CS419M

Lecture 17: Interpolation and Regression revisited

24th March, 2023

Lecturer: Abir De *Scribe: Harshit, Samyak, Satush*

1 Introduction

This is a brief recap of the results of the previous lecture on interpolation.

Conditional Gaussian distribution: Consider the joint distribution between vectors x_1 and x_2:

$$\begin{bmatrix} y_A \\ y_B \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AB} \\ \mathbf{\Sigma}_{BA} & \mathbf{\Sigma}_{BB} \end{bmatrix}\right)$$

We are interested in the conditional distribution, which itself is Gaussian:

$$y_A | y_B \sim \mathcal{N}\left(\boldsymbol{\mu}_{A|B}, \mathbf{\Sigma}_{A|B}\right)$$

where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}(y_B - \boldsymbol{\mu}_B)$$
$$\mathbf{\Sigma}_{A|B} = \mathbf{\Sigma}_{AA} - \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{\Sigma}_{BA}$$

Product of Gaussian distributions: Consider the two distributions:

$$p_1(x) = \mathcal{N}(x; \mu_1, \mathbf{\Sigma}_1), \quad p_2(x) = \mathcal{N}(x; \mu_2, \mathbf{\Sigma}_2)$$

The product is an un-normalised Gaussian:

$$p_1(x)p_2(x) \propto \mathcal{N}(x; \mu, \mathbf{\Sigma})$$

where

$$\boldsymbol{\mu} = \mathbf{\Sigma}(\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2)$$
$$\mathbf{\Sigma} = (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}$$

2 Linear Regression

Consider the supervised training data of n samples, each with an observation $\mathbf{x_i}$ and output y_i. The regression function $f(\mathbf{x})$ is linear if defined as

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

and the target value has Gaussian noise so that

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon$$

where

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

For a given value of w, the likelihood of the outputs can be expressed as

$$p(y_1, \ldots, y_n | x_1, \ldots, x_n, \mathbf{w}) = \prod_{i=1}^{n} p(y_i | x_i, w) = \mathcal{N}(\mathbf{y}; \mathbf{\Phi}^T \mathbf{w}, \sigma^2 I)$$

where I is the $n \times n$ identity matrix and

$$\mathbf{\Phi} = \begin{bmatrix} \phi(\mathbf{x_1}) & \cdots & \phi(\mathbf{x_n}) \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

The value of optimum $\mathbf{w}^*$ can then be found by maximizing this likelihood. This is equivalent to minimizing the least squares cost function

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \phi(\mathbf{x_i}))^2 = \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{\Phi}^T \mathbf{w}\|^2$$

which gives us the following result (derived in a previous lecture):

$$\mathbf{w}^* = (\mathbf{\Phi} \mathbf{\Phi}^T)^{-1} \mathbf{\Phi} \mathbf{y}$$

3 Weight Vector Prior

Now consider a prior distribution over w given by the Gaussian:

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{\Sigma}_p)$$

The result now becomes:

$$\mathbf{w}^* = \left(\frac{\mathbf{\Phi} \mathbf{\Phi}^T}{\sigma^2} + \mathbf{\Sigma}_p^{-1} \right)^{-1} \frac{\mathbf{\Phi} \mathbf{y}}{\sigma^2}$$

We can confirm this intuitively by the following 2 checks:

- For the σ^2 outside the bracket: if σ^2 is large, that means y has a lot of noise so we should discard the data point. Clearly the weights become very small for such a data point.

- For the σ^2 inside the bracket: if we are discarding the data points as in the above point, the weights shouldn't depend on x, thus it is also divided by σ^2.

4 Interpolation and Regression

From the previous setting, lets say we have observed the points $\mathcal{D} = (\mathbf{x_i}, y_i)_{i=1}^N$.

Now let us try to find out the distribution of $y^*|\{\mathbf{x}^*, (\mathbf{x_i}, y_i)_{i=1}^N\}$, or $y^*|\{\mathbf{x}^*, \mathcal{D}\}$, where $(\mathbf{x}^*, y^*)$ is a new unobserved point.

FACT: The distribution will be a Gaussian, hence we just need to find the mean and variance. Since the distribution is a Gaussian, the mean and the mode will be the same. Now we know that the mode is the following (by maximum likelihood estimate done in the previous sections):

$$\hat{y}^* = \mathbf{w}^{*T}\phi(\mathbf{x}^*) = \phi(\mathbf{x}^*)^T\mathbf{w}^*$$

where $\mathbf{w}^*$ is as shown in Section 3. Hence:

$$\hat{y}^* = \phi(\mathbf{x}^*)^T\left(\frac{\mathbf{\Phi\Phi}^T}{\sigma^2} + \mathbf{\Sigma}_p^{-1}\right)^{-1}\frac{\mathbf{\Phi y}}{\sigma^2}$$

Thus we have found the mean of the distribution:

$$y^*|\{\mathbf{x}^*, \mathcal{D}\} \sim \mathcal{N}(\hat{y}^*, \mathbf{\Sigma} =?)$$

Is this similar to the formula we derived in the last lecture? Lets see!

