Bayesian Regression

Assaf Dvora

June 2019

1 The Linear Regression Problem

Assume we have a training set $\mathcal{D}$ of n observations, $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)|i = 1, .., n\}$, where $\boldsymbol{x} \in \mathbb{R}^d$ denotes an input vector (covariates) of dimension $\mathcal{D}$ and y denotes a scalar output or target (dependent variable). The columns vector inputs for all n cases are aggregated in the $D \times n$ design matrix $X \in \mathbb{R}^{D\times n}$ and the targets are collected in the vector $\boldsymbol{y}$. We can write $\mathcal{D} =(X, \boldsymbol{y})$. We are interested in making inferences about the relationship between inputs and targets, i.e. the conditional distribution of the targets given the inputs (but we are not interested in modeling the input distribution itself.

2 Bayesian Linear Regression

Linear regression assumes that there is a linear relationship between inputs and outputs, up to some noise term ϵ:

$$y = \boldsymbol{x}^T\boldsymbol{w} + \epsilon, \tag{1}$$

where $\boldsymbol{w}$ is a vector of weights (parameters) of the linear model. Often bias term is added, but as this can be implemented by augmenting the input vector $\boldsymbol{x}$ with additional element whose value is always one, we do not explicitly include it in our notation. We further assume that this noise term follows an IID Gaussian distribution with zero mean and variance σ_ϵ^2

$$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2). \tag{2}$$

The likelihood results in

$$\begin{aligned} p(\boldsymbol{y}|X, \boldsymbol{w}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\epsilon^2}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2}{2\sigma_\epsilon^2}\right] = \\ &= \frac{1}{(2\pi\sigma_\epsilon^2)^{n/2}} \exp\left[-\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{w}_i)^2\right] = \mathcal{N}(X^T\boldsymbol{w}, \sigma_\epsilon^2 I). \end{aligned} \tag{3}$$

In the Bayesian formalism we need to specify a *prior* over the parameters. We put a zero mean Gaussian prior with covariance matrix Σ_p on the weights

$$\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p). \tag{4}$$

The posterior distribution over the weights is given by,

$$p(\boldsymbol{w}|\boldsymbol{y}, X) = \frac{p(\boldsymbol{y}|X, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|X)} = \frac{p(\boldsymbol{y}|X, \boldsymbol{w})p(\boldsymbol{w})}{\int p(\boldsymbol{y}|X, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}}. \tag{5}$$

Since the prior and the likelihood are conjugate Gaussian distributions, the posterior is Gaussian:

$$p(\boldsymbol{w}|\boldsymbol{y}, X) = \mathcal{N}(\boldsymbol{\mu}_w, \Sigma_w), \tag{6}$$

and "completing the square" we obtain,

$$\mu_w = \frac{1}{\sigma_\epsilon^2} A^{-1} X \boldsymbol{y}, \quad \Sigma_w = A^{-1}, \tag{7}$$

where $A = \sigma_\epsilon^{-2} X X^T + \Sigma_p^{-1}$. We note that for Gaussian posterior the mean vector is also the MAP estimate of $\boldsymbol{w}$.

To make predictions for a test case we average over all possible parameter values, weighted by their posterior probability. Thus, the predictive probability is given by,

$$p(y_*|\boldsymbol{x}_*, X, \boldsymbol{y}) = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{y}, X)d\boldsymbol{w} \tag{8}$$

in the case of linear model with Gaussian posterior, the predictive probability is also Gaussian,

$$p(y_*|\boldsymbol{x}_*, X, \boldsymbol{y}) = \mathcal{N}(\frac{1}{\sigma_\epsilon^2} \boldsymbol{x}_*^T A^{-1} X \boldsymbol{y},\ \boldsymbol{x}_*^T A^{-1} \boldsymbol{x}_*). \tag{9}$$

