CS419-Intro to ML

Lecture 19:

31/03/2023

Lecturer: Abir De *Scribe: Group 21*

1 Recap

Consider the following process :

$$y = w^T\phi(x) + \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p)$ is a d X 1 weight vector, and $\phi(x)$ is the d x 1 feature vector, then the probability of observing y^* given x^* and D, where D = $\{(x_i, y_i)|i = 1, 2, ..., n\}$ is given by:

$$P(y^*|x^*, D) \sim \mathcal{N}(\mu_{y^*}, \sigma^2_{y^*})$$

$$\mu_{y^*} = \phi(x^*)^T \left[\frac{\Phi\Phi^T}{\sigma^2} + \Sigma_p^{-1}\right]^{-1} \frac{\Phi y}{\sigma^2} \tag{2}$$

$$\mu_{y^*} = \phi(x^*)^T \Sigma_p \Phi(\Phi\Sigma_p\Phi + \sigma^2 I)^{-1} y \tag{3}$$

Note that these two expressions of μ_{y^*} will be equal only when both the inverses exist

$$\sigma^2_{y^*} = \phi(x^*)^T \left[\Sigma_p - \Sigma_p\Phi(\Phi^T\Sigma_p\Phi + \sigma^2 I)^{-1}\Phi^T\Sigma_p\right]\phi(x^*) + \sigma^2 \tag{4}$$

Recall ,

$$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(y_B - \mu_B)$$

Hence here $\mu_A = \mu_B = 0$, $\Sigma_{AB} = \phi(x^*)^T\Sigma_p\Phi$, $\Sigma_{BB} = \Phi^T\Sigma_p\Phi$

For σ = 0 we have shown that,

$$\mu_{y^*} = y_i \quad if \;\; (x^*, y^*) = (x_i, y_i) \tag{5}$$

2 Analyzing the Variance

Similar to calculation of μ_{y^*}, for $\sigma = 0 \; and \; (x^*, y^*) = (x_i, y_i)$,

$$\sigma^2_{y^*} = \phi(x^*)^T\Sigma_p\phi(x^*) - \phi(x^*)^T\Sigma_p\Phi(\Phi^T\Sigma_p\Phi + \sigma^2 I)^{-1}\Phi^T\Sigma_p\phi(x^*)$$

As shown in previous lectures , $\phi(x^*)^T\Sigma_p\Phi(\Phi^T\Sigma_p\Phi + \sigma^2 I)^{-1}$ = [1 0 0..0] , when $(x^*, y^*) \in D$

$$\implies \sigma^2_{y^*} = \phi(x^*)^T\Sigma_p\phi(x^*) - [1\;0\;0..0]\Phi^T\Sigma_p\phi(x^*)$$

$$\implies \sigma_{y^*}^2 = \phi(x^*)^T \Sigma_p \phi(x^*) - \phi(x^*)^T \Sigma_p \phi(x^*)$$

$$\implies \sigma_{y^*}^2 = 0$$

Similarly we can also show that if (x^*, y^*) are such that,

$$\phi(x^*) = \Sigma \alpha_i \phi(x_i) \quad , \quad y^* = \Sigma \alpha_i y_i \quad , \quad then \ \ \sigma_{y^*} = 0 \tag{6}$$

3 Finding Variance of Gaussian process when the data belongs to the training data itself

We have to find $\sigma_{y^*}^2 \ of \ P(y^*|x^*, D)$ for $x^* = x_i$
We know that,

$$\sigma_{y^*}^2 = \phi(x^*)^T \left[\Sigma_p - \Sigma_p \Phi (\Phi^T \Sigma_p \Phi + \sigma^2 I) \Phi^T \Sigma_p \right] \phi(x^*) + \sigma_2$$

$$\Sigma_p - \Sigma_p \Phi(\Phi^T \Sigma_p \Phi + \sigma^2 I) = \Sigma_p - \Sigma_p \phi (I + \sigma^2 (\Phi^T \Sigma_p \Phi)^{-1})^{-1} (\Phi^T \Sigma_p \Phi)^{-1} \Phi^T \Sigma_p$$

$$= \Sigma_p - \Sigma_p \phi (I - \sigma^2 (\Phi^T \Sigma_p \Phi)^{-1}) (\Phi^T \Sigma_p \Phi)^{-1} \Phi^T \Sigma_p$$

$$= \Sigma_p - \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-1} \Phi^T \Sigma_p + \sigma^2 \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-2} \Phi^T \Sigma_p$$

Hence,

$$\sigma_{y^*}^2 = \Phi(x^*)^T \Sigma_p \Phi(x^*) - \Phi(x^*)^T \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-1} \Phi^T \Sigma_p \Phi(x^*) + \sigma^2 \Phi(x^*)^T \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-2} \Phi^T \Sigma_p \Phi(x^*) + \sigma^2$$

$$\Phi(x^*)^T \Sigma_p \Phi(x^*) - \Phi(x^*)^T \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-1} \Phi^T \Sigma_p \Phi(x^*) = 0 \qquad for \ x^* = x_i$$

Hence,
$\sigma_{y^*}^2 = \sigma^2 \Phi(x^*)^T \Sigma_p \Phi (\Phi^T \Sigma_p \Phi)^{-2} \Phi^T \Sigma_p \Phi(x^*) + \sigma^2$
= $2\sigma^2$

4 Writing the above variance in terms of kernel function

Let

$$\phi(x^*)^T \Sigma_p \Phi = K(x^*, x)$$
$$\Phi^T \Sigma_p \Phi = K(X, X) = K$$

Hence,

$$\sigma_{y^*}^2 = K(x^*, x + \sigma^2 - K(x^*, x)[K + \sigma^2 I]^{-1} K(x, x^*) y$$

5 Some points to think about

If $\sigma \neq 0$ then for what instance the variance σ_{y^*} will be least?

Lets say we have $x_1, x_2, ..., x_{1million}$ unlabelled points. $x'_1, x'_2, ..., x'_{1000}$ are labelled as $y'_1, y'_2, ..., y'_{1000}$. Using these we want to pick some $x_{i's}$ for labelling , but which ones to pick, because picking all of them is not practically possible?

We will want those $x_{i's}$ which are dissimilar with the given $x_{j's}$. Hence they will have more variance. Therefore we find low variance $x_{i's}$ and just label them using nearest neighbour and discard them from getting picked up to label the hard way.