Advanced probabilistic methods

Lecture 4: Sparse Bayesian linear models

Pekka Marttinen

Aalto University

January, 2018

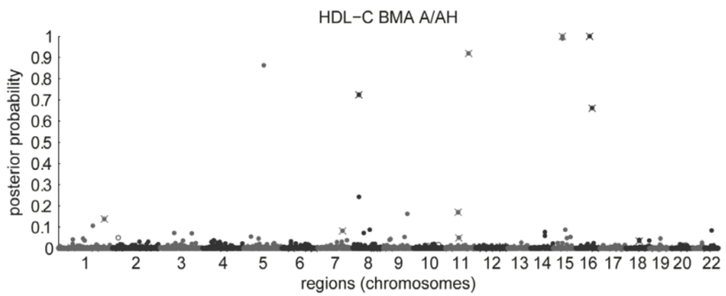
Lecture 4 overview[1]

- Bayesian Linear Parameter Models (LPMs)
  - Posterior computation (given fixed hyperparameters)
  - Determining hyperparameters
  - Example using radial basis functions
- Logistic regression for classification
  - Laplace approximation
- Barber, Ch. 18

[1] These slides build upon the book *Bayesian Reasoning and Machine Learning* and the associated teaching materials. The book and the demos can be downloaded from *www.cs.ucl.ac.uk/staff/D.Barber/brml*.

Example: genetic association studies

- Analysis of $\sim 1,000,000$ genetic polymorphisms in $\sim 50,000$ genomic regions (Peltola et al., 2012, *PLoS ONE*).
- *Spike-and-slab* prior on regression weights



Regression with Gaussian noise

- **Data** $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
  - $\mathbf{x}_i$: the input
  - y_i: the output
- **Model**:

$$y = \underbrace{f(\mathbf{w}, \mathbf{x})}_{\text{clean output}} + \underbrace{\eta}_{\text{noise}}, \quad \eta \sim N(0, \beta^{-1})$$

- In the simplest case

$$\begin{aligned} f(\mathbf{w}, \mathbf{x}) &= \mathbf{w}^T \mathbf{x} \\ &= w_1 x_1 + \ldots + w_D x_D \end{aligned}$$

- The *parameters* w_i are also called the *weights*

Bayesian linear parameter models

- A prior distribution $p(\mathbf{w}|\alpha)$ is placed on the weights $\mathbf{w}$.
- The posterior distribution $p(\mathbf{w}|\mathcal{D}, \Gamma)$ can be computed, and reflects the uncertainty of the parameters.

Prior distribution

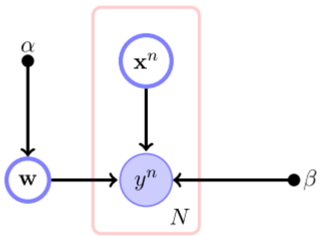
- A Gaussian prior distribution may placed on $\mathbf{w}$:

$$
\begin{aligned}
p(\mathbf{w}|\alpha) &= N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \\
&= \prod_{i=1}^{D} N(w_i|0, \alpha^{-1}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{D}{2}} e^{-\frac{\alpha}{2}\sum_i w_i^2}
\end{aligned}
$$

- Posterior

$$
\log p(\mathbf{w}|\Gamma, \mathcal{D}) = -\frac{\beta}{2}\sum_{i=1}^{N}\left[y_i - \mathbf{w}^T\mathbf{x}_i\right]^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const}
$$

Hyperparameters



- α: *precision* of the *regression weights*
  - determines the amount of regularization
  - large precision $\longrightarrow$ small variance $\longrightarrow$ weights are close to zero
- β: *precision* of the noise
- $\Gamma = \{\alpha, \beta\}$ are called the **hyperparameters** (in the course book...)

Posterior distribution

- Posterior distribution is obtained by completing the square (left as an exercise):

$$
p(\mathbf{w}|\Gamma, \mathcal{D}) = N(\mathbf{w}|\mathbf{m}, S)
$$

where

$$
S = \left(\alpha I + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}, \qquad \mathbf{m} = \beta S\sum_{i=1}^{N} y_i\mathbf{x}_i
$$

- Mean prediction

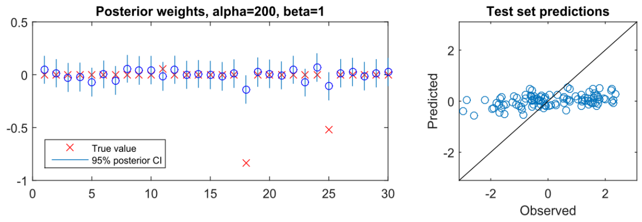
$$
\tilde{y} = \int \mathbf{w}^T\mathbf{x} \times p(\mathbf{w}|\Gamma, \mathcal{D})d\mathbf{w} = \mathbf{m}^T\mathbf{x}
$$

