Property

In the Bayesian framework, and considering the prior distribution $p(\beta)$ introduced before, the maximum a posteriori estimate of β is given by:

$$\hat{\beta}_{\text{MAP}} = (X^\intercal X + \overset{\lambda}{\overset{=}{\alpha\sigma^2}} I_p)^{-1} X^\intercal Y$$

- provided that $\lambda = \alpha\sigma^2$ is large enough, $(X^\intercal X + \lambda I_p)$ is full rank and so $\hat{\beta}_{\text{MAP}}$ can be computed
- simple solution for the high dimensional setting

Bayesian framework: step 2

Property

In the Bayesian framework, and considering the prior distribution $p(\beta)$ introduced before, the posterior distribution of the regression vector given the data has an analytical form:

$$p(\beta|X, Y, \sigma^2) = \mathcal{N}(\beta; m_n, S_n),$$

with

$$S_n = (\frac{X^\intercal X}{\sigma^2} + \alpha I_p)^{-1},$$

and

$$m_n = (X^\intercal X + \alpha\sigma^2 I_p)^{-1} X^\intercal Y$$

Remark

Since $p(\beta|X, Y, \sigma^2)$ is Gaussian, its mode is its expectation:

$$\hat{\beta}_{\mathrm{MAP}} = m_n$$

Outline Part 2

Bayesian linear regression

EM revisited

Gaussian processes

we now want to see α as an (hyper)parameter to be estimated from the training data set (link with ridge regression). So, $p(\beta)$ is replaced by:

$$p(\beta|\alpha) = \mathcal{N}(\beta; 0_p, \frac{I_p}{\alpha}),$$

with $\alpha > 0$ to be estimated.

Bayesian framework: step 3: EM revisited

Seing β as a latent (unknown) random vector, an EM algorithm can be derived to estimate the pair (α, σ^2) on the *full* training data set:

- init: initialise the values of (α, σ^2)

E compute
  - $S_n = (\frac{X^\intercal X}{\hat{\sigma^2}} + \hat{\alpha} I_p)^{-1}$
  - $m_n = (X^\intercal X + \hat{\alpha}\hat{\sigma^2} I_p)^{-1} X^\intercal Y$

M compute
  - $\hat{\alpha} = p/(\mathrm{Tr}(S_n) + m_n m_n^\intercal)$
  - $\hat{\sigma^2} = (1/n)\left\{||Y - Xm_n||^2 + \mathrm{Tr}(X^\intercal X S_n)\right\}$

- if the log-likelihood has changed (or the parameters) (no eps convergence) back to E.

The evidence procedure

This algorithm is referred to as the evidence procedure (Mac92)

- the full training set is used to estimate α and σ^2 !
- no splits of the training data set are used as in cross validation !

Outline Part 2

Bayesian linear regression

EM revisited

Gaussian processes GP

$$\boxed{Y = X\beta + \varepsilon} \quad \text{with } \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$$

$$\beta|\alpha \sim \mathcal{N}(0_p, \frac{I_p}{\alpha})$$

$$\varepsilon \perp\!\!\!\perp \beta$$

As of now, we have:

- $Y|X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$
- $\beta|\alpha \sim \mathcal{N}(0_p, \frac{I_p}{\alpha})$

Reminder

The regression vector β is seen as a latent (unknown) random vector. The hyperparameters are α and σ^2

$$\Rightarrow Y | X, \alpha, \sigma^2 \sim \mathcal{N}\left(0_n, \frac{XX^T}{\alpha} + \sigma^2 I_n\right)$$

$$E[Y|X,\alpha,\sigma^2] = XE[\beta] + E[\varepsilon] = 0_n$$

$$V(Y|X,\alpha,\sigma^2) = E[X\beta\beta^T X^T] + E[\varepsilon] = X\underbrace{E[\beta\beta^T]}X^T + \sigma^2 I_n$$

$$\log p(Y \mid X, \sigma^2, \alpha) = \log\left\{\int p(Y, \beta \mid X, \sigma^2, \alpha)\, d\beta\right\} \longrightarrow \text{EM}$$

$$= \log \mathcal{N}\left(Y; O_n, \frac{XX^\intercal}{\alpha} + \sigma^2 I_n\right)$$

Property

From the Gaussian property, we have:

$$Y|X, \sigma^2, \alpha \sim \mathcal{N}(O_n, \underbrace{\frac{XX^\intercal}{\alpha} + \sigma^2 I_n}_{C_n})$$

Remark

- the associated likelihood $\mathcal{N}(Y; O_n, \frac{XX^\intercal}{\alpha} + \sigma^2 I_n)$ is sometimes referred to as the *type 2* maximum likelihood
- it can be optimised directly using optimisation algorithms
- warning: complexity: $O(n^3)$!

if $i = j$ 1, ,, 0 otherwise

$$(C_n)_{ij} = \frac{x_i^\intercal x_j}{\alpha} + \sigma^2 \delta(i,j)$$

Def

More generally, Gaussian processes can be built directly as:

$$Y|X, \sigma^2, \theta \sim \mathcal{N}(0_n, C_n),$$

where $C_n = K_n + \sigma^2 I_n$ and

$$(K_n)_{ij} = k(x_i, x_j)$$

The function $k(\cdot, \cdot)$ is a kernel function.

Example of a kernel function for Gaussian processes

Def

The exponential quadratic kernel is given by:

$$k(x_i, x_j) = \theta_0 \exp\left\{-\frac{\theta_1}{2}||x_i - x_j||^2\right\} + \theta_2 + \theta_3 x_i^\intercal x_j,$$

with

$$\beta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \in \mathbb{R}^4$$

In Gaussian processes (GP), the optimisation problem is given by:

$$(\hat{\theta}, \hat{\sigma^2}) = \text{argmax}_{\theta, \sigma^2} \log \mathcal{N}(Y; 0_n, C_n)$$

θ, σ^2

Remarks

- again, the complexity is $O(n^3)$
- the parameters θ and σ^2 "only" play a role in the covariance matrix of the model

Training $\longrightarrow \hat{\theta}, \hat{\sigma}^2$

- (X, Y) is the training data set with n elements
- let us consider a new observation x_{n+1} for which we aim at predicting y_{n+1}
- we build

$$X_{n+1} = \begin{pmatrix} X \\ x_{n+1}^{\intercal} \end{pmatrix},$$

and

$$Y_{n+1} = \begin{pmatrix} Y \\ y_{n+1} \end{pmatrix},$$

prediction

- the model becomes:

$$Y_{n+1}|X_{n+1}, \theta, \sigma^2 \sim \mathcal{N}(0_{n+1}, C_{n+1})$$

with

$$C_{n+1} = \begin{pmatrix} C_n & k \\ k^\intercal & c \end{pmatrix},$$

and $k_i = k(x_i, x_{n+1}) = k(x_{n+1}, x_i), \forall i \in \{1, \ldots, n\}$, and $c = k(x_{n+1}, x_{n+1}) + \sigma^2$

Property

From Gaussian property, it follows that:

$$y_{n+1}|X_{n+1}, Y, \theta, \sigma^2 \sim \mathcal{N}(\tilde{m}, \tilde{\sigma^2})$$

References I

- A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood for incomplete data via the em algorithm*, Journal of the Royal Statistical Society **B39** (1977), 1–38.
- D. MacKay, *A practical bayesian framework for backpropagation networks*, Neural Computation **4** (1992), 448–472.