

Lecture 17: Regression

15-10-2022

Lecturer: Abir De

Scribe: Group 35 & 36

We have already discussed about Gaussian Processes in the previous lectures. In this lecture we continue on it and draw comparisons between Linear Regression which is parametric and Gaussian Process which is not parametric. At the end of the lecture we make use of Gaussian Processes to improve upon K-Means Clustering Algorithm. To begin with let us recall Linear Regression

1 Linear Regression

Consider the dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in R^d, y_i \in R$. On applying Linear Regression on this dataset we get loss as follows

$$L(\omega) = \sum_{i=1}^N (y_i - \omega^T x_i)^2 + \lambda \|\omega\|^2$$

In vectorized form we can write loss as

$$L(\omega) = \|\mathbf{y} - \mathbf{X}^T \omega\|^2 + \lambda \|\omega\|^2$$

where $\mathbf{y} \in R^N, \mathbf{X} \in R^{d \times N}, \omega \in R^d$. On minimizing the loss function we get

$$\begin{aligned} \nabla_{\omega} L(\omega) &= 0 \\ \nabla_{\omega} ((\mathbf{y} - \mathbf{X}^T \omega)^T (\mathbf{y} - \mathbf{X}^T \omega) + \lambda \omega^T \omega) &= 0 \\ \nabla_{\omega} (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}^T \omega - \omega^T \mathbf{X} \mathbf{y} + \omega^T \mathbf{X} \mathbf{X}^T \omega + \lambda \omega^T \omega) &= 0 \\ -2\mathbf{X} \mathbf{y} + 2\mathbf{X} \mathbf{X}^T \omega + 2\lambda \omega &= 0 \\ (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I}) \omega &= \mathbf{X} \mathbf{y} \end{aligned}$$

Proof for invertibility of $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$ when $\lambda > 0$

Consider

$$\begin{aligned} \mathbf{v}^T (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) \mathbf{v} &= \mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} + \lambda \mathbf{v}^T \mathbf{v} \\ &= \|\mathbf{X} \mathbf{v}\|^2 + \lambda \|\mathbf{v}\|^2 \\ &> 0 \text{ when } \|\mathbf{v}\| > 0 \end{aligned}$$

Thus $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$ is a positive definite matrix and hence invertible.

Similarly proof for invertibility of $\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I}$ can be done and we will use that in a later section.

When $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$ is invertible we have

$$\begin{aligned}\boldsymbol{\omega} &= (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X} \mathbf{y} \\ y_{pred}^{LR} &= \mathbf{x}_*^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X} \mathbf{y}\end{aligned}$$

2 Bayesian Linear Regression

We have already seen how to obtain the probability distribution $P(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y})$ in the previous lecture on Gaussian Process. Although this was not discussed in class, this alternate way would help us relate Gaussian Process with Linear Regression. Let $f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}$ and $\mathbf{y} = f(\mathbf{x}) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

$$\begin{aligned}P(\mathbf{y} | \mathbf{X}, \mathbf{w}) &= \prod_{i=1}^N P(y_i | \mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y_i - \mathbf{x}_i^T \mathbf{w})^2}{2\sigma_n^2}} \\ &= \frac{1}{(2\pi\sigma_n^2)^{\frac{n}{2}}} e^{-\frac{|\mathbf{y} - \mathbf{X}^T \mathbf{w}|^2}{2\sigma_n^2}} = \mathcal{N}(\mathbf{X}^T \mathbf{w}, \sigma_n^2 \mathbf{I})\end{aligned}$$

Consider a prior on \mathbf{w} as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \Sigma_p)$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \quad P(\mathbf{w} | \mathbf{y}, \mathbf{X}) = \frac{P(\mathbf{y} | \mathbf{w}, \mathbf{X}) P(\mathbf{w})}{P(\mathbf{y} | \mathbf{X})}$$