3 Bayesian Linear Regression Using Gaussian Process.

Lets assume the following additive model for the response variable

$$y = f(\boldsymbol{x}) + \varepsilon, \tag{10}$$

where $f(\boldsymbol{x})$ is unobserved latent **stochastic process** indexed by the input vector $\boldsymbol{x}$, and ε is i.i.d Gaussian noise with variance σ_ε^2. Hence, the response variable satisfies

$$p(y|f(\boldsymbol{x}), \sigma_\varepsilon^2) = \mathcal{N}(f(\boldsymbol{x}), \sigma_\varepsilon^2). \tag{11}$$

We further assume that σ_ε^2 is known and hence we use $p(y|f(\boldsymbol{x}))$ instead.

Denote with $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_n)]^T \in R^n$ the vector of realizations of the random process. The likelihood of the model is given by

$$p(\mathcal{D}|\boldsymbol{f}) = p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{n} p(y_i|f(x_i), \sigma_\varepsilon^2) = \mathcal{N}(\boldsymbol{f}, \sigma_\varepsilon^2 I). \tag{12}$$

To define a prior over $\boldsymbol{f}$ we use the Bayesian Linear Regression model, $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$, with prior $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_w)$. We show in Appendix A that this results in a Gaussian process for $f(\boldsymbol{x})$ with the mean and covariance satisfying

$$\begin{aligned} E[f(\boldsymbol{x})] &= 0, \\ E[f(\boldsymbol{x})f(\boldsymbol{x}')] &\triangleq \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\Sigma_w\boldsymbol{x}'. \end{aligned} \tag{13}$$

Thus, the prior probability of the latent vector $\boldsymbol{f}$ is given by

$$\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, K), \tag{14}$$

where $K = X^T\Sigma_w X$. Based on Bayes theorem, the posterior probability can be written as

$$p(\boldsymbol{f}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{f})p(\boldsymbol{f})}{p(\mathcal{D})} \tag{15}$$

since the prior probability $p(\boldsymbol{f})$ and the likelihood function $p(\mathcal{D}|\boldsymbol{f})$ are conjugate Gaussian distributions, the posterior is Gaussian and satisifies

$$p(\boldsymbol{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p). \tag{16}$$

We show in Appendix B that the mean vector, $\boldsymbol{\mu}_p$, and the covaraince matrix, Σ_p are given by

$$\boldsymbol{\mu}_p = K(\sigma_\varepsilon^2 I + K)^{-1}\boldsymbol{y} \tag{17}$$

$$\Sigma_p = \sigma_\varepsilon^2(\sigma_\varepsilon^2 I + K)^{-1}K \tag{18}$$

For a test case, $\boldsymbol{x}_*$, we would like to predict the latent varaibale $f_* \triangleq f(\boldsymbol{x}_*)$ given the knowledge we have about our data. The posterior predictive distibution $p(f_*|\boldsymbol{x}_*\mathcal{D})$ can be written as

$$p(f_*|\mathcal{D}) = \int p(f_*|\boldsymbol{f})p(\boldsymbol{f}|\mathcal{D})d\boldsymbol{f} \tag{19}$$

where $p(f_*|\boldsymbol{f})$ is the (prior) conditional distibution of f_* given $\boldsymbol{f}$ and $p(\boldsymbol{f}|\mathcal{D})$ is the posterior distrbution of $\boldsymbol{f}$. We note that $\boldsymbol{f}$ and f_* are jointly Gaussian

$$\begin{bmatrix} \boldsymbol{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{0} \\ 0 \end{pmatrix}, \begin{pmatrix} K & \boldsymbol{k_*} \\ \boldsymbol{k}_*^T & \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{pmatrix}\right] \tag{20}$$

where $\boldsymbol{k}_* = [\mathcal{K}(\boldsymbol{x}_1, \boldsymbol{x}_*), \mathcal{K}(\boldsymbol{x}_2, \boldsymbol{x}_*), ..., \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}_*)]$. The conditional distribution $p(f_*|\boldsymbol{f})$ is also Gaussian and is given by (Rasmussen et al.)

