CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 5: Regression

Lecturer: Abir De *Scribe: Lecture 5 - Group 2*

5.1 Introduction to Regression

Let's say we are given a dataset {(x,y)} where $y \in \mathbb{R}$. Till now we have considered $y \in \{-1, +1\}$ but here we are considering $y \in \mathbb{R}$.

We need to find mapping from x to y that is $x \mapsto y$

Applications of Regression

1. House price prediction
2. Time series prediction (predicting stocks, loans etc.)
3. Sentiment detection

If we are given a set a data points $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$ we need to find y given x for an unseen sample. So here y is not known and x is not present during training.

Our goal is to come up with some function h(x) so that $h(x) \approx y$.

5.2 When h(x) is linear

If the function $h(x)$ is linear, then

$$h(x) = w^T x$$

For each data point (x_i, y_i), we get $y_i \approx w^T x$, i.e,

$$\begin{aligned} [y_1, y_2, \ldots, y_n] &= w^T [x_1, x_2, \ldots, x_n] \\ Y_{1\times n} &= w^T_{1\times d} X_{d\times n} \end{aligned}$$

To get w, we need to solve the above equation. So, for solution to exist,

$$Y \in \mathbb{R}(X)$$

We have

$$rank(X) = rank(X^T) \leq d$$

Since $d << n$, columns of X^T cannot span entire R^n and as y is n-dimensional vector, we can have $y \in \mathbb{R}^n \backslash \mathbb{R}(X)$. Therefore, solution may not exist.
Also, we need to consider the noise.

5.3 Linear Regression with Error

Let us now consider the equation for linear regression, when including error along with it.

The equation for linear regression when including noise is given by

$$Y = \omega^T x + \epsilon$$

Now, because of this ϵ term, the vector Y no longer lies in the rowspace of x, as y can be contaminated with noise.

$$Y \notin R(X)$$

Now, how do we estimate ω from the given equation for linear regression?

- **Gaussian Distribution**
  Here, if ϵ $N(0, \sigma^2)$, then the optimization function would be the one which maximises the probability that ϵ is 0
  This would be given by:

$$\max_{\omega} \exp \frac{-\epsilon^2}{2\sigma^2}$$
$$\therefore \max_{\omega} \prod_{i \in D} \exp \frac{-(y_i - \omega^T x)^2}{2\sigma^2}$$
$$\therefore \min_{\omega} \sum_{i \in D} (y_i - \omega^T x)^2$$

Using this optimization function, let us solve for the optimal value of ω, which would be ω_1 in our case. This can be done by simply differentiating the optimization function and putting it equal to 0.

$$\frac{d}{d\omega}\sum_{i\in D}(y_i - \omega^T x_i)^2 = 0$$

$$\therefore \frac{d}{d\omega}\sum_{i\in D}(y_i^2 + \omega^T x_i x_i^T \omega - 2y_i\omega^T x_i) = 0$$

$$\therefore \sum_{i\in D}(0 + 2\omega(x_i x_i^T) - 2x_i y_i) = 0$$

$$\therefore \sum_{i\in D}\omega(x_i x_i^T) = \sum_{i\in D} x_i y_i$$

$$\therefore \omega_1 = \sum_{i\in D}(x_i x_i^T)^{-1}\sum_{i\in D} x_i y_i$$

$$\mathbf{W}^* = (\sum_{\mathbf{i}\in\mathbf{D}}(\mathbf{x_i x_i^T}))^{\mathbf{-1}}\sum_{\mathbf{i}\in\mathbf{D}}(\mathbf{x_i y_i}) \tag{5.1}$$

5.4 Existence of the inverse $(\sum_{i\in D}(x_i x_i^T))^{-1}$

We derived :

$$\mathbf{W}^* = (\sum_{\mathbf{i}\in\mathbf{D}}(\mathbf{x_i x_i^T}))^{\mathbf{-1}}\sum_{\mathbf{i}\in\mathbf{D}}(\mathbf{x_i y_i}) \tag{5.2}$$

Let $\mathbf{V} = \sum_{i\in D}(x_i x_i^T)$.
For existence of V^{-1}, we need that $|V| \neq 0$. Now, let us analyze what is the probability of having $|V| = 0$.
Let

$$|\mathbf{V}| = \mathbf{f}(\mathbf{x_1^1}, \mathbf{x_1^2}, ...\mathbf{x_n^{d-1}}, \mathbf{x_n^d}) \tag{5.3}$$

, where $x_i^j = j^{th}$ element of x_i .

Claim: Probability that $f(x_1^1, x_1^2, ...x_n^{d-1}, x_n^d) \to 0$ is 0.

Argument: Say we have generated each x_i^j from $Uni(\cdot)$. It is highly improbable that the multi-variable polynomial function $f(\cdot)$ in $n \cdot d$ variables turns out to be 0. Hence, $Pr[f(\cdot) \to 0] \to 0$

Hence, we conclude that the inverse V^{-1} exists at all times.
Another case of interest is that (ϵ is a very small finite value)

$$|\mathbf{V}| \in [-\epsilon, \epsilon] \tag{5.4}$$

when the V^{-1} will blow up. This has a substantial probability. If this happens, W^* will become very large, which might become a problem as it will cause further calculations to amplify in size Hence, we need a method to tackle this problem.

