathbf{f}_* \,|\, \mathbf{f}, \mathbf{m}, \mathbf{m}_*)$, where $\mathbf{m}$ collects the training inputs and $\mathbf{f}_*$ is a $C \times 1$ vector of latent values related to an unlabeled test input $\mathbf{m}_*$. Using the Laplace approximation also a marginal likelihood approximation $q(\mathbf{y} \,|\, \mathbf{m}, l_{\mathrm{se}}, \sigma_{\mathrm{se}}^2)$ can be obtained to determine point estimates of the parameters l_{se} and σ_{se}^2. However, optimiz-

ing the marginal likelihood resulted in a very small length scale and instead more conservative estimates $l_{\text{se}} = 2$ and $\sigma^2_{\text{se}} = 1$ were selected based on cross-validated predictive tests with the data from the second day. In practise, both the dimensionality reduction as well as the multi-class classification were implemented with the freely available GPstuff software package (`http://www.lce.hut.fi/research/mm/gpstuff/`).

Bibliography

[1] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, (2006)

[2] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, (2006)