Example, impact of hyperparameters (1/3)

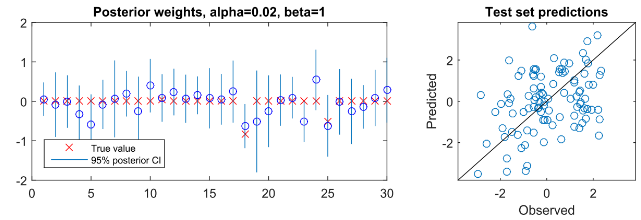
- Setup: simulate $y = \mathbf{w}_{true}^T \mathbf{x} + \epsilon$, where $\epsilon \sim N(0, \beta^{-1})$ and $\beta = 1$
- The goal is to investigate how hyperparameter α affects the posterior distribution of the parameters $\mathbf{w}$

Example, impact of hyperparameters (2/3)

- Too large α, $Var(y - \tilde{y}) = 1.54$ (Original $Var(y) = 1.75$)

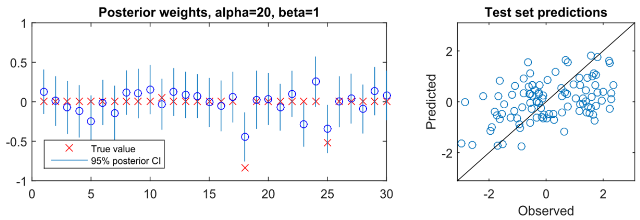


- Too small α, $Var(y - \tilde{y}) = 2.48$

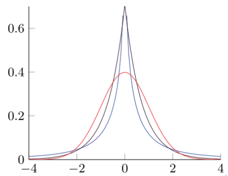


Example, impact of hyperparameters (3/3)

- About good α, $Var(y - \tilde{y}) = 1.46$
- A compromise between bias and variance

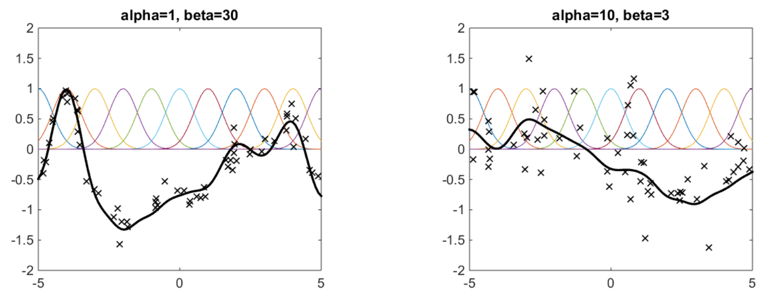


- Other sparse priors (e.g., Laplace, horse-shoe, spike-and-slab):



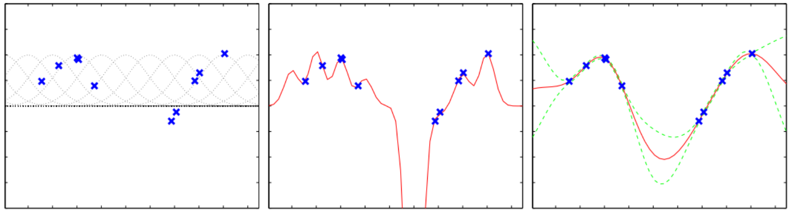
Non-linear transformation of the inputs

- Select $f(\mathbf{w}, \mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$
- $\phi(\mathbf{x})$ are the *basis functions*
- Example
  - weights drawn from $N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$; β is the noise precision.
  - $\mathbf{w} = (-0.7, 1.1, -0.8, -1.1, -0.8, -0.6, -0.6, 0.2, -0.2, 0.6, -0.9)$ for basis functions ordered from left to right (left panel)



Importance of learning hyperparameters

- (a): raw data and 15 radial basis functions $\phi_i(x) = \exp\left(-0.5(x - c_i)^2/\lambda^2\right)$ with $\lambda = 0.03^2$ and c_i spread evenly over the input space
- (b): predictions with $\beta = 100$ and $\alpha = 1$ (severe overfitting)
- (c): predictions with ML-II fitted hyperparameter values



(a) (b) (c)

Determining hyperparameters

- The hyperparameter posterior distribution is

$$p(\Gamma|\mathcal{D}) \propto p(\mathcal{D}|\Gamma)p(\Gamma)$$

- If $p(\Gamma) \approx$ *const* this is equivalent to

$$\Gamma^* = \arg\max_{\Gamma} p(\mathcal{D}|\Gamma),$$

where the **marginal likelihood**

$$p(\mathcal{D}|\Gamma) = \int p(\mathcal{D}|\Gamma, \mathbf{w})p(\mathbf{w}|\Gamma)d\mathbf{w}$$

- Selecting hyperparameters that maximize the marginal likelihood is called ML-II approach (in the book...)