5.5 Regularization

We saw in the previous section, using the notion of probability of a multi-variable function's zeros being obtained from a random distribution to be 0, that the probability of $det(x_i x_i^T)$ being 0 was negligible.

$$
\begin{aligned}
|\sum_{i\in D} x_i x_i^T| &= f(x_{1i}, x_{2i}, x_{3i}..., x_{dd}) \\
Pr(f(x_{1i}, x_{2i}, x_{3i}..., x_{dd})) &\to 0 \\
\therefore Pr(|\sum_{i\in D} x_i x_i^T|) &\to 0 \\
\therefore (\sum_{i\in D} x_i x_i^T)^{-1} &\quad exists
\end{aligned}
$$

However, we may see that the above function may have a very small value with finite, non-zero probability. In fact, if we have all $x_{ij}\ N(0,1)$, the probability that:

$$|\sum_{i\in D} x_i x_i^T| \in [-\epsilon, \epsilon]$$

is significant.

The issue with this is that according to the previously derived equation for $\omega*$, if the aforementioned determinant is very small, then the inverse of $\sum_{i\in D} x_i x_i^T$ would be very large. This causes the issue that even for a small point (x_i, y_i), the value of ω would be very large, thus making the equation numerically unstable.

Hence, we employ the method of **Regularization**; whereby we obtain the following equation for the ideal parameter ω, by adding an extra hyperparameter λ to our optimization equation, to obtain

$$\omega_2 = (\lambda I + \sum_{i\in D} x_i x_i^T)^{-1} . \sum_{i\in D} x_i y_i$$

Clearly, if the value of λ is large enough, then the value of ω will be numerically stable.

However, how large or small should the value of λ be?

For this, let us take an example scenario, where we have only one sample, that is, $|D| = 1$.

$$
\begin{aligned}
L(w) &= \sum_{i \in D} (y_i - \omega_2^T x_i)^2 \\
\therefore L(w) &= (y_1 - \omega_2^T x_1)^2
\end{aligned}
$$

$$
\begin{aligned}
\lambda &\to \infty \\
\therefore \omega_2 &\to 0 \\
\therefore L(w) &\to y_1^2
\end{aligned}
$$

$$
\begin{aligned}
\lambda &\to 0 \\
\therefore \omega_2 &\to (x_1 x_1^T)^{-1} \\
\therefore L(w) &\to 0
\end{aligned}
$$

However, do note that for a dataset of just a single sample, $x_1 x_1^T$ would clearly not be invertible, because the rank of $x_1 x_1^T$ is 1
Hence, we need to take care of how we set the regularization constant, because if it is higher, then the loss function would not return a value of 0, but if it is too small, then the previous problem of the matrix being non-invertible may creep up.

Now, let us try to find the optimization function for which we obtain ω_2 as the solution.

We already know that the initial optimization problem was given by

$$\min_{\omega} \sum_{i \in D} (y_i - \omega^T x_i)^2$$

The solution for the above problem was given by

$$\omega_1 = (\sum_{i \in D} x_i x_i^T)^{-1} \sum_{i \in D} x_i y_i$$

Now, for the equation obtained after regularization

$$\omega_2 = (\lambda I + \sum_{i \in D} x_i x_i^T)^{-1} \sum_{i \in D} x_i y_i$$

$$\therefore 2\lambda I \omega_2 + 2 \sum_{i \in D} x_i x_i^T \omega = 2 \sum_{i \in D} x_i y_i$$

$$\therefore 2\lambda I \omega_2 + 2 \sum_{i \in D} x_i x_i^T \omega - 2 \sum_{i \in D} x_i y_i = 0$$

$$\therefore \frac{d}{d\omega}(\omega^T (\lambda I)\omega + \sum_{i \in D}(y_i^2 + \omega^T x_i x_i^T \omega - 2\omega^T x_i y_i)) = 0$$

$$\therefore \frac{d}{d\omega}(\sum_{i \in D}(y_i^2 + \omega^T x_i x_i^T \omega - 2\omega^T x_i y_i) + \lambda ||\omega||^2) = 0$$

$$\therefore \frac{d}{d\omega}(\sum_{i \in D}(y_i - \omega^T x_i)^2 + \lambda ||\omega||^2) = 0$$

Hence, we obtain that the optimization problem for the given ω_2 obtained after regularization is given by:

$$\min_{\omega} \sum_{i \in D}(y_i - \omega^T x_i)^2 + \lambda ||\omega||^2$$

5.6 Group Details and Individual Contribution

1. **Mayank Gupta**: Linear Regression with Error

2. **Malhar Kulkarni**: Regularization to prevent determinant of $x_i x_i^T$ to be too small

3. **N Vishal**: When $h(x)$ is linear (section 5.2)

4. **Pradyumna Atreya**: Section 5.4, Existence of the inverse

5. **Tanisha Khandelwal**: Introduction to Regression (section 5.1)