Statistical Learning

Pranat Sharma And Aaryamitra Pateriya

17th Jan, 2022

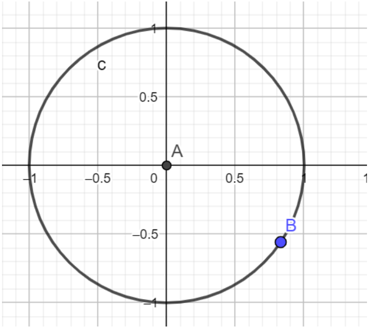
Prelims

We simply start by the definition of what's an expectation value. In simple words expectation value of an event can be defined as a generalization weighted average. Mathematically, expectation value of an event as:
$E(X) = \Sigma_{i=0}^{i=n} x_i p_i$ where $\Sigma_{i=0}^{i=n} p_i = 1$ for obvious reasons.
and $x_i^{\prime} s$ are constants.Let's understand this with the help of an example
A dice has a probability of getting 6 is $1/6$ and $1/6$ for 5 and so on. So, we can write the expectation value of this event as:

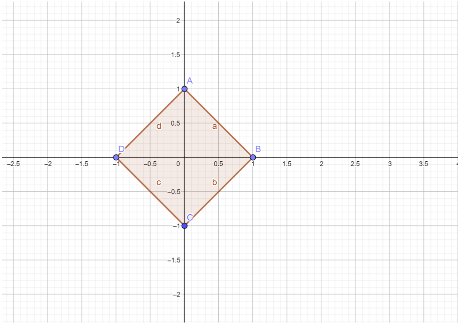
$E(X) = 1.\frac{1}{6} + 5.\frac{1}{6} + 4.\frac{1}{6} + 3.\frac{1}{6} + 2.\frac{1}{6} + 6.\frac{1}{6} = 3.5$
We define variance $Var(X) = \sigma^2(X) = \sigma(x)$ and standard deviation as $\sqrt{Var(X)} = \sigma$
Random Variables and Probability Functions: Suppose you've a sample set S and a real-valued function (X) which assigns each $s \in S$ are real value.
L2-norm of a vector given by $X = (x_1, x_2, x_3)$ is
$||X||_2 = \sqrt{|x_1|^2 + |x_2|^2 + |x_3|^2}$
L1-norm of norm of a vector given by $X = (x_1, x_2, x_3)$ is
$||X||_1 = |x_1| + |x_2| + |x_3|$
Vectors can be represented by a column matrix.
for ex: $\vec{X} = 2\hat{i} + 4\hat{j} + 6\hat{k}$ then we can represent Vector as

$$X = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

NOTE!: Not all column matrices are vectors, but vectors can be represented by them. But, today for our discussion we'll assume all the mentioned column matrices are vectors.


c
A
B
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
1

L2


A
B
C
D
a
b
c
d
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1.5
−1
−0.5
0.5
1
1.5
2
2.5
3
3.5

L1

Linear Regression

Regression models between variables by fitting a straight line to the observed data.Linear regression uses a straight line to fit the data where as nonlinear regression like polynomial regression and logistic regression use a curved line to fit the data.

Types of Linear Regression models:

- Simple Linear Regression
- Multiple Linear Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression

Now, let's talk about them in a bit more detail.

Simple Linear Regression

Simple linear regression formula:
$y = \beta_0 + \beta_1 X + \epsilon$
where y is the predicted value for the dependent variable(y) for any given value of an independent variable(X).
β_0 is the interecept, the predicted value of y when $X = 0$
β_1 is the regression coefficient, how much we expect y to change as x increases.
X is the independent variable.
ϵ is the error estimate or how much variation there is in our estimate of regression coefficients.

Multiple Linear Regression

This model relates y-variables to (p-1)x-variables and can be written mathematically as:

$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p-1} + \epsilon_i$

Error Analysis:

for each data point, using coefficients β there exists some error in our prediction

$\epsilon(\beta) = y - x\beta_1 - \beta_0$

It's mean squared(MSE) and root-mean squared error(RMSE) are given as follows:

$MSE(\beta) = \frac{1}{n}\Sigma_{i=1}^{n}\epsilon_i^2(\beta)$

and $RMSE(\beta) = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\epsilon_i^2(\beta)}$

We also calculate something known as R^2 for our linear regression model. It is mathematically defined as:

$R^2 = 1 - \frac{\Sigma \epsilon_i^2}{\Sigma (y_i - \hat{y})^2}$

R-squared is the measure of how well a linear regression model fits the data. It is a no. between 0 and 1 if it's closer to 1 means more variability i.e. not the best model for our dataset and vice-versa if it's closer to 0. If $R^2 = 0$ the model cannot explain any variability in the outcome . Matrix formulation can be derived as follows:

for $i = 1, ..., n$

$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_1$

$y_2 = \beta_0 + \beta_2 x_2 + \epsilon_2$

...and so on. Thus we can write:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$X\beta = \begin{bmatrix} \beta_0 + \beta_1 . x_1 \\ \beta_0 + \beta_1 x_2 \\ . \\ . \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

