Ref for Bayesian estimation: Bishop 2.3.6

From Last Time (with corrections): Bayesian Regression relies on integration

$$p(w|y) = N(w|m_n, S_n)$$

if $p(w) = N(w, 0, \alpha^{-1} I)$

then

$$m_n = \beta S_n \Phi^T y$$

$$S_n^{-1} = \alpha I + \beta \Phi^T \Phi$$

substituting

$$m_n = \beta(\alpha I + \beta \Phi^T \Phi)^{-1} \Phi^T y$$

$$= \left(\frac{1}{\beta}\right)^{-1} (\alpha I + \beta \Phi^T \Phi)^{-1} \Phi^T y$$

$$= \left(\frac{\alpha}{\beta} I + \Phi^T \Phi\right)^{-1} \Phi^T y$$

if not scalar $(AB)^{-1} = B^{-1} A^{-1}$

We can also directly maximize the posterior:

Maximum a posteriori (MAP) estimate: $\longrightarrow$ based on optimizing something

$$\max_w \; p(w|y)$$

$$\min_w \; -\log p(w|y)$$

$$= \min_w \frac{\beta}{2} \sum (y_i - w^T \phi(x_i))^2 + \frac{\alpha}{2} w^T w$$

Covariance Matrix Review

what is covariance?

Def.

$$\mathrm{cov}(x,y) = \mathbb{E}[(x - \mathbb{E}x)(y - \mathbb{E}y)]$$

$$= \mathbb{E}[xy] - \mathbb{E}x\,\mathbb{E}y$$

ex: If we have a population

$$w_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad i = 1 \dots n$$

$$\mathrm{cov}(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

if $x_i > \bar{x}$ and $y_i > \bar{y}$ then $\mathrm{cov}(x,y) > 0$

$x_i < \bar{x}$ and $y_i < \bar{y}$ " " > 0

$x_i > \bar{x}$ and $y_i < \bar{y}$ " " < 0

With some data with $\bar{x} = 10$, $\bar{y} = 20$

$$w_1 = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \quad w_2 = \begin{bmatrix} 11 \\ 21 \end{bmatrix}, \quad w_3 = \begin{bmatrix} 9 \\ 19 \end{bmatrix}$$

$\downarrow$

$$cov(x,y) = \frac{1}{3}\Big((0)(0) + (1)(1) + (-1)(-1)\Big) = \frac{2}{3}$$

Now

$$w_1 = \begin{bmatrix}10\\20\end{bmatrix}, \quad w_2 = \begin{bmatrix}11\\19\end{bmatrix}, \quad w_3 = \begin{bmatrix}9\\21\end{bmatrix} \quad cov(x,y) = \cdots = -\frac{2}{3}$$

We can also calculate the <u>covariance matrix</u>

$$\frac{1}{n}\sum(w_i - \bar{w})(w_i - \bar{w})^T = 2 \times 2 \text{ matrix}$$

first case

$$\frac{1}{3}\left[\begin{bmatrix}0\\0\end{bmatrix}\begin{bmatrix}0 & 0\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} + \begin{bmatrix}-1\\-1\end{bmatrix}\begin{bmatrix}-1 & -1\end{bmatrix}\right] = \frac{1}{3}\begin{bmatrix}2 & 2\\2 & 2\end{bmatrix}$$

Symmetric

Second case

$$\begin{bmatrix}1\\-1\end{bmatrix}\overset{\begin{bmatrix}1 & -1\end{bmatrix}}{\begin{bmatrix}1 & -1\\-1 & 1\end{bmatrix}} = \frac{1}{3}\begin{bmatrix}2 & -2\\-2 & 2\end{bmatrix}$$

where non-diagonal entries are covariances

Independent $\rightarrow$ uncorrelated $\quad cov(x,y) = 0$

uncorrelated $\not\rightarrow$ independent

... but if $w \sim N(\mu, \Sigma)$, Σ diagonal (all off diag entries are zero)

then uncorrelated $\rightarrow$ independent

reason:

