Bayesian Linear Regression

Why there is a need for a Bayesian formulation?

- MLE (maximum likelihood estimation) is prone over-fitting (especially when the observed data is scarce)
- Oftentimes it desirable to produce uncertainty estimates

Idea: Place a "prior" over the unkown model parameters θ,
↪ $p(\theta)$

and the use Bayes rule to estimate the optimal paramete

via the principle of maximum a-posteriori estimation.

$$\underbrace{p(\theta \mid \mathcal{D})}_{\substack{\text{posterior} \\ \text{pdf over the} \\ \text{model params}}} = \frac{\overbrace{p(\mathcal{D} \mid \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{D})}_{\substack{\text{model} \\ \text{evidence}}}} = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{\underbrace{\int p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta}_{\substack{\text{marginal} \\ \text{likelihood}}}}$$

- Recall linear regression:

Setup: Given $\mathcal{D} := \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Model: $y_i = w^T x_i + \varepsilon$, if we assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$w^T \phi(x_i)$ in case we use basis fun

$\overset{\text{likehood}}{\Longrightarrow}$ $y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(y_i \mid w^T x_i, \alpha^{-1})$, where $\alpha = \frac{1}{\sigma^2}$ is the precision.

- Unkown params: $\theta := \{w_1, \ldots, w_d, \cancel{\alpha^{-1}}\}$ (for now let us assume that the noise precisi is known)

In a Bayesian formulation we assume

a prior: $w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, b^{-1} I_d)$, $\underset{d \times d}{I_d} := \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

i.e. a multi-variate Gaussian prior on $w \in \mathbb{R}^d$.

Likelihood: $p(\mathcal{D}|w) \propto \exp\left[-\frac{\alpha}{2}(y - Xw)^T(y - Xw)\right]$ (see l... lectu...

where $\underset{n\times 1}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\underset{\substack{\text{design}\\\text{matrix}}}{\overset{n\times d}{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & & \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$, $\underset{d\times 1}{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$

Posterior:

$$\underbrace{p(w|\mathcal{D})}_{\text{posterior}} \overset{\text{Bayes}}{\propto} \underbrace{p(\mathcal{D}|w)}_{\text{likelihood}} \underbrace{p(w)}_{\text{prior}}$$

(i.e. we ommited the denominator in Bayes rule

$$\Longrightarrow \quad p(w|\mathcal{D}) \propto \exp\left[-\frac{a}{2}(y - Xw)^T(y - Xw) - \frac{b}{2} w^T w\right] \quad \text{(verify}$$

1

Notice that the exponent is quadratic in w. This hints that the posterior $p(w|\mathcal{D})$ is Gaussian. To see this, let us derive the result by "caupleting the square".

First, let us re-writte:

$$\alpha(y - Xw)^T(y - Xw) + b w^T w \overset{\text{(verify!)}}{=} a y^T y - 2\boxed{a w^T X^T y} + w^T(\alpha X^T X + bI$$

Recall the form of the exponent of a multi-variate Gaussian

$X \sim \mathcal{N}(\mu, \Lambda^{-1})$, then $(X - \mu)^T \Lambda (X - \mu) =$

$$= X^T \Lambda X - 2 X^T \Lambda \mu + \underbrace{\mu^T \Lambda \mu}_{\text{constant}} \quad \text{(i.e. does not involve } X\text{)}$$

To "match" the terms, let:

$$\boxed{\Lambda := \alpha X^T X + bI \quad \text{(precision)}}$$

we also want: $\alpha w^T X^T y = w^T \Lambda \mu \Longrightarrow \boxed{\mu := \alpha \Lambda^{-1} X^T y \quad \text{(mea}}$

Based on these newly defined variables, we can re-write

$$\textcircled{1} \Longrightarrow \quad p(w|\mathcal{D}) \propto \exp\left[(w-\mu)^{T} \Lambda (w-\mu) \right] \quad \textcircled{2}$$

$$\boxed{p(w|\mathcal{D}) \propto \mathcal{N}(w|\underline{\underline{\mu}}, \Lambda^{-1}), \quad \begin{cases} \Lambda = aX^TX + bI \\ \mu = \alpha \Lambda^{-1} X^T y \end{cases}}$$

This derivation was possible because:

(i) We assumed a linear model (i.e. model depends linearly on

(ii) We assumed a Gaussian likelihood (i.e. we assumed a Gaussian model for the observation noise)

(iii) We assumed a Gaussian prior over w.

- Maximum a-posteriori estimation for w (MAP):

$$w_{MAP} = \arg\max_{w} \underbrace{p(w|\mathcal{D})}_{\text{posterior}}$$

Recall, :

$$w_{MLE} = \arg\max_{w} \underbrace{p(\mathcal{D}|w)}_{\text{likelihood}}$$

$$\Rightarrow \quad w_{MAP} = \mu = \alpha(\alpha X^TX + bI)^{-1} X^T y$$

$$\Rightarrow \quad \boxed{w_{MAP} = \left(X^TX + \frac{b}{\alpha} I\right)^{-1} X^T y} \longrightarrow$$

The Bayesian approach naturally introduces regularization.

Compare this to

$$\boxed{w_{MLE} = (X^TX)^{-1} X^T y}$$

(*) Equivalently one can see the distinction between MLE vs MA by noticing the following:

$$w_{MLE} = \arg\min_{w} \|y - Xw\|_2^2$$

$$w_{MAP} = \arg\min_{w} \|y - Xw\|_2^2 + \underbrace{\lambda \|w\|_2^2}_{\text{regularization}}, \quad \lambda = \frac{b}{\alpha}$$

At the end, all we really care about is making predictions (ideally with quantified uncertainty), i.e.

$$y^* \quad f(x^*) \quad p(y^* | x^*, \mathcal{D})$$

$\underbrace{y = f(x^*)}_{\text{point prediction}}$,

statistical / probabilistic prediction.

Specifically to the Bayesian linear regression model defined abov

$$p(y^*|x^*, \mathcal{D}) = \int \underbrace{p(y^*|x^*, \mathcal{D}, w)}_{\substack{\text{likelihood} \\ \text{(Gaussian)}}} \underbrace{p(w|\mathcal{D})}_{\substack{\text{posterior} \\ \text{(Gaussian)}}} dw$$

$$\propto \int \exp\left[-\frac{\alpha}{2}(y^* - X^* w)^T (y^* - X^* w)\right] \exp\left[-\frac{1}{2}(w-\mu)^T \Lambda (w-\mu)\right]$$

$$\cdots \implies \boxed{p(y^*|x^*, \mathcal{D}) = \mathcal{N}(y^*|u, 1/\lambda)}, \quad \text{where}$$

predictive posterior distribution

$$\begin{cases} u = \mu^T x^* \\ \dfrac{1}{\lambda} = \dfrac{1}{\alpha} + x^{*T} \Lambda^{-1} x^* \end{cases}$$

where
$$\begin{cases} \mu = w_{MAP} = \left(X^T X + \frac{b}{\alpha} I\right)^{-1} X^T y \\ \Lambda = \alpha X^T X + bI \end{cases}$$