AIML - CS 419

Lecture 4: Probabilistic Interpretation of Regularization in Regression

25 January 2023

Lecturer: Abir De *Scribe: Groups 7 & 8*

Till now, we have studied about error functions, minimization of error functions and linear regression model. In the last lecture, we discussed maximum likelihood estimation and modeling of non-linear relations. In this lecture, we will study overfitting, regularization and entropy function.

1 Recap of the last lecture

In the last class, we discussed how to model a non-linear relation between input and output.
For that, we can model y as given below:

$$y = \sum_{i=0}^{n} w_i \phi_i(x) \tag{1}$$

where ϕ_i's are non-linear functions of x. now the problem is the same as the linear regression problem seen before. So model y is:

$$y = w^T \phi(x) \tag{2}$$

For minimization of mean squared error, we arrived at a solution

$$w = (\phi^T \phi)^{-1} \phi^T y \tag{3}$$

2 Entropy Function

We have a model as given below:

$$y_{actual} = w^T \phi(x) + \epsilon \tag{4}$$

Although the predicted output is $y = w^T \phi(x)$, we are likely to encounter some **noise** ϵ in the data. Distribution of noise can also affect our model. Let's first look at noises with different variance.

If the noise has high variance, it can lead to overfitting, where the model becomes too complex and starts to fit the noise instead of the underlying patterns in the data, resulting in poor generalization performance on unseen data. On the other hand, if the variance of the noise is low, the model may underfit the data, not capturing important patterns and leading to a poor fit to the training data, which will also lead to poor performance on the test set.

When discussing randomness, we might think that the standard deviation of a given probability distribution would give us a good idea about the type of distribution to work with. The **uniform distribution** :

$$P(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$$

is the probability distribution function with maximum standard deviation. However, the range of values for which it is non-zero is bounded and so, this distribution is not practically useful to us.

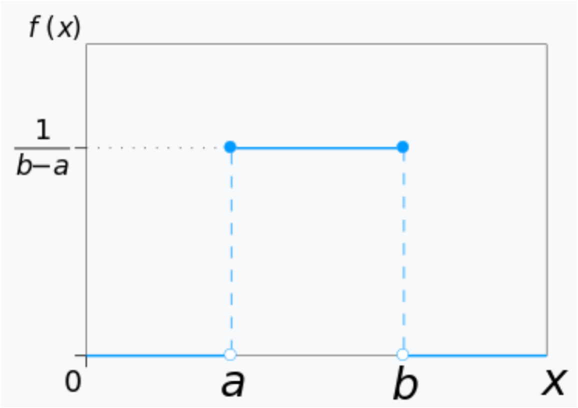

Figure 1: Uniform distribution *source: www.wikkipedia.org*

Now suppose we fix the variance and see how the randomness of data can affect the model. Low randomness noise refers to random and irrelevant information in a dataset that is highly homogeneous, making it difficult for a machine learning model to identify patterns and relationships. Providing some level of randomness and uncertainty in the data can prevent the model from becoming too complex and overfitting to the training data. Additionally, some randomness in the data can also help the model to generalize better to new, unseen data.

Now, what is another good measure of randomness? The answer is **entropy**. *The principle of maximum entropy states that subject to precisely stated prior data (such as a proposition that expresses testable information), the probability distribution which best represents the current state of knowledge is the one with the largest entropy.* So, we can define randomness in terms of the entropy function of distribution $p(x)$ given below:

$$H(p) = -\int_{-\infty}^{\infty} p(x)\log(p(x))dx \tag{5}$$

The **Gaussian** distribution has maximum entropy among all real-valued functions with the same variance. We can intuitively understand why the Gaussian would have a larger entropy than the uniform distribution. Outside the range of non-zero values, the uniform distribution's entropy will evaluate to zero, whereas that of the Gaussian will not instantly vanish as we move away from the mean. Let's compare the entropies of the Gaussian and uniform distributions by evaluating the values of H(p) in each case

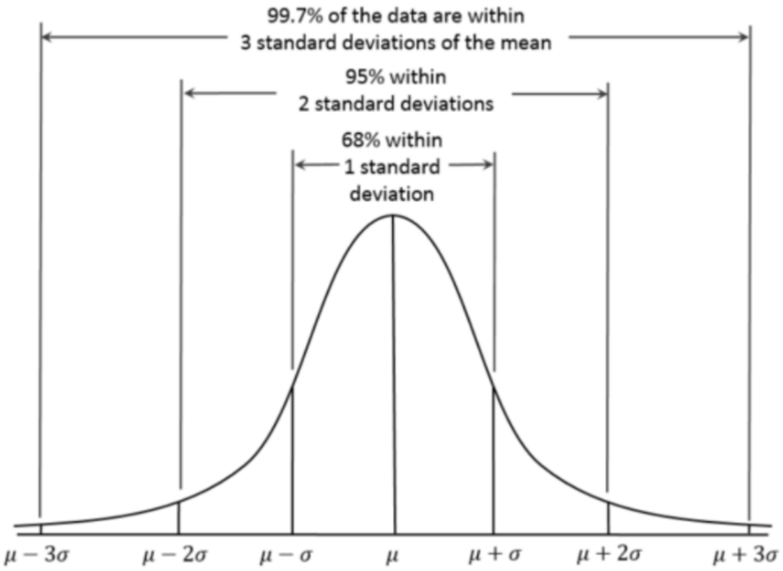

Figure 2: Gaussian distribution *source: www.wikipedia.org*

For a uniform distribution $p(X) \neq 0$ only for a finite interval and so, we only need to calculate the integral over that integral, whereas for the Gaussian, we will have to evaluate the integral over the entire extended real line.
The entropy for the Gaussian distribution can be calculated as :

