AI 511 – Machine Learning

Assignment 1

Name: Sani A. Ranpariya

Roll No: MT2021116

TA: Anurag P.

Prof.: Viswanth G.

1. Classification

1.1 Problem statement

The dataset bank auth.csv contains data to classify whether a bank note is authentic or not. The dataset was created by extracting features from images of bank notes. The features extracted are listed below:

- **variance** of Wavelet Transformed image
- **skewness** of Wavelet Transformed image
- **curtosis** of Wavelet Transformed image
- **entropy** of image

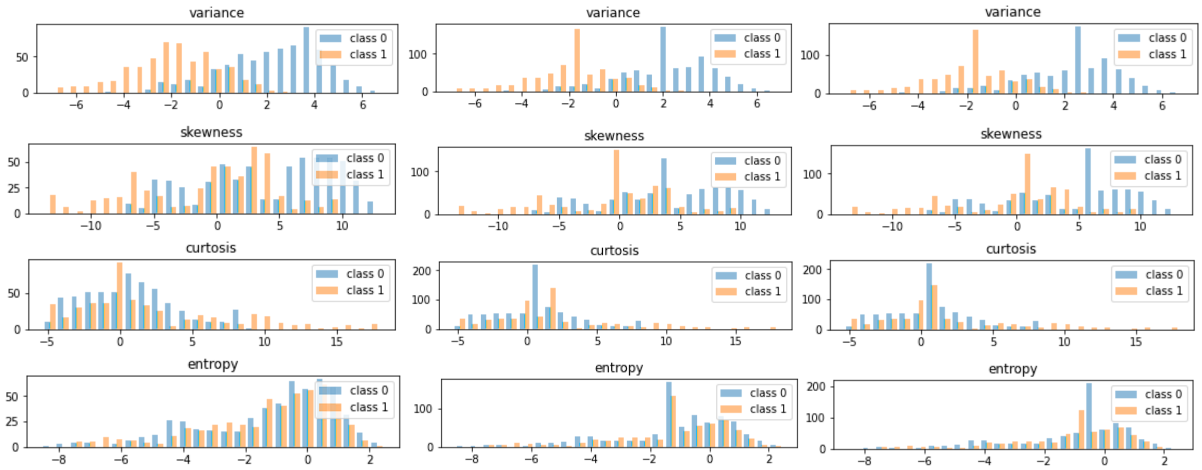
Use **accuracy** as the metric for evaluation.

1.2 Dataset Pre-processing

In most datasets, there are mostly two things we need to handle it which are missing value and non-numeric data. In the bank-auth data set, we don't have any non-numeric data, but we have some missing values. To dealing with the missing value we can drop that row or column or replace it with a mean or median value of that data member.

- **Data field distribution**

Drop null value | Replace with mean | Replace with median



1.3 Naive Bayes Classifier for univariate gaussian

$$P(c = c_k/x) = \frac{p(x/c = c_k) * P(c = c_k)}{p(x)}$$

Naïve Bayes classifier is probabilistic classifier. Therefore, we try to fit data distribution in any probabilistic distribution that fits our data. here we try to fit data in Gaussian distribution. From data distribution visualization, we can see that the variance data field is good to fit in a gaussian distribution.

Normal (Gaussian) Distribution probability:

$$p(x_i/\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

Accuracy:

	Drop null value	Replace with mean	Replace with median
Variance	**0.85**	**0.88**	**0.88**
Skewness	0.61	0.67	0.67
Curtosis	0.62	0.59	0.61
Entropy	0.51	0.54	0.54

Conclusion:

Here we can see that because of variance data field look similar to Gaussian distribution, thus it gives better accuracy compare to skewness, curtosis and entropy.

1.4 Logistic Regression

$$\sigma(\mu) = \frac{1}{1 + e^{-u}}$$

In logistic regression instead of fitting the model into the distribution model, we try to find parameters such that we get a probability of each class of given feature. We use sigmoid function for find probability. We optimize this parameter using the gradient descent method and newton's method.