Note that $P(\mathbf{y} | \mathbf{X})$ is independent of \mathbf{w} . Thus we have

$$\begin{aligned}P(\mathbf{w} | \mathbf{X}, \mathbf{y}) &\propto e^{-\frac{(\mathbf{y} - \mathbf{X}^T \mathbf{w})^T (\mathbf{y} - \mathbf{X}^T \mathbf{w})}{2\sigma_n^2}} e^{-\frac{\mathbf{w}^T \Sigma_p^{-1} \mathbf{w}}{2}} \\ &\propto e^{-\frac{(\mathbf{w} - \bar{\mathbf{w}})^T (\frac{1}{\sigma_n^2} \mathbf{X} \mathbf{X}^T + \Sigma_p^{-1}) (\mathbf{w} - \bar{\mathbf{w}})}{2}}\end{aligned}$$

Let $A = \frac{1}{\sigma_n^2} \mathbf{X} \mathbf{X}^T + \Sigma_p^{-1}$ and $\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} (\frac{1}{\sigma_n^2} \mathbf{X} \mathbf{X}^T + \Sigma_p^{-1})^{-1} \mathbf{X} \mathbf{y}$

$$\begin{aligned}P(\mathbf{w} | \mathbf{X}, \mathbf{y}) &= \mathcal{N}\left(\frac{1}{\sigma_n^2} A^{-1} \mathbf{X} \mathbf{y}, A^{-1}\right) \\ P(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) &= \int P(f_* | \mathbf{x}_*, \mathbf{w}) P(\mathbf{w} | \mathbf{X}, \mathbf{y}) d\mathbf{w} \\ &= \mathcal{N}\left(\frac{1}{\sigma_n^2} \mathbf{x}_*^T A^{-1} \mathbf{X} \mathbf{y}, \mathbf{x}_*^T A^{-1} \mathbf{x}_*\right)\end{aligned}$$

Note that $\bar{\mathbf{w}} = \frac{1}{\sigma_n^2} (\frac{1}{\sigma_n^2} \mathbf{X}\mathbf{X}^T + \Sigma_p^{-1})^{-1} \mathbf{X}\mathbf{y} = (\mathbf{X}\mathbf{X}^T + \sigma_n^2 \Sigma_p^{-1})^{-1} \mathbf{X}\mathbf{y}$. Comparing it with the result of first section we obtain

$$\boxed{\lambda I = \sigma_n^2 \Sigma_p^{-1}}$$

2.1 Bayesian Linear Regression to Gaussian Regression

Going to higher dimensional space, we can just replace \mathbf{x} by $\phi(\mathbf{x})$. Define $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \Sigma_p \phi(\mathbf{x}')$. As Σ_p is positive definite we can find a symmetric matrix $\Sigma_p^{1/2}$ (using SVD) so that $(\Sigma_p^{1/2})^2 = \Sigma_p$. Define $\psi(\mathbf{x}) = \Sigma_p^{1/2} \phi(\mathbf{x})$. Thus $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^T \psi(\mathbf{x}')$. Also let $\mathbf{k}_* = [\mathbf{k}(\mathbf{x}_*, \mathbf{x}_1), \mathbf{k}(\mathbf{x}_*, \mathbf{x}_2), \dots, \mathbf{k}(\mathbf{x}_*, \mathbf{x}_N)]^T$. Substituting the above values in the obtained formula of $P(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y})$ along with some mathematical manipulations [1] we obtain

$$P(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{y}, \mathbf{k}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{k}_*)$$

This is the exact same equation which we had obtained earlier for Gaussian Processes.

3 Gaussian Processes

Recall that in Gaussian Processes we have the following

$$P(f_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{y}, \mathbf{k}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{k}_*)$$