ML vs. ML-II

- In **maximum likelihood**, we select parameter values **w** that maximize the log-likelihood

$$\log p(y|\mathbf{w}, \mathbf{x}) = \sum_{i=1}^{N} \log N(y_i|\mathbf{w}^T\phi(\mathbf{x}_i), \beta^{-1})$$
$$\widehat{\mathbf{w}} = \arg\max_{\mathbf{w}}\{\log p(y|\mathbf{w}, \mathbf{x})\} \quad \text{(does not depend on } \beta\text{)}$$

- In **ML-II**, we select hyperparameter values α and β that *maximize the (log-)marginal likelihood* (parameters **w** integrated out)

$$p(y|\Gamma, \mathbf{x}) = \int p(y|\Gamma, \mathbf{w}, \mathbf{x})p(\mathbf{w}|\Gamma)d\mathbf{w}$$
$$\Gamma^* = \arg\max_{\Gamma}\{\log p(y|\Gamma, x)\}$$

Hyperparameter optimization in practice

- EM-algorithm
- using the gradient
- compute log-marginal likelihood over a grid of values and choose the best value
- use some standard optimization routine (e.g. *fminunc*)

Alternative to ML-II: validation data (1/2)*

- Set the hyperparameters Γ to the value that minimizes the prediction error in the validation data

$$\{\mathcal{X}_{val}, \mathcal{Y}_{val}\} = \left\{ (\mathbf{x}_j^{val}, y_j^{val}), j = 1, \dots, M \right\}.$$

- Mean squared error (MSE)

$$\mathrm{MSE}(\Gamma) = \frac{1}{M} \sum_{j=1}^{M} (y_j^{val} - \tilde{y}_j^{val})^2,$$

where

$$\tilde{y}_j^{val} = \mathbf{m}^T \phi(\mathbf{x}_j^{val}), \qquad \mathbf{m} = E(\mathbf{w} | \Gamma, \mathcal{X}_{train}, \mathcal{Y}_{train})$$

Alternative to ML-II: validation data (2/2)*

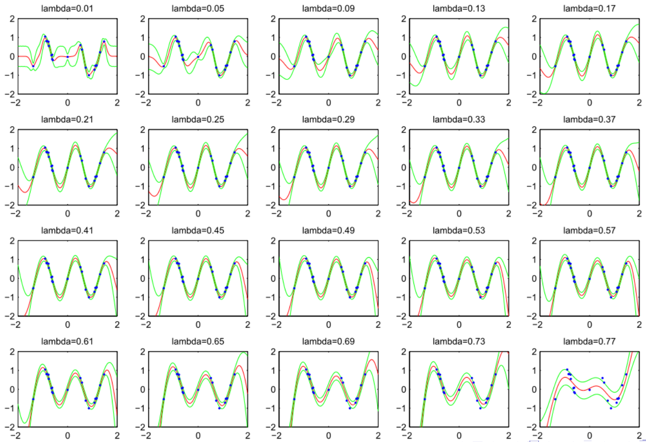
- Or by maximizing the validation data marginal likelihood

$$p(\mathcal{Y}_{val} | \Gamma, \mathcal{D}_{train}, \mathcal{X}_{val}) = \int_{\mathbf{w}} p(\mathcal{Y}_{val} | \mathbf{w}, \mathcal{X}_{val}, \Gamma) p(\mathbf{w} | \Gamma, \mathcal{X}_{train}, \mathcal{Y}_{train}) d\mathbf{w}$$

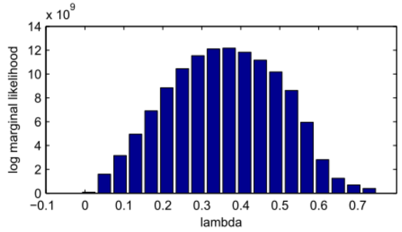
- Possible extension: *cross-validation*

Learning radial basis function width (1/2)

- A set of 10 evenly spaced radial basis functions is used $\phi_i(x) = \exp\left(-0.5(x - c_i)^2 / \lambda^2\right)$
- $\Gamma = (\alpha, \beta)$ optimized for different width parameters λ



Learning radial basis function width (2/2)



- The log marginal likelihood

$$\log p(\mathcal{D} | \lambda, \alpha^*(\lambda), \beta^*(\lambda))$$

having optimized α and β using ML-II. These values depend on λ.

