Project 10

Santiago Castro Dau, June Monge, Rachita Kumar, Sarah Lötscher

2022-05-05

Problem 28: Bayesian priors as regularizers

To recover the least squared estimator in the Bayesian setting one seeks to perform MAP estimation on the probability of the weights given the data $P(w|y_{1:n}, x_{1:n})$, since maximizing this probability is equivalent to finding the most probable weights $\hat{w}$, where w is the weight vector (note that we use w instead of β). Here $y_{1:n}$ are the labels and $x_{1:n}$ are the corresponding feature vectors. Decomposing $P(w|y_{1:n}, x_{1:n})$ using Bayes rule yields the following.

$$P(w|y_{1:n}, x_{1:n}) = \frac{P(w, y_{1:n}, x_{1:n})}{P(y_{1:n}, x_{1:n})}$$
$$= \frac{P(y_{1:n}|w, x_{1:n})P(w|x_{1:n})P(x_{1:n})}{P(y_{1:n}|x_{1:n})P(x_{1:n})}$$

Since the prior doesn't depend on the data we can simplify $P(w|x_{1:n})$ to $P(w)$.

$$= \frac{P(y_{1:n}|w, x_{1:n})P(w)}{P(y_{1:n}|x_{1:n})}$$

Since we are performing MAP estimation we seek to maximized this probability by finding the best set of weights $\hat{w}$.

$$\hat{w} = \underset{w}{\operatorname{argmax}} \frac{P(y_{1:n}|w, x_{1:n})P(w)}{P(y_{1:n}|x_{1:n})}$$

First we will derive the expression for the likelihood $P(y_{1:n}|w, x_{1:n})$ by performing MLE, which we can then use in this primary optimization problem.

The likelihood

The linear regression model is given by $Y = h(X) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $h(X) = w^T X$. Then we seek to find a simplified expression of this quantity $P(y_{1:n}|w, x_{1:n})$ by performing MLE, which means finding w such that $P(y_{1:n}|w, x_{1:n})$ is maximized. Since σ^2 is assumed to be fixed, we can omit it in our expressions of conditional probability.

$$\hat{w} = \underset{w}{\operatorname{argmax}} P(y_{1:n}|x_{1:n}, w)$$

Assuming that all n data points are $i.i.d$ and by applying the log, we can rewrite this expression as follows.

$$= \underset{w}{\operatorname{argmin}} - \sum_{i=1}^{n} \log P(y_i|x_i, w)$$

Then we can plug in our assumptions about the model for each data point.

$$\begin{aligned}
-\log P(y_i|x_i, w, \sigma^2) &= -\log \mathcal{N}(y_i; w^T x_i, \sigma^2) \\
&= -\log \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(y - w^T x)^2}{2\sigma^2} \\
&= \frac{1}{2}\log 2\pi\sigma^2 + \frac{(y - w^T x)^2}{2\sigma^2}
\end{aligned}$$

Then for the whole data set we can wirte the following.

$$\begin{aligned}
\hat{w} &= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (\frac{1}{2}\log 2\pi\sigma^2 + \frac{(y - w^T x)^2}{2\sigma^2}) \\
&= \underset{w}{\operatorname{argmin}} \frac{n}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y - w^T x)^2
\end{aligned}$$

Since the firs term is independent of w we can eliminate it, yielding the expression for the least square loss. From this result we can deduce that least squares linear regression is then equivalent to performing MLE for a conditional linear Gaussian likelihood. One can also clearly see from this demonstration that by performing OLS regression one implicitly places some statistical assumptions on the data, namely that the error is normally distributed and independent of the input (homoscedasticity). This assumption is characteristic of the least squared loss.

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n}(y - w^T x)^2$$

Comming back to our original problem

Now that we have an expression for the likelihood $P(y_{1:n}|x_{1:n}, w, \sigma^2)$ we can plug it into our original problem along with a Laplassian prior and aplaying the log function to the whole expression. By doing this we obtain the Lasso regression estimator.

$$\begin{aligned}
\hat{w} &= \underset{w}{\operatorname{argmin}} -\log P(w) - \log P(y_{1:n}|w, x_{1:n}) + \underbrace{\log P(y_{1:n}|x_{1:n})}_{\text{indp. of w}} \\
&= \underset{w}{\operatorname{argmin}} -\log \prod_{j=1}^{p} \frac{1}{2b}\exp{(\frac{-|w_j|}{b})} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y - w^T x)^2
\end{aligned}$$

$$-\log \prod_{j=1}^{p} \frac{1}{2b}\exp{(\frac{-|w_j|}{b})} = -(\underbrace{\sum_{j=1}^{p}\log\frac{1}{2b}}_{\text{indp.of w}} + \sum_{j=1}^{p}\log\exp{(\frac{-|w_j|}{b})})$$

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{b}\sum_{j=1}^{p}|w_j| + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y - w^T x)^2$$

Finally we can group together the constants into a single one λ and write the expression in matrix notation.

$$\hat{w} = \underset{w}{\operatorname{argmin}} \lambda||w||_1 + \frac{1}{2}||\mathrm{y} - \mathrm{X}w||_2^2$$