AIML - CS 337

Lecture 15: Gaussian processes and Kernel applications

7th Oct 2022

Lecturer: Abir De *Scribe: Anubhab, Govind, Hemendra, Shubham*

Kernel methods have various applications, such as capturing non linearity and being able to substitute ∞ dimensional features into an implementable form. In this lecture we look at another application of kernels in the context of Gaussian Processes and how to deal with smaller training sets to still give fair results.

1 Formulation of the problem

In typical machine learning applications, we aim to perform well on unseen data by learning on the training data. This is easier to accomplish when we have a large training data, however it is tough when we are faced with a smaller training set. An alternative approach to this could be to obtain a distribution on the function we are trying to predict such that every point in the training data must have exactly the same output in the hypothesis as the training label. Consider the standard linear regression model with the optimal function f defined as

$$\begin{aligned} w^{regression} &= \text{argmin} \quad \sum_{i \in D} (y_i - w^\top x_i)^2 \\ &= \left(\sum_{i \in D} x_i x_i^\top \right)^{-1} \sum_{i \in D} x_i y_i \\ f(x) &= (w^{regression})^\top x \end{aligned}$$

With a new data point x^*, $f(x^*)$ reduces to

$$f(x^*) = x^{*\top} \left(\sum_{i \in D} x_i x_i^\top \right)^{-1} \sum_{i \in D} x_i y_i$$

However, if $x^* \in D$, for e.g. if $x^* = x_1$, then $f(x_1) \neq y_1$. We would like to overcome this problem by designing a non linear estimator f to model the training data with the additional restriction that $\forall x_i \in D \; f(x_i) = y_i$. For the other points $x \notin D$, $f(x)$ is a random variable with an associated probability distribution. For this case, we look at the set of functions that follow a gaussian distribution. We can visualise such a function as shown in the figure below:

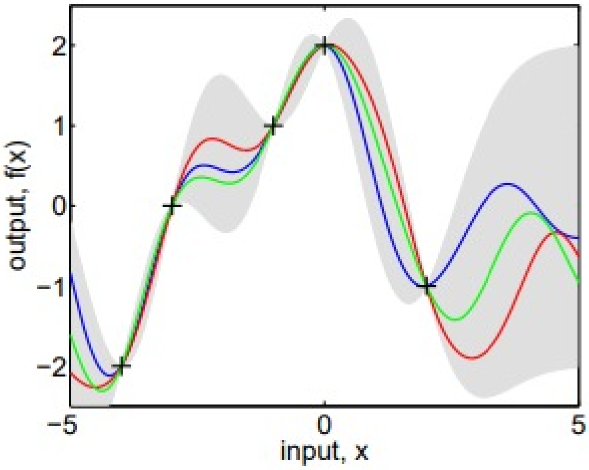

Figure 1: Graphical Representation

Here in this figure you can see that the points marked as + are the points in our dataset, for which the output is exactly one value while it is a distribution (as given by the shaded area) for all the other points

2 Gaussian Processes

Gaussian processes are a method for non parametric estimation to provide confidence on the seen data and some kind of distribution on unseen data. For any subset of the training data, we must have that the joint prior distribution of this subset is normally distributed for some mean and covariance matrix. For any subset $\{x_1...x_m\}$ of the training data, the prior distribution follows:

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix} \quad \sim \quad \mathcal{N}(\vec{\mu}(x_1, \dots, x_m), \Sigma(x_1, \dots, x_m))$$

where $\vec{\mu}$ and Σ are deterministic functions.

On introducing a new data point into any subset of the training data, we expect the resulting conditional distribution to also follow the normal distribution. For the data point x^*

$$f(x^*)|(f(x_1), \dots f(x_m), x^*) \sim \mathcal{N}(\vec{\mu}(x_1, \dots, x_m, x^*), \Sigma(x_1, \dots, x_m, x^*))$$

As described earlier, we expect that if a new data point introduced is already in the training data, then we expect that $f(x^*)$ takes the value that was present in the training set. This means that for any x^* such that $x^* \in \{x_1, \dots, x_m\}$

$$f(x^*)|(f(x_1), \dots f(x_m), x^*) \sim \mathcal{N}(f(x^*), 0)$$

Our aim is to design a matrix Σ that satisfies such a property i.e. posterior for any point in training data must have zero variance. For the sake of simplicity, we consider $\mu = 0$ in the prior for the rest of the section.

Notation Let $K(x, y)$ denote the kernel function we are using. Then the notation $K(x_1 \ldots x_m, y_1 \ldots y_n)$ denotes the the m x n matrix M with $M_{i,j} = K(x_i, y_j)$.

One such covariance matrix proposed in the class was the kernel matrix.

$$\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \\ f(x^*) \end{bmatrix} \sim \mathcal{N}(0, \begin{bmatrix} K(x_1, x_1) & \ldots & K(x_1, x_m) & K(x_1, x^*) \\ \vdots & \ddots & \vdots & \vdots \\ K(x_m, x_1) & \ldots & K(x_m, x_m) & K(x_m, x^*) \\ K(x^*, x_1) & \ldots & K(x^*, x_m) & K(x^*, x^*) \end{bmatrix})$$

Thus now we have our posterior distribution to be of the form:

$$f(x^*)|(f(x_1) \ldots f(x_m), x^*) \sim \mathcal{N}(\mu, \Sigma)$$

where

$$\mu = K(x^*, x_1 \ldots x_m) \times K^{-1}(x_1 \ldots x_m, x_1 \ldots x_m) \times \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}$$

and

$$\Sigma = K(x^*, x^*) - K(x^*, x_1 \ldots x_m) \times K^{-1}(x_1 \ldots x_m, x_1 ... x_m) \times K(x_1 \ldots x_m, x^*)$$

Here you can see that our μ and Σ are of such a form that upon substituting any $x_i \in \{x_1 \ldots x_m\}$ in place of x^* we get as output $\mu = f(x_i)$ and $\Sigma = 0$, which was indeed what we desired in our hypothesis while constructing such a distribution.

The complete proof for this can be found in chapter 2 of the reference [1].

3 Conclusion

Gaussian process regression is one application of kernel methods, in which we are able to obtain meaningful results when the data set is small. It provides us an assurance on seen data points, while providing a distribution of values over other unseen points. A detailed study of Gaussian processes can be found in the reference [1]. With the end of this lecture, the topic on Kernel Methods was brought to a close.

References

[1] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.