$$\begin{aligned} H(N(0,\sigma^2)) &= -\mathbb{E}[ln\mathcal{N}(0,\sigma^2)] \\ &= -\mathbb{E}[ln[(2\pi\sigma^2)^{-1/2}exp(-\frac{1}{2\sigma^2}(x-0)^2)]] \\ &= \frac{1}{2}ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[x^2] \\ &= ln(\sigma\sqrt{2\pi}) + \frac{1}{2} \end{aligned} \tag{6}$$

and the entropy of uniform distribution with variance σ^2

$$\begin{aligned} H(U) &= -\int_a^b \frac{1}{b-a} ln(\frac{1}{b-a})\, dx \\ &= -ln(\frac{1}{b-a}) \\ &= ln(\sigma) \end{aligned} \tag{7}$$

which is less than the entropy for a Gaussian distribution. The Gaussian distribution has higher entropy and is non-zero throughout the real line and is therefore chosen to model the noise ϵ in our model y_{actual}.
We have :

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{8}$$

3 Solving for least-square error

Coming back to the least-square error minimized model of our model :

$$w = (\phi^T \phi)^{-1} \phi^T y \tag{9}$$

3.1 Invertibility of $\phi^T \phi$

We can observe that the above equation is meaningless (and hence, will not have a solution) unless the matrix $\phi^T \phi$ is invertible. A square matrix M is invertible (non-singular) *iff* *det*(M) $\neq$ 0.

We have :

$$x_i = \begin{bmatrix} x_i^1 & x_i^2 & x_i^3 & \dots & x_i^d \end{bmatrix}^T \in \mathbb{R}^d$$

and

$$\phi = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \text{; where n = number of elements in the dataset}$$

The above equation (9) can be written as :

$$w = [\sum_{i=1}^{n} x_i x_i^T]^{-1} \sum_{i=1}^{n} (x_i y_i) \tag{10}$$

Here, each of the x_i^j s are continuous random variables (since they are drawn randomly from a continuous distribution) that can take any real value and so each vector x_i has its entries as continuous random variables. Hence, $x_i x_i^T$ is also a matrix of continuous random variables. Therefore the determinant of the matrix $\phi^T \phi$ is also a continuous random variable.
Since the probability of a continuous random variable being equal to a specific value (in this case, zero) is infinitesimal, i.e.

$$P(det(\phi^T \phi) = 0) \sim 0$$

we can safely assume that the matrix $\phi^T \phi$ is invertible.

3.2 Conditioning of Data

One more problem we have to deal with is **poorly conditioned data**. To understand this better, let's take the simplistic example for $n = 2$: We have:

$$\left(\sum_{i=1}^{2} x_i x_i^T\right) = \begin{bmatrix} 0.0001 & 0 \\ 0 & 20 \end{bmatrix} \tag{11}$$

In this case, one can very easily interpret the fact that this solution can result in a highly skewed model which will give large importance to a single parameter while ignoring the other. So in order to tackle this problem, we perform the following manipulation:

$$\left(\sum_{i=1}^{2} x_i x_i^T\right) + \lambda I_2 = \begin{bmatrix} 0.0001 + \lambda & 0 \\ 0 & 20 + \lambda \end{bmatrix} \tag{12}$$

This will reduce the disparity between the values in our original matrix as one can always choose λ large enough so that the resulting matrix will have entries that are of the same order of magnitude. For a general scenario, we use:

$$\omega = \left(\sum_{i=1}^{n} x_i x_i^T + \lambda I_n\right)^{-1} \sum x_i y_i \tag{13}$$

where I_n is the identity matrix of order n.

3.3 Regularization

Now that we have carefully analyzed the solution we obtained for **Minimum Mean Square Error** we will try to simplify it further,

$$MMSE = \min_{\omega}\left(\sum_{i=1}^{n}(y_i - 1 - \omega_1 x_i - \omega_2 x_i^2 ... - \omega_d x_i^d)^2 + \lambda \sum_{j=1}^{n} \omega_j^2\right) \tag{14}$$

For the above expression, if we consider $y_i = \omega_1 x_i + \omega_0$:

$$MMSE = \min_{\omega_0, \omega_1}\left(\sum_{i=1}^{n}(y_i - \omega_1 x_i - \omega_0)^2 + \lambda(\omega_1^2 + \omega_0^2\right) \tag{15}$$

which can be written as,

$$\max_{\omega_0, \omega_1}\left(\prod_{i=1}^{n} e^{-(y_i - \omega_1 x_i - \omega_0)^2} e^{-\lambda(\omega_1^2)} e^{-\lambda(\omega_0^2)}\right) \tag{16}$$

From the above equation, we can conclude that the parameters ω_0, ω_1 are not picked freely from $\mathbb{R}$ but have been sampled from a **Gaussian Distribution**, $\mathcal{N}(0, \frac{1}{\lambda})$. λ is called the regularizing parameter.