AIML - CS 337

Lecture 17: Regression

15-10-2022

Lecturer: Abir De *Scribe: Margav, Vedant, Nishant, Praneeth, Durgam*

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

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

In vectorized form we can write loss as

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

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

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

Proof for invertiblity of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ when $\lambda > 0$
Consider

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

Thus $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ is a positive definite matrix and hence invertible.
Similarly proof for invertibility of $\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I}$ can be done and we will use that in a later section.

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

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

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

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

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

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

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

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

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

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

Note that $\overline{\mathbf{w}} = \frac{1}{\sigma_n^2}(\frac{1}{\sigma_n^2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1})^{-1}\mathbf{X}\mathbf{y} = (\mathbf{X}\mathbf{X}^T + \sigma_n^2\Sigma_p^{-1})^{-1}\mathbf{X}\mathbf{y}$. 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 Regresion

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 $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x}).\psi(\mathbf{x}')$. Also let $\mathbf{k}_* = [\mathbf{k}(\mathbf{x}_*, \mathbf{x}_1), \mathbf{k}(\mathbf{x}_*, \mathbf{x}_2), ...\mathbf{k}(\mathbf{x}_*, \mathbf{x}_N)]^T$. 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 $\mathbf{k}_* = [\mathbf{k}(\mathbf{x}_*, \mathbf{x}_1), \mathbf{k}(\mathbf{x}_*, \mathbf{x}_2), ...\mathbf{k}(\mathbf{x}_*, \mathbf{x}_N)]^T$. 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 KK^{-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.

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

4 Linear Regression with $\lambda = 0$

. Without regularisation the solution of Linear Regression looks like

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

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

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

Normally we don't expected 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 definetly not a great achievement. We show a more stronger result below which shows the equality of both if $d >> N$.

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($d >> N$). 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 $y^{LR}_{pred} = \mathbf{x}_*^T\mathbf{w} = \mathbf{x}_*^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{y} = y^{GP}_{pred}$

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 ($d >> N$ or ($d \to \infty$)) 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

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

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

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

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

$Proof :-$ Using **X** = **X**I = I**X** for all matrices **X**

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

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

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

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

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

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

7 Inference of w for Gaussian Process

Mean of the probability distribution $P(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})$ is $\mathbf{k}_*^T(K + \sigma_n^2 I)^{-1}\mathbf{y}$. For Bayesian Linear Regression we saw that Mean($P(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})$) = $\mathbf{x}_*^T$Mean($P(\mathbf{w}|\mathbf{X}, \mathbf{y})$).Taking inspiration from the facts that $\lambda > 0$ and thus Linear Regression and Gaussian Process are equivalent, for Gaussian Process we can write $\mathbf{k}_*^T(K + \sigma_n^2 I)^{-1}\mathbf{y}$=$\mathbf{x}_*^T$Mean($P(\mathbf{w}|\mathbf{X}, \mathbf{y})$). To find the i^{th} component of the mean we can take $\mathbf{x}_*$ to be a vector of all zeros except for the i^{th} index whose value is 1. This idea can also be vectorized and it is left as an Exercise.

8 Practical Tips

If training dataset and test dataset are small then Gaussian Process is better to use then Deep Neural Networks whereas the latter is a better choice in presence of large amount of data. To measure the performace of a Gaussian process we can consider the variance of Gaussian process at points closer to the training dataset. The smaller the variance the better is the model. In Gaussian Process we model about the probability distribution of labels. In K-Means Clustering we model about the features. K-Means Clustering gives high variance on y. Thus a solution to this is to fit a Gaussian process on each cluster. If we fix the prior and Kernel there are no trainable parameters left in Gaussian Process. We will continue on this further in the future lectures.

References

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