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

We derived :

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

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 $|V| = 0$.

$$|\mathbf{V}| = \mathbf{f}(\mathbf{x_1^1}, \mathbf{x_1^2}, \ldots \mathbf{x_n^{d-1}}, \mathbf{x_n^d}) \tag{5.2}$$

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

1. For two square matrices $A, B$of size $n \times n$ if $AB = BA$ for all B, then show that $A = cI_n$ for some $c \in \mathbb{R}$
   Pick B to be a diagonal matrix with pair-wise distinct elements. Then it can be shown that A is also a diagonal matrix. Now pick B to be a matrix with all ones, i.e. $B = [1]_{ij}$. As A is diagonal, $AB = BA$ implies all diagonal entries of A are equal i.e., $A = cI_n$ for some $c \in \mathbb{R}$

2. If $x^\top Ax = 0 \ \forall x \in \mathbb{R}^n$ then show that A is skew-symmetric.
   On differentiating the above equation we get

   $$(A + A^\top)x = 0 \ \forall x \in \mathbb{R}^n$$

   This implies $A = -A^\top$

3. Show that $\text{rank}(AB) \leq \text{rank}(A)$
   Each column of AB can be viewed as a linear combination of columns of A. Hence, if the dimension of column space of A is r, the dimension of column space of AB cannot be more than r. In other words, $\text{rank}(AB) \leq \text{rank}(A)$

4. Suppose you have a uniform sampler which samples uniformly from $[0, 1]$. Propose an algorithm which uses this uniform sampler to generate samples from any given distribution.
   Suppose we have a uniform random variable $U \sim \text{Uniform}([0, 1])$. We need to find a function g such that the PDF of the random variable $g(U)$ will be same as that of the given distribution, say f. That is $g(U) \sim f$. Which is same as

   $$\begin{aligned} \mathbb{P}(g(U) \leq x) &= \mathbb{P}(U \leq g^{-1}(x)) \\ \implies \int_{-\infty}^{x} f(x)\, dx &= \int_{-\infty}^{g^{-1}(x)} 1\, du \\ \implies F(x) &= g^{-1}(x) \\ \implies g &= F^{-1} \end{aligned}$$

5.4 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.

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

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, $|D| = 1$.

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

$$\lambda \to \infty$$
$$\therefore \omega_2 \to 0$$
$$\therefore L(w) \to y_1^2$$

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

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

$$
\begin{aligned}
\omega_2 &= (\lambda I + \sum_{i \in D} x_i x_i^T)^{-1} \sum_{i \in D} x_i y_i \\
\therefore 2\lambda I \omega_2 + 2 \sum_{i \in D} x_i x_i^T \omega &= 2 \sum_{i \in D} x_i y_i \\
\therefore 2\lambda I \omega_2 + 2 \sum_{i \in D} x_i x_i^T \omega - 2 \sum_{i \in D} x_i y_i &= 0 \\
\therefore \frac{d}{d\omega}(\omega^T (\lambda I)\omega + \sum_{i \in D}(y_i^2 + \omega^T x_i x_i^T \omega - 2\omega^T x_i y_i)) &= 0 \\
\therefore \frac{d}{d\omega}(\sum_{i \in D}(y_i^2 + \omega^T x_i x_i^T \omega - 2\omega^T x_i y_i) + \lambda ||\omega||^2) &= 0 \\
\therefore \frac{d}{d\omega}(\sum_{i \in D}(y_i - \omega^T x_i)^2 + \lambda ||\omega||^2) &= 0
\end{aligned}
$$

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

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

5.5 Group Details and Individual Contribution

1. **Mayank Gupta**:

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**:

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