$$p(f_*|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{k}_*^T K^{-1}\boldsymbol{f}, \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T K^{-1}\boldsymbol{k}_*). \tag{21}$$

In Appendix C we show that the posterior predictive distribution, $p(f_*|\mathcal{D})$ is Gaussian and satisfies

$$\mathbb{E}[f_*|\mathcal{D}] = \boldsymbol{k}_*^T K^{-1}\boldsymbol{\mu}_p \tag{22}$$

$$\text{var}[f_*|D] = \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T(K^{-1} - K^{-1}\Sigma_p K^{-1})\boldsymbol{k}_* \tag{23}$$

where $\boldsymbol{\mu}_p$ and Σ_p are the mean vector and covariance matrix of the posterior $p(\boldsymbol{f}|\mathcal{D})$ (respectively). Substituitng (17) to (22) we obtain the mean value of the posterior predictive distribution

$$\mathbb{E}[f_*|\mathcal{D}] = \boldsymbol{k}_*^T(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{y}. \tag{24}$$

To obtain an expression for the posterior predictive variance we first note that the posterior covariance matrix, Σ_p can be rearranged to

$$\begin{aligned}\Sigma_p &= \sigma_\varepsilon^2(\sigma_\varepsilon^2 I + K)^{-1}K = (I + \sigma_\varepsilon^{-2}K)^{-1}K \\ &= \left[K^{-1}(I + \sigma_\varepsilon^{-2}K)\right]^{-1} = (K^{-1} + \sigma_\varepsilon^{-2}I)^{-1}.\end{aligned} \tag{25}$$

Substituting (25) to (23) we obtain,

$$\text{var}[f_*|D] = \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T(K^{-1} - K^{-1}(K^{-1} + \sigma_\varepsilon^{-2}I)^{-1}K^{-1})\boldsymbol{k}_* \tag{26}$$

Using the matrix inversion lemma (Rasmussen et al., Appendix A), we obtain

$$\text{var}[f_*|D] = \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{k}_*. \tag{27}$$

Thus, the posterior predictive distribution satisfies

$$p(\boldsymbol{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{k}_*^T(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{y}, \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{k}_*) \tag{28}$$

Next we show that the posterior predictive distribution (28) is equal to the predictive probability (9) obtained for the Bayesian Linear Regression problem (Section 2). Using the linear covaraince defined in (13) we obtain

$$K = X^T\Sigma_w X \tag{29}$$

$$\boldsymbol{k}_* = X^T\Sigma_w\boldsymbol{x}_* \tag{30}$$

$$\mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) = \boldsymbol{x}_*^T\Sigma_w\boldsymbol{x}_* \tag{31}$$

Substituting into the varaince expression (27) and using the matrix inverse lemma (see Appendix D) we obtain,

$$\begin{aligned}\text{var}[f_*|D] &= \boldsymbol{x}_*^T\Sigma_w\boldsymbol{x}_* - \boldsymbol{x}_*^T\Sigma_w X(X^T\Sigma_w X + \sigma_\varepsilon^2 I)^{-1}X^T\Sigma_w\boldsymbol{x}_* = \\ &= \boldsymbol{x}_*^T(\Sigma_w - X(X^T\Sigma_w X + \sigma_\varepsilon^2 I)^{-1}X^T)\boldsymbol{x}_* = \\ &= \boldsymbol{x}_*^T(\Sigma_w^{-1} + \sigma_\varepsilon^{-1}XX^T)^{-1}\boldsymbol{x}_*\end{aligned}$$

which is equal to the varaince of the predictive probability in (9). Similarly, substituting (29)-(31) into the mean (24) we obtain,

$$\begin{aligned}\mathbb{E}[f_*|\mathcal{D}] &= \boldsymbol{k}_*^T(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{y} = \\ &= \boldsymbol{x}_*^T\Sigma_w X(X^T\Sigma_w X + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{y} =\end{aligned}$$

Since Σ_w is symmetric and postive-definite we can use a varaint of the matrix inverse lemma (Appendix D) to obtain

$$\mathbb{E}[f_*|\mathcal{D}] = \frac{1}{\sigma_\varepsilon^2}\boldsymbol{x}_*^T(\sigma_\varepsilon^{-2}XX^T - \Sigma_w^{-1})^{-1}X\boldsymbol{y}$$

which is equal to the mean of the predictive probability in (9).