Since $N(w|\mu,\Sigma) = \prod_{i=1}^{m} N(w_i | \mu_i, \sigma_i^2)$ $\quad$ Recall $p(x,y) = p(x)\,p(x)$

$$\mu = \begin{bmatrix}\mu_1\\ \vdots \\ \mu_m\end{bmatrix}, \quad \Sigma = \begin{bmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2\end{bmatrix}$$

uncorrelated: $\mathbb{E}(x,y) - \mathbb{E}x\,\mathbb{E}y$

Back to Bayesian Regression

We no longer have a point prediction for $\vec{w}$

We instead have a distribution => this will allow us to put levels of confidence on our estimates

<u>Predictive distribution</u>

defn.

$$p(\hat{y}|y) = \int \underbrace{p(\hat{y}|w)}_{N(\hat{y}\,|\,w^T\phi(x),\,\beta^{-1})}\ \underbrace{p(w|y)}_{\text{posterior}}\,dw \qquad N(w|m_n, S_n)$$

remember this is a result of the noise

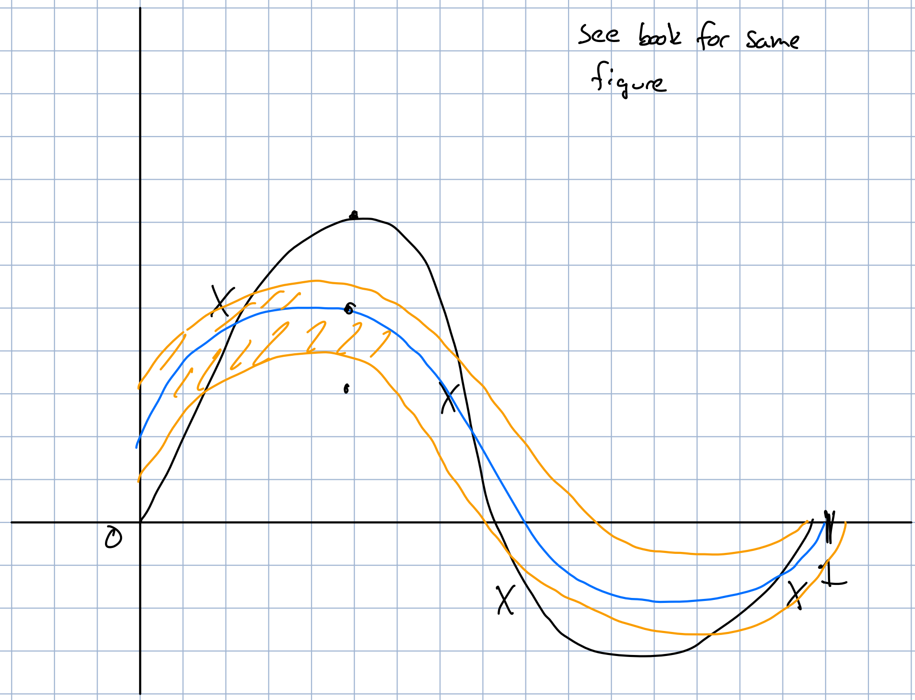
$$p(\hat{y}|y) = N(\underbrace{\hat{y}\,|\,m_n^T\phi(x),\ \sigma_n^2(x)}_{\text{these are all scalars}})$$

where $\sigma_n^2 = \frac{1}{\beta} + \phi(x)^T S_n \phi(x)$

noise

arises from uncertainty in w

See book for same figure



in the demo we see that we have larger uncertainty in areas where we don't have data

For different values of x we have predictions with different variances $\sigma_n^2(x)$

Instead of a single curve that fits the data we have a family of curves

The mean of the family is $m_n^T \phi(x)$

another example of the family is $w^T \phi(x)$

$w \sim N(m_n, S_n)$

Unpacking the predictive distribution some more

Take

$$\hat{y} = w_0 + w_1 x$$

$$= \int \left(w_0 \, p(w_0 | y) + w_1 \, p(w_1 | y) \, x \right) dw_0 \, dw_1$$

$$= \int w^T \phi(x) \, p(w | y) \, dw$$

=> the integral is prediction averaged over all w

like formula for expectation