- The best model corresponds to $\lambda = 0.37$.

Linear parameter models for classification

- Binary classification problem: $\mathcal{D} = \{(\mathbf{x}_i, c_i), i = 1, \ldots, N\}$, where the output $c \in \{0, 1\}$.
- Let p denote the probability that $p(c = 1|\mathbf{x})$
- Logistic (linear) regression

$$\log \frac{p}{1-p} = \mathbf{w}^T \mathbf{x}$$

- Or, equivalently

$$p(c = 1|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}),$$

where $\sigma(\cdot)$ is the so-called *logistic sigmoid*

$$\sigma(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

Logistic regression for classification

- When used for classification, the decision boundary is defined by $p(c = 1|\mathbf{x}) = p(c = 0|\mathbf{x}) = 0.5$. This corresponds to a hyperplane

$$\mathbf{w}^T \mathbf{x} = 0.$$

Classification rule

$$\mathbf{w}^T \mathbf{x} > 0 \rightarrow c = 1$$
$$\mathbf{w}^T \mathbf{x} < 0 \rightarrow c = 0$$

- Note: $\mathbf{x}$ can include a constant term, $\mathbf{x} = (1, x_1, \ldots, x_D)$, such that the *intercept* is automatically included

$$\mathbf{w}^T \mathbf{x} = w_0 + w_1 x_1 + \ldots + w_D x_D$$

Logistic regression, interpretation of parameters*

$$\log \left(\frac{p}{1-p} \right) = w_0 + w_1 x$$
$$\Leftrightarrow \frac{p}{1-p} = \exp(w_0 + w_1 x)$$

- Interpretation: when x increases by one unit, the **odds** $\frac{p}{1-p}$ of belonging in class 1 increases by a factor equal to e^{w_1}.
- If x is binary itself, $x \in \{0, 1\}$, then e^{w_1} is the **odds ratio** between classes $x = 1$ and $x = 0$.
  - a common term in medical literature, e.g., X='smoking', C='cancer'.

Prior for logistic regression

- Gaussian prior

$$p(\mathbf{w}|\alpha) = N_D(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \alpha^{\frac{D}{2}} (2\pi)^{-\frac{D}{2}} e^{-\frac{\alpha}{2}\mathbf{w}^T \mathbf{w}}$$

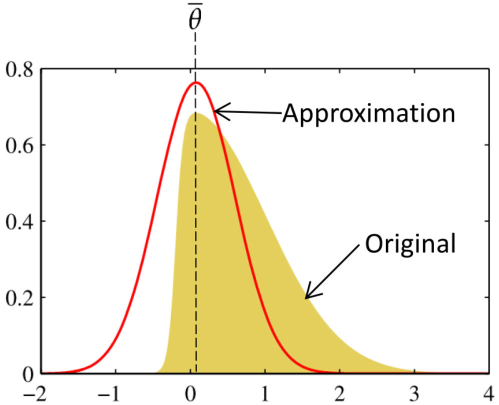
where α is the precision.
- Given $\mathcal{D} = \{(\mathbf{x}_i, c_i), i = 1, \ldots, N\}$ the posterior equals

$$p(\mathbf{w}|\alpha, \mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w}, \alpha) p(\mathbf{w}|\alpha)}{p(\mathcal{D}|\alpha)} = \frac{1}{p(\mathcal{D}|\alpha)} p(\mathbf{w}|\alpha) \prod_{i=1}^{N} p(c_i|\mathbf{x}_i, \mathbf{w})$$

(not of standard form, Laplace approximation is feasible to compute).

Laplace approximation

- Gaussian approximation at the mode



modified from Bishop, Fig. 4.14

Laplace approximation to the posterior

- In general,

$$p(\mathbf{w}|\alpha, \mathcal{D}) \propto \exp(-E(\mathbf{w})), \quad E(\mathbf{w}) = -\log p(\mathbf{w}|\alpha, \mathcal{D}).$$

- For logistic regression,

$$E(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \log \sigma(\mathbf{w}^T\mathbf{h}_i), \quad \mathbf{h}_i \equiv (2c_i - 1)\mathbf{x}_i.$$

1. approximate $E(\mathbf{w})$ by a quadratic function $\widetilde{E}(\mathbf{w})$ around the minimum $\overline{\mathbf{w}}$

$$\widetilde{E}(\mathbf{w}) = E(\overline{\mathbf{w}}) + \frac{1}{2}(\mathbf{w}-\overline{\mathbf{w}})^T H_{\overline{\mathbf{w}}}(\mathbf{w}-\overline{\mathbf{w}})$$