Thus now we can write the formula for simple linear regression as follows:
$Y = X\beta + \epsilon$
Hence, MSE can be written as follows:
$MSE(\beta) = \frac{1}{n}\epsilon^T\epsilon$
Now, we define what's known as a "loss" or "cost" function as follows:
$L(\beta) = (X\beta - Y)^T(X\beta - Y) = ||X\beta - y||_2^2$
So, quite naturally we want to minimize the loss function by using the optimum value for β.(For those of you who failed to notice we're just minimizing our before stated MSE). So, the loss function can be minimized for the choice of β as given follows:
$\beta = (X^TX)^{-1}X^TY$

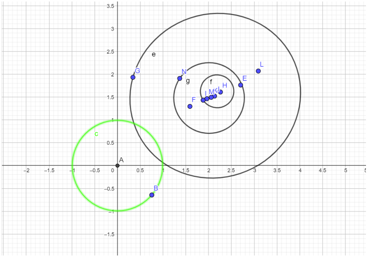
Ridge Lasso Regression

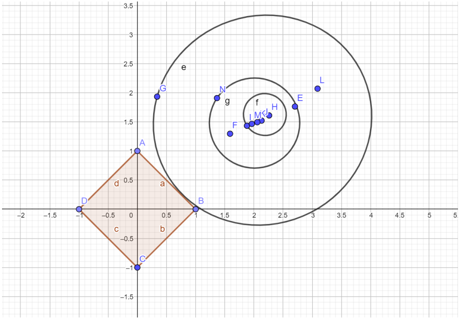
Now, Let's discuss the problems we might face using OLS:

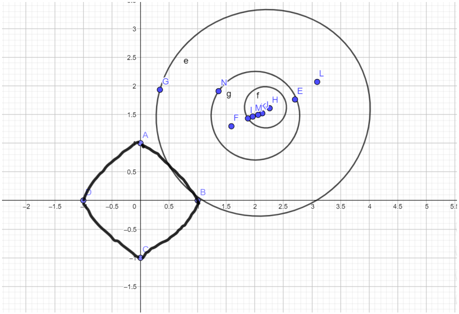
- If X^TX is a singular matrix then $(X^TX)^{-1}$ wouldn't even exist.
- It's not unusual to see no. of input variables exceed the no. of observation.
- Highly correlated data.

Thus to deal with these issues we arrive at what's known as ridge regression:

So, in order to cope harder we introduce a new loss function given by:
$L(\beta_{Ridge}) = ||y - x\beta||_2^2 + \lambda||\beta||_2^2$
where, λ is known as a hyperparameter which is just a numerical no. greater than 0. This is also known as L2-regularized loss function. The choice of β to minimize this new loss function is given by:
$\beta_{Ridge} = (X^T X + \lambda.1_p)^-1X^T Y$
Similarly, we can define L1-regularized loss function which is used in LASSO regression. Furthermore, One can can use both L1 and L2 regularization to yield what's known as Elastic net Regression.But, we won't be talking about them here, today. However, a graphical representation of Ridge,LASSO and Elastic Net regression respectively is be given as follows:


e
g
f
c
A
B
G
N
L
E
F
I
M
K
J
H
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


e
g
f
G
N
L
E
F
I
M
K
J
H
A
B
C
D
a
b
c
d
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


e
g
f
G
N
L
E
F
I
M
K
J
H
A
B
C
D
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

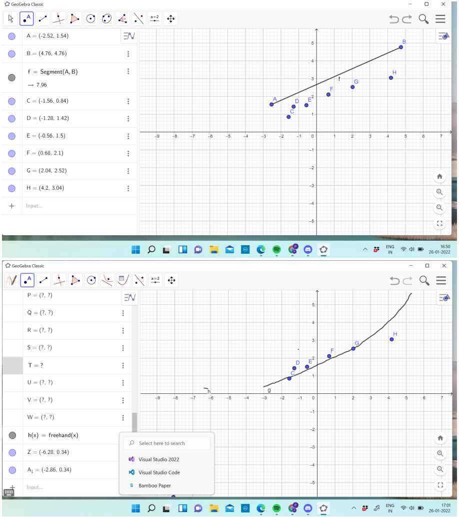
(1)

Non-Linear Regression

Polynomial RegressionIn polynomial regression we model a relationship

between y(dependent variable) and x (independent variable) using a nth degree polynomial. It is given as:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + ... + \beta_n x_n^n$

- It is the linear model,but we've added some new higher order terms to increase accuracy.
- In polynomial regression the original features are converted into polynomial features of required degree and then, modeled using a linear model.
- If we apply linear regression on a non-linear data set, we might not get the desired accuracy. Hence, model like these are of great help for non-linear data set.



A = (-2.52, 1.54)

B = (4.76, 4.76)

f = Segment(A, B)

→ 7.96

C = (-1.56, 0.84)

D = (-1.28, 1.42)

E = (-0.56, 1.5)

F = (0.68, 2.1)

G = (2.04, 2.52)

H = (4.2, 3.04)

P = (?, ?)

Q = (?, ?)

R = (?, ?)

S = (?, ?)

T = ?

U = (?, ?)

V = (?, ?)

W = (?, ?)

h(x) = freehand(x)

Z = (-6.28, 0.34)

A_1 = (-2.86, 0.34)

We see how a polynomial regression might fit a non-linear data set better.