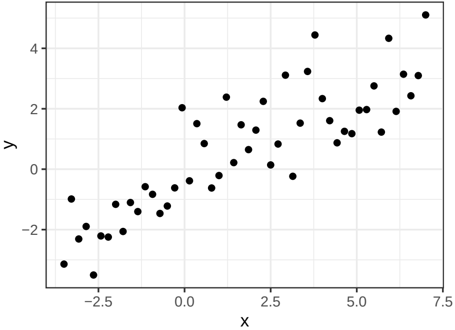
AutoML: Gaussian Processes

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

$$y = f(\mathbf{x}) + \epsilon = \boldsymbol{\theta}^\top \mathbf{x} + \epsilon$$

or on the data:

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

We now assume (from a Bayesian perspective) that also our parameter vector $\boldsymbol{\theta}$ is stochastic and follows a distribution.
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

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

and independent of $\mathbf{x}$ and θ.

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

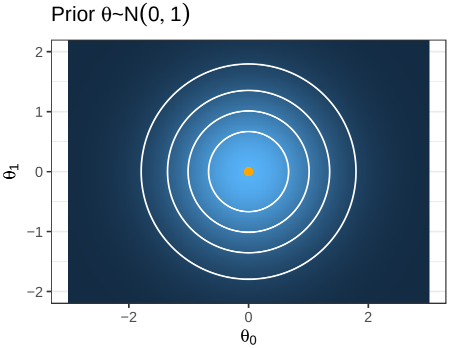
The posterior distribution of the parameter $\boldsymbol{\theta}$ is again normal distributed (the Gaussian family is self-conjugate):

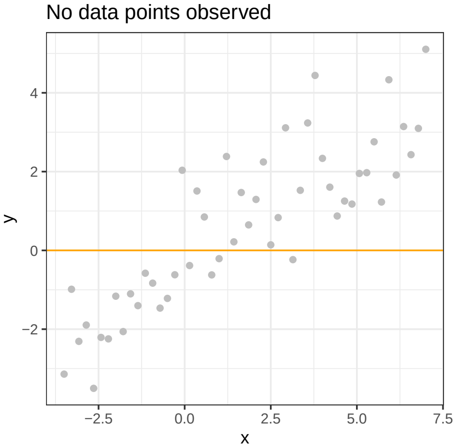
$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{A}^{-1}), \text{ where } \boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p.$$

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 $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})$.

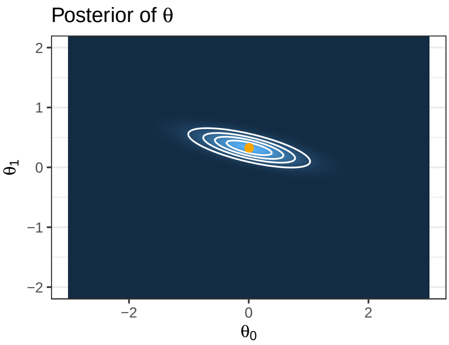
Note: The Gaussian family is **self-conjugate** with respect to a Gaussian likelihood function: choosing a Gaussian prior for a Gaussian likelihood ensures that the posterior is also Gaussian.