Appendix A

Using the Bayesian Linear Regression model, the process $f(\boldsymbol{x})$ is defined as

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}, \quad \boldsymbol{w} \sim \mathcal{N}(0, \Sigma_w).$$

We would like to show that $f(\boldsymbol{x})$ is Gaussian process.
Definition A.1: A Gaussian process is a collection if random variables, any finite number of which have a joint Gaussian distribution.
Lemma A.1: A collection of Gaussian random variables are jointly Gaussian and form a joint Gaussian distribution if any linear combination of the variables is a Gaussian random variable.
Denote $f(x_i), i = 1, .., n$ the set of random variables corresponding to n input point x_i. Further denote aith $a_i \in \mathbb{R}, i = 1, .., n$ a set of deterministic values. Thus,

$$\sum_{i=1}^{n} a_i f(\boldsymbol{x}_i) = \sum_{i=1}^{n} a_i \boldsymbol{w}^T \boldsymbol{x}_i = \sum_{i=1} \boldsymbol{w}^T a_i \boldsymbol{x}_i = \boldsymbol{w}^T \underbrace{\sum_{i=1}^{n} a_i \boldsymbol{x}_i}_{\boldsymbol{Y}} = \boldsymbol{w}^T \boldsymbol{Y}.$$

Since $\boldsymbol{w}$ is Gaussian random vector,$\boldsymbol{w}^T \boldsymbol{Y}$ is Gaussian variable (linear combination of Gaussian variables) and according to Lemma A.1 $f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), ..., f(\boldsymbol{x}_n)$ are jointly Gaussian. Hence, according to Definition A.1 $f(\boldsymbol{x})$ is Gaussian process with mean and covariance function satisfying

$$E[f(\boldsymbol{x})] = 0$$
$$E[f(\boldsymbol{x}) f(\boldsymbol{x}')] = \boldsymbol{x}^T \Sigma_w \boldsymbol{x'}$$

Appendix B

From Eq. (15) we obtain

$$\begin{aligned}
\log p(\boldsymbol{f}|\mathcal{D}) &= C - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{f})^T(\boldsymbol{y} - \boldsymbol{f}) - \frac{1}{2}\boldsymbol{f}^T K^{-1} \boldsymbol{f} = \\
&= C - \frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{y}^T \boldsymbol{f} + \boldsymbol{f}^T \boldsymbol{f}) - \frac{1}{2}\boldsymbol{f}^T K^{-1}\ \boldsymbol{f} = \\
&= C + \frac{1}{\sigma_\varepsilon^2}\boldsymbol{y}^T\ \boldsymbol{f} - \frac{1}{2}\boldsymbol{f}^T \left[\sigma_\varepsilon^{-2} I + K^{-1}\right] \boldsymbol{f} = \\
&= C - \frac{1}{2}\{\boldsymbol{f}^T \left[\sigma_\varepsilon^{-2} I + K^{-1}\right] \boldsymbol{f} - 2\frac{1}{\sigma_\varepsilon^2}\boldsymbol{y}^T \boldsymbol{f}\}
\end{aligned}$$

