**Machine Learning -2**

**Assignment-2**

**Joint Gaussian PDF**

**Namra Aziz (09590)**

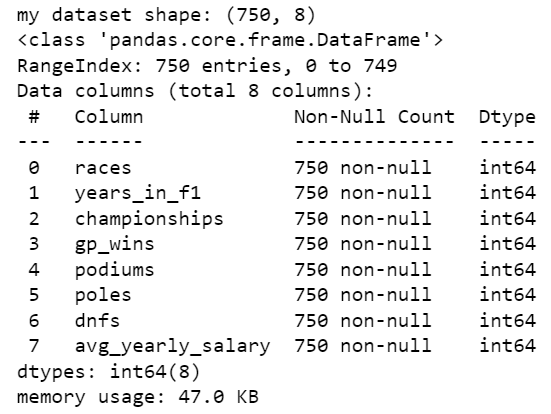
**Dataset:** Formula1

Predicting yearly average salary of car racers.

**Column names: (750,13)**

1. Unnamed
2. Name
3. Country
4. Team
5. Races
6. Years\_in\_f1
7. Championships
8. Gp\_wins
9. Podium
10. Poles
11. Dnfs
12. Fav\_circuit
13. Avg\_yearly\_salary (Target)

**Only numerical columns from this dataset were considered**

****

**Ran regression on the above 7 columns and target:**

Linear Regression Test Accuracy: 0.11404626535031914

Ridge Regression Test Accuracy: 0.11404637375978466

Lasso Regression Test Accuracy: 0.11486775523145476

Elastic Net Regression Test Accuracy: 0.11470748173767498

**Best performing regressor:** Lasso Regression

**Now I located columns that followed gaussian distribution**

I plotted histplots of the above columns and noted their skew value:

1. races, years\_in\_f1, gp\_wins, podiums, poles, dnfs, had very low skewness.

2. These were assumed to be following gaussian distribution.

3. Target variable also had very low skewness.

4. We further calculated shapiro-wilk test to check if these columns were normally distributed**.**

**Shapiro-Wilk Test:** Columns having p-value<0.05 are **NOT** following gaussian distribution.

Highlighted Columns follow gaussian distribution and target is taken as it is.

|  |  |  |
| --- | --- | --- |
| **Column names** | **Shapiro test statistics** | **P-value** |
| races | 0.9982183575630188 | 0.638690710067749 |
| years\_in\_f1 | 0.9785652756690979 | 4.974123335443892e-09 |
| championships | 0.9449057579040527 | 4.4349504270734333e-16 |
| gp\_wins | 0.9975486397743225 | 0.3351175785064697 |
| podiums | 0.9985681772232056 | 0.8132235407829285 |
| Poles | 0.9977558851242065 | 0.4169639050960541 |
| dnfs | 0.984850287437439 | 5.238597964307701e-07 |
| avg\_yearly\_salary | 0.9779317378997803 | 3.2592388787833215e-09 |

**New dataset has 4 features (highlighted) and target.**

**Ran Regression on these normally distributed 4 columns and target:**

Linear Regression Test Accuracy: 0.06245414418000439

Ridge Regression Test Accuracy: 0.06245408940192332

Lasso Regression Test Accuracy: 0.061489628077834424

Elastic Net Regression Test Accuracy: 0.062014702092689

**Best performing regressor**: Lasso Regression

**Generating a sample from the above normally distributed dataset by calculating their mean and covariance**

Mean: 4x1 vector

covariance: 4x4 matrix

Sample: has 4 attributes and target is the same as of above dataset.

**Ran Regression on sample and target:**

Linear Regression Test Accuracy: -0.0014596548351288874

Ridge Regression Test Accuracy: -0.001459679434165917

Lasso Regression Test Accuracy: -0.004580220274946534

Elastic Net Regression Test Accuracy: -0.003518396046212402

**Best performing regressor:** Lasso Regression

**Now, I dropped the normally distributed columns from my actual dataset and substituted the sample columns in place of them.**

**Ran Regression on this new dataset:**

Linear Regression Test Accuracy: -0.004839481633327258

Ridge Regression Test Accuracy: -0.004839405569240007

Lasso Regression Test Accuracy: -0.0033643119970290236

Elastic Net Regression Test Accuracy: -0.003836732801181686

**Best performing regressor:** Lasso Regression

Comparing Results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regressor** | **Accuracy (7 numerical columns with normal and not normal)** | **Accuracy (Normally distributed 4 columns)** | **Accuracy (sample generated of 4 columns)** | **Accuracy (Original data with numerical columns and sample’s columns substituted in place of normally distributed columns)** |
| **Linear Regression** | 0.11404626535031914 | 0.06245414418000439 | -0.0014596548351288874 | -0.004839481633327258 |
| **Ridge Regression** | 0.11404637375978466 | 0.06245408940192332 | -0.001459679434165917 | -0.004839405569240007 |
| **Lasso Regression** | 0.11486775523145476 | 0.061489628077834424 | -0.004580220274946534 | -0.0033643119970290236 |
| **Elastic Net Regression** | 0.11470748173767498 | 0.062014702092689 | -0.003518396046212402 | -0.003836732801181686 |
| **Best performing** | Lasso Regression | Lasso Regression | Lasso Regression | Lasso regression |

**We can see that results declined after generating a sample and in the actual data and generated sample, we have best regressor as Lasso.**

**Highlighted columns compared**

**For actual dataset:**

Accuracy:Lasso>Elastic Net>Ridge>Linear

**For new dataset:**

Accuracy:Lasso>Elastic Net>Ridge>Linear