2. obtain a Gaussian approximation $q(\mathbf{w}|\alpha, \mathcal{D}) \propto \exp(-\widetilde{E}(\mathbf{w}))$ to $p(\mathbf{w}|\alpha, \mathcal{D})$

Laplace approximation in practice

- In practice:
  - Find the minimum of $E(\mathbf{w})$ by numerical optimization, e.g. Newton's method:
  $$\mathbf{w}^{new} = \mathbf{w} - \mathbf{H}_w^{-1}\nabla E$$
  - When converged, compute the Hessian $H_{\overline{\mathbf{w}}}$ of $E(\mathbf{w})$ at $\overline{\mathbf{w}}$.
  - The posterior approximation is

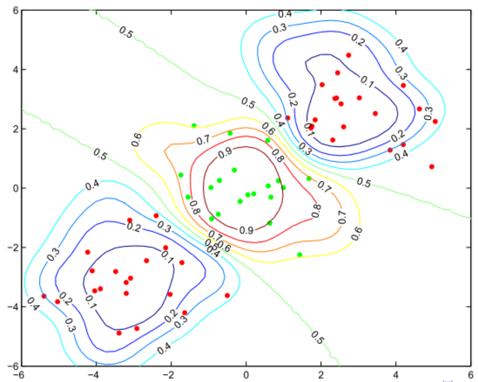
$$q(\mathbf{w}|\alpha, \mathcal{D}) = N(\mathbf{w}|\mathbf{m}, \mathbf{S}), \quad \mathbf{m} = \overline{\mathbf{w}}, \quad \mathbf{S} = \mathbf{H}_{\overline{\mathbf{w}}}^{-1}.$$

- Reminder: if $f \equiv f(x_1, \ldots, x_n)$

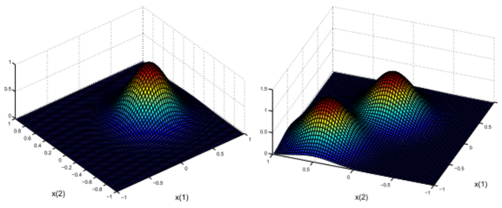
$$H_f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Example

- Bayesian logistic regression with RBF functions $\phi_i(\mathbf{x}) = \exp(-\lambda(\mathbf{x} - \mathbf{m}_i)^2)$.
- $\mathbf{m}_i$ placed on a subset of training points, λ set to 2
- Hyperparameter α optimized as with the Bayesian linear regression by maximizing the approximated marginal likelihood ($\rightarrow \alpha = 0.45$).



Curse of dimensionality



- In the 1-dimensional example, we used 10 radial basis functions
- To cover 2D region with same resolution, we would need 10^2 basis functions
- **Curse of dimensionality**: the number of basis functions required scales exponentially w.r.t. the dimension

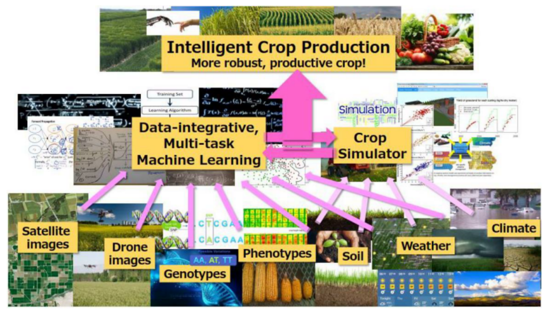
General comments on usage

- Curse of dimensionality limits the use of RBFs to low-dimensional cases
  - Possible remedy: place basis functions on observations
  - Alternatives: kernel methods, Gaussian processes
- With sparse priors, standard linear models can be used with very large D
  - $y = \sum_{i=1}^{D} w_i x_i + \epsilon$

Important points

- By placing a Gaussian prior on the parameters of linear regression, the posterior is also Gaussian.
- In classification, no closed form solution is available for logistic regression and approximations, e.g., the Laplace approximation, are needed.
- Hyperparameters can be set by maximizing the marginal likelihood (either exact or approximate).

Advertisement: PhD student position in machine learning



4-year funded PhD student positions in machine learning.
The student can start by doing a Master's thesis.

Goal: to develop new machine learning techniques and combine these with climate and ecological simulation models, in order to develop methods for intelligent plant breeding, which are needed in response to the changing climate to feed the rapidly growing global population.

Further information:
Hiroshi Mamitsuka (hmamitsuka.work@gmail.com)
Pekka Marttinen (pekka.marttinen@aalto.fi)
Jussi Gillberg (jussi.gillberg@aalto.fi)