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}(y_B - \boldsymbol{\mu}_B)$$

Here, A is equivalent to $\mathbf{x}^*$ and B is equivalent to the y_i's in $\mathcal{D}$. Now, μ_A and μ_B are both equal to 0 as we do not have any observation and both y_A and y_B are sampled from distributions with mean 0. Thus,

$$\boldsymbol{\mu}_{A|B} = \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}y_B$$

$$\overline{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \mathbf{\Phi}^T\mathbf{w} + \epsilon$$

Hence,

$$\begin{aligned} \mathbb{E}[\mathbf{y}\mathbf{y}^T] &= \mathbf{\Phi}^T\mathbb{E}[\mathbf{w}\mathbf{w}^T]\mathbf{\Phi} + \mathbb{E}[\epsilon\epsilon^T] \\ &= \mathbf{\Phi}^T\mathbf{\Sigma}_p\mathbf{\Phi} + \sigma^2 I \end{aligned}.$$

Here, $\mathbb{E}[\mathbf{y}\mathbf{y}^T]$ denotes the covariance matrix and row x^* and column $1, 2, \ldots N$ of the covariance will be $\mathbf{\Sigma}_{AB}$. Therefore,

$$\begin{aligned} y_{A|B} &= \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}\mathbf{y}_B \\ &= [\phi(x^*)^T\mathbf{\Sigma}_p\mathbf{\Phi}][\mathbf{\Phi}^T\mathbf{\Sigma}_p\mathbf{\Phi} + \sigma^2 I]^{-1}\overline{\mathbf{y}} \end{aligned}$$

We define $\mathbf{K}$ as $\mathbf{\Phi}^T\mathbf{\Sigma}_p\mathbf{\Phi}$

It is important to note that we simply can't take the inverse of $\mathbf{\Phi}$ as $\mathbf{\Phi}$ is not a square matrix.

Theorem 4.1. *Let* $A = \frac{\mathbf{\Phi}\mathbf{\Phi}^T}{\sigma^2} + \mathbf{\Sigma}_p^{-1}$, *then*

$$A\mathbf{\Sigma}_p\mathbf{\Phi} = \frac{1}{\sigma^2}\mathbf{\Phi}\left(\boldsymbol{K} + \sigma^2 I\right)$$

Proof.

$$\begin{aligned}
\frac{1}{\sigma^2}\mathbf{\Phi}\left(\mathbf{K} + \sigma^2 I\right) &= \frac{1}{\sigma^2}\mathbf{\Phi}\left(\mathbf{\Phi}^T\mathbf{\Sigma}_p\mathbf{\Phi} + \sigma^2 I\right) \\
&= \frac{\mathbf{\Phi}\mathbf{\Phi}^T\mathbf{\Sigma}_p\mathbf{\Phi}}{\sigma^2} + \mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_p\mathbf{\Phi} \\
&= \left(\frac{\mathbf{\Phi}\mathbf{\Phi}^T}{\sigma^2} + \mathbf{\Sigma}_p^{-1}\right)\mathbf{\Sigma}_p\mathbf{\Phi}
\end{aligned}$$

□

Therefore, we have

$$\begin{aligned}
&\frac{1}{\sigma^2}\mathbf{\Phi}\left(\mathbf{K} + \sigma^2 I\right) = A\mathbf{\Sigma}_p\mathbf{\Phi} \\
\implies &\frac{1}{\sigma^2}A^{-1}\mathbf{\Phi} = \mathbf{\Sigma}_p\mathbf{\Phi}\left(\mathbf{K} + \sigma^2 I\right)^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{\sigma^2}\phi(x^*)^T A^{-1}\mathbf{\Phi}\overline{\mathbf{y}} &= \phi(x^*)^T\mathbf{\Sigma}_p\mathbf{\Phi}\left(\mathbf{K} + \sigma^2 I\right)^{-1}\overline{\mathbf{y}} \\
&= y_{A|B}
\end{aligned}$$

This proves that in linear regression also we are interpolating. But how is this possible? The answer lies in a small assumption we have made.

The matrix A (from regression formula) is always invertible. If $\sigma^2 = 0$ then the formula has $\mathbf{K}^{-1}$ (for the interpolation- $y_{A|B}$ case). Dimension of $\mathbf{\Phi}$ is $d * N$, where d is the dimension of the feature set ϕ. Thus, K is a reduced rank matrix for $d < N$ and isn't invertible. Hence the analogy of both being the same fails. So, for the analogy to work, d should be greater than **any** N, that means it should be infinite.

This means we have assumed that K is invertible, or if $\sigma^2 = 0$, ϕ is of infinite dimension!

We will prove that the Variance is also the same in both cases in the next lecture.