- **Gradient Descent:**

$$\boldsymbol{\theta}_{(new)=}\boldsymbol{\theta}_{(old)} - \boldsymbol{\mu}\Delta\boldsymbol{\theta}$$

Here, $\Delta\theta_j = \sum_{i=i}^{N}(y_i - P(x_i)) * x_{ij}$

- **Newton's Method:**

$$\boldsymbol{\theta}_{(new)=}\boldsymbol{\theta}_{(old)} - \boldsymbol{H}^{-1}\Delta\boldsymbol{\theta}$$

Here, H = Hessian Matrix

$= [\Delta\theta_{jk}]$

$$\Delta\theta_{jk} = -\sum_{i=1}^{N} P(x_i)\big(1 - p(x_i)\big) * x_{ij} * x_{ik}$$

1.5 Accuracy Result:

Feature	Drop null value	Replace with mean	Replace with median
Gradient Descent	0.92	0.91	0.91
Newton's Method	0.98	0.95	0.96

1.6 Conclusion:

Newton's method uses the second derivative of loss function instead of the first derivative like the gradient descent method, thus it converges faster than gradient descent with better accuracy.

2. Regression

2.1 Problem statement

The dataset garments worker productivity.csv contains required data as features to predict the actual productivity of the workers in the factory. Few of the features, which are not self-explanatory, that are used to predict the actual productivity are listed below:

- **smv:** Standard Minute Value, allocated time for the task
- **wip:** Work in Progress, the number of un_nished items for products.
- **idle men:** The number of workers who were idle due to production interruption.

Use **Mean Squared Error** and **Mean Absolute Error** as the metrics for evaluation.

2.2 Dataset Pre-processing

In given dataset, Quarter column null value replace with “Quarter5”, reason behind going this is to when we count values, only Quarter5 count is very -less compare to other, thus for equalize column value we do this. we drop all row which have null value of day column because we can predict value based on day column. We also drop "targeted_productivity", "wip", "idle_time", "idle_men", "department","day" column because these columns have not proper value or unrelated value. We convert categorical data into one hot encoding data for better use in modelling.

2.3 Linear Regression

Linear regression is used when feature and outcome are linearly related to each other.

Model: $\boldsymbol{y = w_0 + w_1 x_1 + w_2 x_2 + \dots \dots \dots}$

Parameter of the model can be found by closed-form, gradient descent method and newton’s method, in which closed-form give a direct solution of optimal parameter, but it will difficult to find when we have too many features with related parameter. Instead of closed-form use gradient descent and newton’s method which its iterative methods.

- **Closed Form:**

Solution of $\boldsymbol{A\,\vec{w} = B}$

Here A = [x_{ij}]; i = 0…d, j = 0…d

$$\boldsymbol{x_{ij} = \sum_{n=0}^{N} x_{in}x_{jn}}$$

$\boldsymbol{\vec{w} = [w_i]}$**; I = 0…d**

$\boldsymbol{B = [b_{i,}]}$**; I = 0…d**

$$\boldsymbol{b_i = \sum_{i=0}^{d} x_i y_i}$$

- **Gradient Descent:**

$$\boldsymbol{\theta_{(new)=}\theta_{(old)} - \mu\Delta\theta}$$

Here, $\Delta\theta_j = 2\sum_{i=1}^{N}\left(y_i - (\sum_{k=0}^{d} w_k x_k)\right) * x_{ij}$