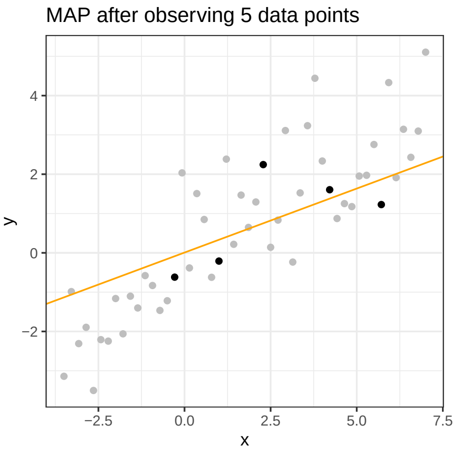
Review: The Bayesian Linear Model V



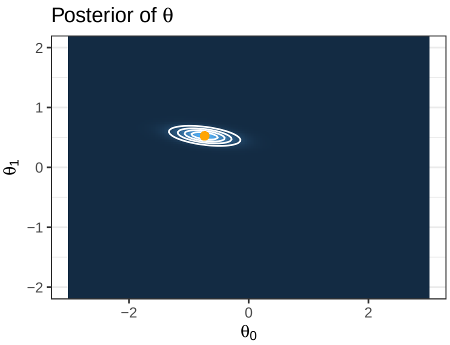


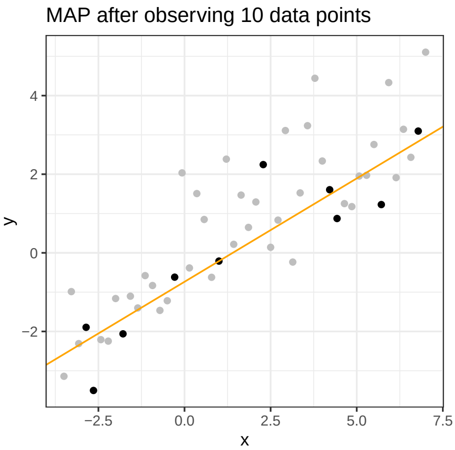
Review: The Bayesian Linear Model VI



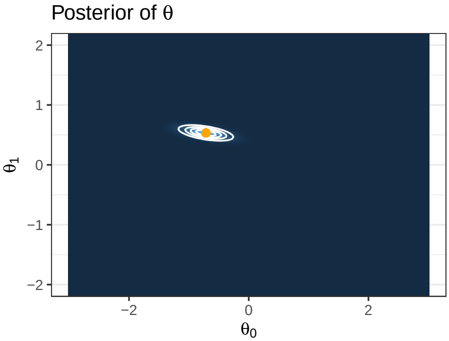


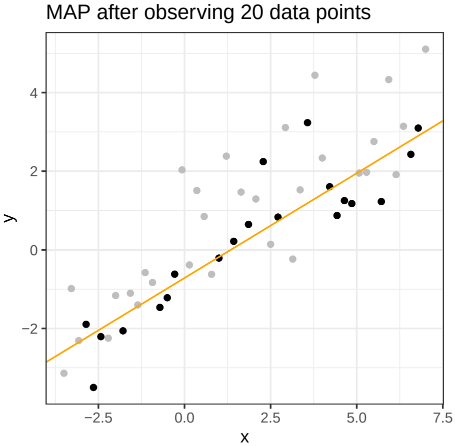
Review: The Bayesian Linear Model VII





Review: The Bayesian Linear Model VIII





Review: The Bayesian Linear Model IX

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

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

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

Review: The Bayesian Linear Model X

Note: We need not worry about the normalizing constant since its mere role is to convert probability functions to density functions with a total probability of one.

We subtract a (not yet defined) constant c while compensating for this change by adding the respective terms ("adding 0"), emphasized in green:

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

If we choose c such that $-c^\top \mathbf{A}\boldsymbol{\theta} + \sigma^{-2}\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} = 0$, the posterior is normal with mean c and covariance matrix $\mathbf{A}^{-1}$. Taking into account that $\mathbf{A}$ is symmetric, this is if we choose

$$
\begin{aligned}
& \sigma^{-2}\mathbf{y}^\top \mathbf{X} = c^\top \mathbf{A} \\
\Leftrightarrow \quad & \sigma^{-2}\mathbf{y}^\top \mathbf{X}\mathbf{A}^{-1} = c^\top \\
\Leftrightarrow \quad & c = \sigma^{-2}\mathbf{A}^{-1}\mathbf{X}^\top \mathbf{y}
\end{aligned}
$$

as claimed.