$$y_{pred}^{GP} = \mathbf{k}_*^T (K + \sigma_n^2 I)^{-1} \mathbf{y}$$

where $\mathbf{k}_* = [\mathbf{k}(\mathbf{x}_*, \mathbf{x}_1), \mathbf{k}(\mathbf{x}_*, \mathbf{x}_2), \dots, \mathbf{k}(\mathbf{x}_*, \mathbf{x}_N)]^T$. Also observe that if $\sigma_n = 0$ for a $\mathbf{x}_* = \mathbf{x}_i \in D$ the value of $\mathbf{k}_*^T (K + \sigma_n^2 I)^{-1}$ is a row vector with all values zero except the i^{th} index which has value one (Think in terms of matrix multiplication of $K K^{-1}$ and focus on the i^{th} row of the output). Thus the mean of the distribution(prediction) for a point in training dataset is the true label and the variance at that point is zero. Note that y_{pred}^{GP} is a linear combination of observations \mathbf{y} . Another way to look at this equation is to see it as a linear combination of N kernel functions, each one centered on a training point, by writing

$$y_{pred}^{GP} = \sum_{i=1}^N \alpha_i \mathbf{k}(\mathbf{x}_i, \mathbf{x}_*)$$

This equation can be arrived at pretty simply.

$$\begin{aligned}\mathbf{k}_*^T &= [\mathbf{k}(\mathbf{x}_*, \mathbf{x}_1), \mathbf{k}(\mathbf{x}_*, \mathbf{x}_2), \dots, \mathbf{k}(\mathbf{x}_*, \mathbf{x}_N)] \\ \boldsymbol{\alpha} &= (K + \sigma_n^2 I)^{-1} \mathbf{y} \\ y_{pred}^{GP} &= \mathbf{k}_*^T \boldsymbol{\alpha} \\ &= \sum_{i=1}^N \alpha_i \mathbf{k}(\mathbf{x}_i, \mathbf{x}_*)\end{aligned}$$

4 Linear Regression with $\lambda = 0$

. Without regularisation the solution of Linear Regression looks like

$$y_{pred}^{LR} = \mathbf{x}_*^T (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \mathbf{y}$$

For $\sigma_n = 0$ and kernel $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$ we have

$$\begin{aligned}y_{pred}^{GP} &= \mathbf{k}_*^T (K + \sigma_n^2 I) \mathbf{y} \\ &= \mathbf{x}_*^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}\end{aligned}$$

Normally we don't expect Linear Regression to completely fit the training dataset whereas we know that Gaussian Process (with $\sigma_n = 0$) fits the training dataset completely. Let us try if we can show that $y_{pred}^{LR} = y_{pred}^{GP}$. In class we gave the following argument.

$$\begin{aligned}\mathbf{X}(\mathbf{X}^T \mathbf{X}) &= (\mathbf{X} \mathbf{X}^T) \mathbf{X} \\ (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}\end{aligned}$$

The second equation is obtained by multiplying first equation with $(\mathbf{X} \mathbf{X}^T)^{-1}$ from left and $(\mathbf{X}^T \mathbf{X})^{-1}$ from right. Thus the second equation shows the equivalence of y_{pred}^{LR} and y_{pred}^{GP} . But the above derivation would be valid only if both $\mathbf{X} \mathbf{X}^T$ and $\mathbf{X}^T \mathbf{X}$ are invertible which requires that $d=N$ and \mathbf{X} is invertible. This is definitely not a great achievement. We show a more stronger result below which shows the equality of both if $d \gg N$.

There is a high probability that \mathbf{X} (which is a $d \times N$ matrix) is rank N i.e. there is a high probability we get N linearly independent vectors out of d vectors ($d \gg N$). If \mathbf{X} is rank N , then $\mathbf{X}^T \mathbf{X}$ would be rank N and invertible (Tutorial 1 Problem 11)

Proof : \mathbf{X} is rank N . This is equivalent to saying $\mathbf{X} \mathbf{v} = \mathbf{0} \iff \mathbf{v} = \mathbf{0}$. To show that $\mathbf{X}^T \mathbf{X}$ is invertible, we need to show that its Null Space is $\{\mathbf{0}\}$.

$$\begin{aligned}
\mathbf{X}^T \mathbf{X} \mathbf{v} &= 0 \\
\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v} &= 0 \\
(\mathbf{X} \mathbf{v})^T (\mathbf{X} \mathbf{v}) &= 0 \\
\|\mathbf{X} \mathbf{v}\| &= 0 \\
\mathbf{X} \mathbf{v} &= 0 \\
\mathbf{v} &= 0
\end{aligned}$$

