CS419M

Lecture 18: Mean and Variance

02/04/23

Lecturer: Abir De *Scribe: Groups 11 & 12*

This is some warmup discussion before the first section.

1 Recap

Consider the following equation:

$$y = w^\top \phi(x) + \epsilon$$

This is our standard regression model, with $\phi(x)$ being the $d \times 1$ feature vector.

$\epsilon \sim \mathcal{N}(0, \sigma^2)$
is the noise in the model, modelled as a Gaussian with 0 mean and variance σ^2.
$w \sim \mathcal{N}(0, \sum_p)$
is the $d \times 1$ weight vector, drawn from a Gaussian distribution.

A Gaussian process is a collection of random variables which have a joint Gaussian distribution.

Given N observations $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$, for a new observation $(x*, y*)$, we have:

$$y * /x*, D \sim \mathcal{N}(\mu, \Sigma)$$

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$$

We want to find out μ and Σ in the above equation.
We have seen that

$$\mathbf{E}(y * /x*, D) = \Phi(x*)^\top \Sigma_p \Phi(\Phi^\top \Sigma_p \Phi + \sigma^2 \boldsymbol{I})^{-1} y \tag{1}$$

Here $\Phi^\top \Sigma_p \Phi$ may be invertible only if $d \to \infty$.

2 Analysing the mean further

Suppose $x* \in D$. Without loss of generality, let $x* = x_1$. If $\epsilon = 0$ (which implies $\sigma = 0$), we expect $y*$ to be exactly equal to y_1. If the noise was present even for an $x_i \in D$, the measured y can be different from y_i. Let us try to verify this.

Putting $x* = x_1$ and $\sigma = 0$ in (1), we have:

$$\mathbf{E}(y_1/x_1, D, \sigma = 0) = \Phi(x_1)^\top \Sigma_p \Phi(\Phi^\top \Sigma_p \Phi)^{-1} y \tag{2}$$

Now $\Phi(x_1)^\top \Sigma_p \Phi$ is the first row of $\Phi^\top \Sigma_p \Phi$.

If B is an invertible matrix and $B_{1,.}$ is its first row, then

$$(AB)_{1,.} = A_{1,.}B$$

We can write:

$$BB^{-1} = \boldsymbol{I} \implies B_{1,.}B = \boldsymbol{I}_{1,.} = [1, 0, ..., 0]_{1\times n}$$

So if we take the matrix $\Phi^\top \Sigma_p \Phi$ as B above we obtain the same row vector as above.*(Note that we have assumed $\Phi^\top \Sigma_p \Phi$ to be invertible, which may not always be the case)*. Finally, multiplying with y which is a $n \times 1$ column vector, we obtain y_1 on the RHS of (2).

Now lets investigate what happens if $\sigma \neq 0$.

Again, taking $B = \Phi^\top \Sigma_p \Phi$ and B_1 as its first row, we have the RHS of (2) as

$$B_1(B + \sigma^2 I)^{-1}y = B_1(B + \sigma^2 BB^{-1})y = B_1(\boldsymbol{I} + \sigma^2 B^{-1})^{-1}B^{-1}y = y_1 - \sigma^2 B_1 1 B^{-2}y$$

Here $(\boldsymbol{I} + \sigma^2 B^{-1})^{-1}$ was expanded as $\boldsymbol{I} - \sigma^2 B^{-1}$ using Taylor's theorem, under the assumption that σ is small enough for the expension to be valid.

3 Variance

$$y = w^T.\phi(x) + \epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$w \sim \mathcal{N}(0, \epsilon_P$$

What would the value of $var(y|D)$ be? Where D = $(x_i, y_i)_{i=1}^N$ P($y^*|x^*$, D)

$$var(y|D) = var(w^T\phi(x)) + \sigma^2$$
$$= \mathbb{E}(\phi(x^*)^T(w - \overline{w})(w - \overline{w})^T\phi(x^*)|D) + \sigma^2$$

Here w is kind of stochastic.

$$= \phi(x^*)^T\mathbb{E}((w - \overline{w})(w - \overline{w})^T|D)\phi(x^*) + \sigma^2$$

For now, we rather focus on P(w|D). The problem of finding the variance of y^* reduces to finding the covariance matrix of w.

If we know w, we can easily find the distribution of D.

$$P(w|D) = \frac{P(D|w).P(w)}{P(D)}$$

$$\implies P(w|D) \propto P(D|w).P(w)$$

Now,

$$P(D|w).P(w) = exp[-\frac{(\overrightarrow{y} - \phi^T w)^T(\overrightarrow{y} - \phi^T w)}{2\sigma^2}].exp[-w^T\epsilon_p^{-1}w/2]$$

$$z^{-1} = \phi\phi^T/\sigma^2 + \epsilon_P^{-1}$$

Confirming $\overline{w}$ is the same that we found earlier.

$$\overline{w} = \frac{z\sum_{i=1}^{N}\phi(x_i)(y_i)}{\sigma^2} = \frac{Z\phi.y}{\sigma^2}$$

$$\mathcal{E}(x^*, y^*|D) = \phi(x^*)^T[\phi\phi^T/\sigma^2 + \epsilon_P^{-1}]\frac{\phi.y}{\sigma^2}$$