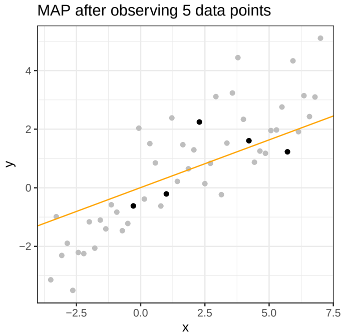
Advanced Machine Learning

The Bayesian Linear Model: Deep Dive



Learning goals

- Know the proof the posterior of bayesian linear model.
- Know how to derive the predictive distribution of bayesian linear model.

PROOF OF THE POSTERIOR OF BAYSIAN LM

Proof:
We want to show that

- for a Gaussian prior on $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$
- for 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 $\boldsymbol{A} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$.
Plugging in Bayes' rule and multiplying out yields

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

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

PROOF OF THE POSTERIOR OF BAYSIAN LM / 2

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:

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

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.

PREDICTIVE DISTRIBUTION

Based on the posterior distribution

$$\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{A}^{-1}\mathbf{X}^\top \mathbf{y}, \boldsymbol{A}^{-1})$$

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 that $y_* = \boldsymbol{\theta}^T \mathbf{x}_* + \epsilon$, where both the posterior of $\boldsymbol{\theta}$ and ϵ are Gaussians. By applying the rules for linear transformations of Gaussians, we can confirm that $y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_*$ is a Gaussian, too.