Review: The Bayesian Linear Model XI

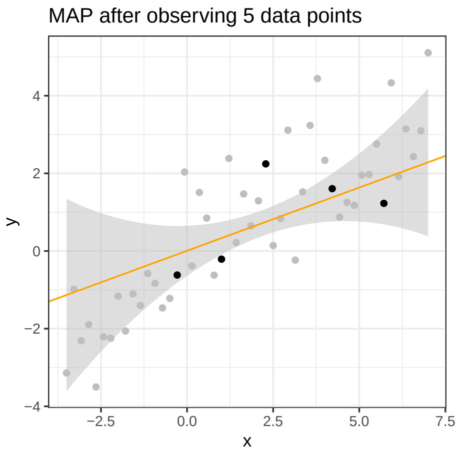
- Based on the posterior destribution, $\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top \mathbf{y}, \boldsymbol{A}^{-1})$, we can derive the predictive distribution for a new observations $\mathbf{x}_*$.

- The predictive distribution for the Bayesian linear model, i.e. the distribution of $\boldsymbol{\theta}^\top \mathbf{x}_*$, is

$$y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\sigma^{-2}\mathbf{y}^\top \mathbf{X}\mathbf{A}^{-1}\mathbf{x}_*, \mathbf{x}_*^\top \mathbf{A}^{-1}\mathbf{x}_*).$$

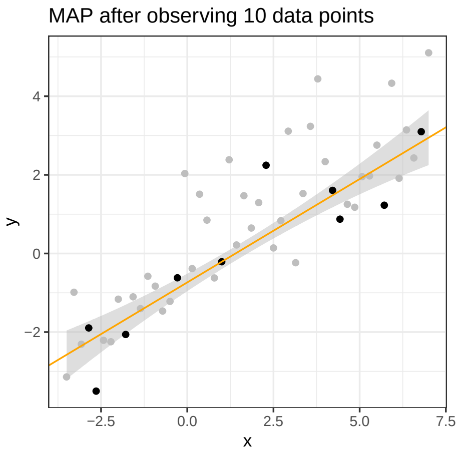
Note: This can be obtained by applying the rules for linear transformations of Gaussians.

Review: The Bayesian Linear Model XII



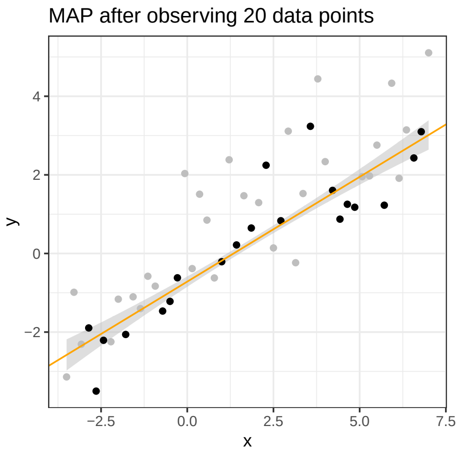
For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (the grey region, which equals $+/-$ two times the standard deviation).

Review: The Bayesian Linear Model XIII



For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (the grey region, which equals $+/-$ two times the standard deviation).

Review: The Bayesian Linear Model XIV



For every test input $\mathbf{x}_*$, we get a distribution over the prediction y_*. In particular, we get a posterior mean (orange) and a posterior variance (the grey region, which equals $+/-$ two times the standard deviation).

Summary: The Bayesian Linear Model

- By switching to a Bayesian perspective, we have not only point estimation for the parameter $\boldsymbol{\theta}$ but also whole **distributions**.

- From the posterior distribution of $\boldsymbol{\theta}$, we can derive a predictive distribution for $y_* = \boldsymbol{\theta}^\top \mathbf{x}_*$.

- We can perform online updates: the **posterior distribution** of $\boldsymbol{\theta}$ can be updated whenever new datapoints are observed.

- In the next step, we would like go beyond the linear funtions and develop a theory for functions with general shapes.