Thus $\mathbf{X}^T \mathbf{X}$ is full rank and invertible. From the first section we know that $\mathbf{X} \mathbf{y} = \mathbf{X} \mathbf{X}^T \mathbf{w}_*$

$$\begin{aligned}
\mathbf{X} \mathbf{y} &= \mathbf{X} \mathbf{X}^T \mathbf{w}_* \\
\mathbf{X}^T \mathbf{X} \mathbf{y} &= \mathbf{X}^T \mathbf{X} \mathbf{X}^T \mathbf{w}_* \\
(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{y} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{X}^T \mathbf{w}_* \\
\mathbf{y} &= \mathbf{X}^T \mathbf{w}_*
\end{aligned}$$

The equation above would have many solutions for \mathbf{w}_* (The dimension of Null Space of \mathbf{X}^T is $d-N$). Consider a particular solution : $\mathbf{w} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y}$

Thus $y_{pred}^{LR} = \mathbf{x}_*^T \mathbf{w} = \mathbf{x}_*^T \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{y} = y_{pred}^{GP}$

d is a rough proxy for capacity of the model. Thus for large d Linear Regression will interpolate (overfit) and memorize all points in the training dataset and thus for large d Linear Regression performs similar to Gaussian Process.

Summarizing the discussion above if $\lambda = 0$ and ($d \gg N$ or ($d \rightarrow \infty$)) Linear Regression is same as Gaussian Process, i.e. Linear Regression on infinite space is same as Gaussian Process. If $\lambda = 0$ and $d < N$ the model does not have enough capacity to memorize all the points and hence Linear Regression is not equal to Gaussian Process.

5 Linear Regression with $\lambda > 0$

From the first section, we already know

$$y_{pred}^{LR} = \mathbf{x}_*^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X} \mathbf{y}$$

For $\sigma_n^2 = \lambda$ and kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$ we have

$$\begin{aligned}
y_{pred}^{GP} &= \mathbf{k}_*^T (K + \sigma_n^2 \mathbf{I}) \mathbf{y} \\
&= \mathbf{x}_*^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{y}
\end{aligned}$$

Now we will show that $y_{pred}^{LR} = y_{pred}^{GP}$ for the given case

Proof : – Using $\mathbf{X} = \mathbf{X}\mathbf{I} = \mathbf{I}\mathbf{X}$ for all matrices \mathbf{X}

$$\mathbf{X}(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}) = (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I}) \mathbf{X}$$

From our earlier discussion we know that both $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})$ and $(\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})$ are invertible. Multiying $(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1}$ from right and $(\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1}$ from left, we will get

$$(\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X} = \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1}$$

Thus we can write y_{pred}^{LR} as

$$\begin{aligned} y_{pred}^{LR} &= \mathbf{x}_*^T (\mathbf{X} \mathbf{X}^T + \lambda \mathbf{I})^{-1} \mathbf{X} \mathbf{y} \\ &= \mathbf{x}_*^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{y} \\ &= y_{pred}^{GP} \end{aligned}$$

Thus we have shown that when regularised, even for finite dimension Linear Regression and Gaussian Process are equivalent.

6 Interpretation of λ

We know that $\lambda \mathbf{I} = \sigma_n^2 \Sigma_p^{-1}$. Thus λ is proportional to inverse of variance of \mathbf{w} (of prior distribution) when there is no data. If the data is good and we put $\lambda=0$ it means that we have confidence on data and hence allow the model to choose from all \mathbf{w} as posterior of \mathbf{w} would have variance $\mathbf{0}$ (Dirac Delta). But when data is not enough (d comparable to N) we lack confidence on \mathbf{w} and thus we set $\lambda > 0$ telling the model to pick \mathbf{w} from a distribution thereby preventing overfitting on the training data.