- **Newton's Method:**

$$\boldsymbol{\theta_{(new)=}\theta_{(old)} - H^{-1}\Delta\theta}$$

Here, H = Hessian Matrix

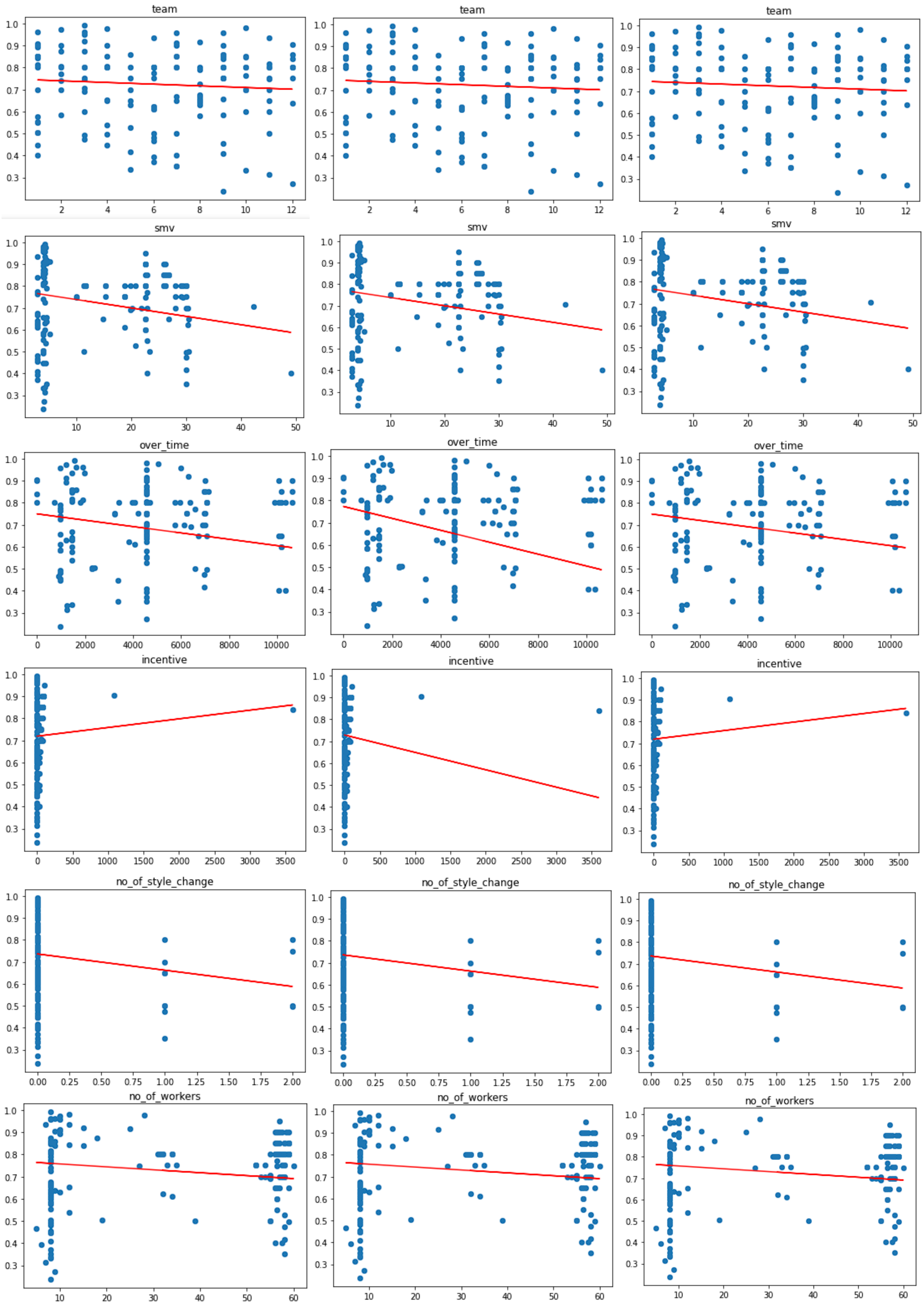
$= [\Delta\theta_{jk}]$

$$\Delta\theta_{jk} = -2\sum_{i=1}^{N} x_{ij} * x_{ik}$$

2.4 Result

- **Univariate linear line**

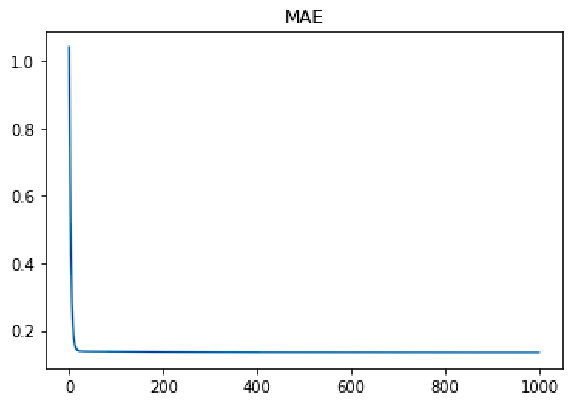
Closed form | Gradient Descent | Newton's method

