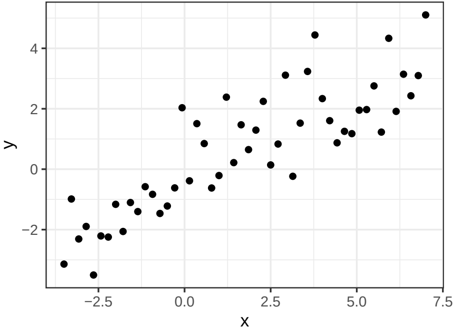
AutoML: Gaussian Processes

The Bayesian Linear Model

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Review: The Bayesian Linear Model I

Let $\mathcal{D}_{\text{train}} = \left\{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(n)}, y^{(n)})\right\}$ be a training set of i.i.d. observations from some unknown distribution.



Let $\mathbf{y} = (y^{(1)}, ..., y^{(n)})^\top$ and $\mathbf{X} \in \mathbb{R}^{n \times p}$ be the design matrix where the i-th row contains vector $\mathbf{x}^{(i)}$.

Review: The Bayesian Linear Model II

The linear regression model is defined as

$$y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)} = \boldsymbol{\theta}^\top \mathbf{x}^{(i)} + \epsilon^{(i)}, \text{ for all } i \in \{1, \ldots, n\}.$$

The observed values $y^{(i)}$ differ from the function values $f(\mathbf{x}^{(i)})$ by some additive noise, which is assumed to be i.i.d. Gaussian

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2).$$

Review: The Bayesian Linear Model III

Let us assume we have **prior beliefs** about the parameter $\boldsymbol{\theta}$ that are represented in a prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$.

Whenever data points are observed, we update the parameters' prior distribution according to Bayes' rule

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{q(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y} \mid \mathbf{X})}_{\text{marginal}}}.$$

Review: The Bayesian Linear Model IV

The posterior distribution of the parameter $\boldsymbol{\theta}$ is again normal distributed (the Gaussian family is self-conjugate):

$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{A}^{-1}), \text{ where } \boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p.$$

Note: If the posterior distributions $p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$ are in the same probability distribution family as the prior $q(\boldsymbol{\theta})$, the prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})$.

Note: The Gaussian family is **self-conjugate** with respect to a Gaussian likelihood function: choosing a Gaussian prior for a Gaussian likelihood ensures that the posterior is also Gaussian.

Review: The Bayesian Linear Model V

Theorem:

- For a Gaussian prior on $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$, and
- a Gaussian likelihood $y \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}^\top \boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n)$,

the resulting posterior is Gaussian: $\mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top \mathbf{y}, \boldsymbol{A}^{-1})$, with $\boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$.

Proof:
Plugging in Bayes' rule and multiplying out yields

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) \quad &\propto \quad p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) - \frac{1}{2\tau^2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right] \\
&= \quad \exp\left[-\frac{1}{2}\left(\underbrace{\sigma^{-2}\mathbf{y}^\top \mathbf{y}}_{\text{doesn't depend on } \boldsymbol{\theta}} - 2\sigma^{-2}\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \sigma^{-2}\boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} + \tau^{-2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right)\right]
\end{aligned}
$$

This expression resembles a normal density - except for the term in red!