Denotes $A = \left[\sigma_\varepsilon^{-2} I + K^{-1}\right]$ and $\boldsymbol{b} = \frac{1}{\sigma_\varepsilon^2}\boldsymbol{y}$. Completing the square yiedls

$$\begin{aligned}
\log p(\boldsymbol{f}|\mathcal{D}) &= C - \frac{1}{2}(\boldsymbol{f}^T A \boldsymbol{f} - 2\boldsymbol{b}^T \boldsymbol{f}) \\
&= C - \frac{1}{2}\left[(\boldsymbol{f} - A^{-1}\boldsymbol{b})^T A (\boldsymbol{f} - A^{-1}\boldsymbol{b}) + \boldsymbol{b}^T A^{-1} \boldsymbol{b}\right]
\end{aligned}$$

and hence,

$$\mathbb{E}[\boldsymbol{f}|\mathcal{D}] = K(\sigma_\varepsilon^2 I + K)^{-1}\boldsymbol{y}$$
$$\text{cov}[\boldsymbol{f}|\mathcal{D}] = \sigma_\varepsilon^2(\sigma_\varepsilon^2 I + K)^{-1}K$$

Appendix C

We would like to compute the posterior predictive distribution defined in Eq. (19) (reproducted here for clarity)

$$p(f_*|\mathcal{D}) = \int p(f_*|\boldsymbol{f})p(\boldsymbol{f}|\mathcal{D})d\boldsymbol{f}$$

where the distributions $p(f_*|\boldsymbol{f})$ and $p(\boldsymbol{f}|\mathcal{D})$ are Gaussian with

$$p(\boldsymbol{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$$
$$p(f_*|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{k}_*^T K^{-1}\boldsymbol{f}, \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{k}_*^T K^{-1}\boldsymbol{k}_*)$$

We use the follwoing property (Bishop et al., Section 2.3.3) of Gaussian distribution.
Property B.1: Consider a marginal Gaussian $p(\boldsymbol{x})$ and a Gaussian conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ in which $p(\boldsymbol{y}|\boldsymbol{x})$ has a mean that is a linear function of $\boldsymbol{x}$, and a covariance which is independent of $\boldsymbol{x}$

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \Lambda^{-1}),$$
$$p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|A\boldsymbol{x} + \boldsymbol{b}, L^{-1}).$$

Here $\boldsymbol{\mu}, A, \boldsymbol{b}$ are parameters governing the means, and Λ, L are the precision matrices (inverse coavariance matrices). If $\boldsymbol{x}$ has dimensionality M and $\boldsymbol{y}$ has dimensionality D then the matrix A has size $D \times M$. The marginal distribution $p(\boldsymbol{y})$ is Gaussian with mean and covaraince given by

$$\mathbb{E}[\boldsymbol{y}] = A\,\boldsymbol{\mu} + b,$$
$$\text{cov}[\boldsymbol{y}] = L^{-1} + A\Lambda^{-1}A^T$$

using property B.1 we obtain

$$\mathbb{E}[f_*] = \boldsymbol{k}_*^T K^{-1}\boldsymbol{\mu}_p$$
$$\text{var}[f_*] = \mathcal{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) + \boldsymbol{k}_*^T(K^{-1}\Sigma_p K^{-1} - K^{-1})\boldsymbol{k}_*$$

Appendix D

The *matrix inverse lemma*, also known as the Woodbury, Sherman & Morrison formula states that

$$(Z + UWV^T)^{-1} = Z^{-1} - Z^{-1}U(W^{-1} + V^T Z^{-1}U)^{-1}V^T Z^{-1}, \tag{32}$$

assuming the relevant inverses all exist. Here Z is $n \times n$, W is $m \times m$ and U and V are both of size $n \times m$.
For positive-definite matrices P and R we obtain the following varaint of the matrix inverse lemma

$$(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}.$$

We note that if P and R are also symmetric, $(P^{-1} + B^T R^{-1} B)^{-1}$ is symmetric function (symmetry is closed under summation and inverse) and hence

$$R^{-1} B (P^{-1} + B^T R^{-1} B)^{-1} = (B P B^T + R)^{-1} B P^T$$