6.1 Posterior of weights using GP

Gaussian Process Regression gives distribution of $f_* = f(\mathbf{x}_*) = \mathbf{x}_*^T \mathbf{w}$ as:

$$Pr(f_* | \mathbf{x}_*, X, \mathbf{y}) = \mathcal{N}(\mathbf{k}_*^T (K + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (K + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_*) \quad (1)$$

where $k(., .)$ is kernel function, $K(X, X_*)$ denotes $N \times N_*$ matrix of co-variances evaluated for every pair and \mathbf{k}_* denotes the vector of co-variances between the test point and the N training points. We can get this expression through function-space view as described in [?].

Note Mean prediction is a linear combination of observations \mathbf{y} , this is sometimes referred to as a linear predictor. Another way to look at this equation is to see it as a linear combination of N kernel functions, each one centered on a training point, by writing

$$\bar{f}_* = \sum_{i=1}^N \alpha_i k(\mathbf{x}_i, \mathbf{x}_*) \quad (2)$$

where $\alpha = (K + \sigma^2 I)^{-1} \mathbf{y}$.

Now we want to compute posterior for weights i.e. $Pr(\mathbf{w}|X, \mathbf{y})$ using Gaussian Process Regression. But GPR gives us $Pr(\mathbf{x}_*^T \mathbf{w} | \mathbf{x}_*, X, \mathbf{y})$. To obtain desired results, we can have $\mathbf{x}_* \in \{(\mathbf{1}_i)_{i=1}^d\}$ where $\mathbf{1}_i$ is $d \times 1$ vector with all entries 0 except i^{th} entry which is 1. For $\mathbf{x}_* = \mathbf{1}_i$, we will get distribution for w_i . We can do the above evaluation in vectorized form also by giving input as I_d which is Identity matrix of size $d \times d$ (some manipulations must be done since input is changed from vector to matrix). Posterior of weights takes exact same form as equation ?? with $\bar{\mathbf{w}} = K(I_d, X)(K(X, X) + \sigma^2 I)^{-1} \mathbf{y}$ for $K(X, X) = X^T X$ and $\lambda = \sigma^2$. Prediction \hat{y} using GPR is following which is same as that of LR:

$$\begin{aligned} \hat{y} &= \hat{\mathbf{x}}^T \bar{\mathbf{w}} = \hat{\mathbf{x}}^T X (X^T X + \sigma^2 I)^{-1} \mathbf{y} \\ \hat{y} &= \hat{\mathbf{x}}^T (X X^T + \lambda I)^{-1} X \mathbf{y} \end{aligned} \quad (3)$$

Homework Exercise Prove that predicted variance cannot be less than data variance in GPR.

6.2 Summarizing equivalence of LR and GP

$d < N, \lambda = 0$	Linear Regression \neq Gaussian Process Regression
$d < N, \lambda > 0$	Linear Regression = Gaussian Process Regression
$d \rightarrow \infty, \lambda = 0$	Linear Regression = Gaussian Process Regression
$d \rightarrow \infty, \lambda > 0$	Linear Regression = Gaussian Process Regression

Observation is that Linear Regression is equivalent to Gaussian Process Regression in all aspects except when $d < N, \lambda = 0$.

7 Is GP good enough?

If we have small training and test dataset, then should we go for Deep Learning or is GP good enough? Well, GP turns out to be sufficient for such data sets, and DL is not required.

So, in which conditions GP will perform good/bad? GP works well if it gives low variance on test set for points close to training dataset points.

What if GP gives high variance? Then GP has no confidence, it is as good as no prediction. Solution is to combine K-Means Clustering and GP to enhance GP.

7.1 K-means clustering

Consider cluster of data points, to classify a new point \mathbf{x}' , we find \mathbf{x}^* such that $\|\mathbf{x}' - \mathbf{x}\|$ is minimum across all \mathbf{x}_i for \mathbf{x}^* . Predicted cluster y' for \mathbf{x}' will be $y(x^*)$.

Note Gaussian Process tries to induce based on labels (target values) but K-means clustering operates on features (example values).

7.2 Combining Gaussian Process with Clustering

To summarise, for small number of points, rather than applying GP directly, we can do some operation and then perform GP on the dataset. We will be continuing discussion on this in further classes.

References

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