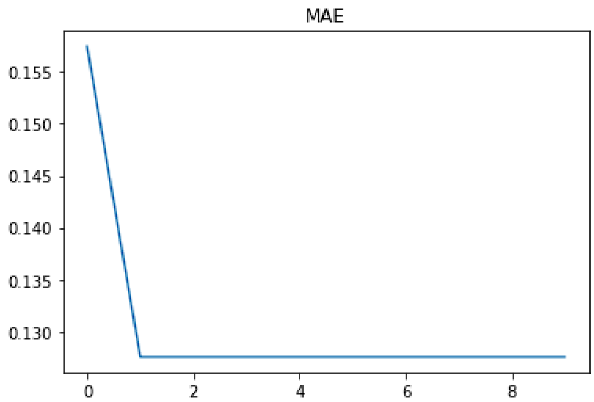


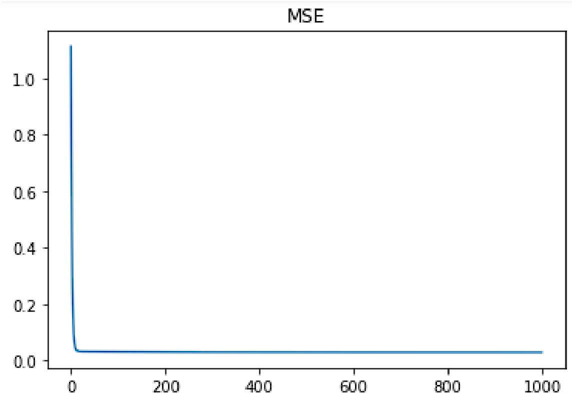
- **Multivariate MAE and MSE error**

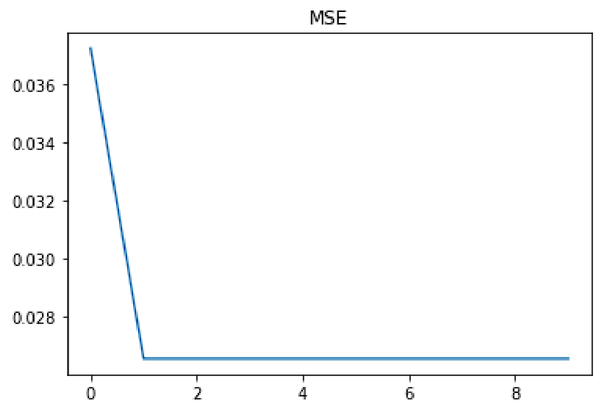
Feature = ['team', 'smv', 'over_time', 'incentive', 'no_of_style_change', 'no_of_workers']

Gradient Descent **Newton's method**









2.5 Error Table

- **Closed form**

Feature	Mean Absolute Error	Mean Squared Error
team	0.138	0.028
smv	0.146	0.031
over_time	0.152	0.032
incentive	0.138	0.028
no_of_style_change	0.134	0.028
no_of_workers	0.144	0.031
Multivariate case	0.174	0.041

- **Gradient descent**

Feature	Mean Absolute Error	Mean Squared Error
team	0.138	0.028
smv	0.146	0.031
over_time	0.135	0.029
incentive	0.139	0.028
no_of_style_change	0.134	0.028
no_of_workers	0.144	0.031
Multivariate case	0.148	0.031

- **Newton's Method**

Feature	Mean Absolute Error	Mean Squared Error
team	0.138	0.028
smv	0.146	0.031
over_time	0.152	0.032
incentive	0.138	0.028
no_of_style_change	0.134	0.028
no_of_workers	0.144	0.031
Multivariate case	0.174	0.041

2.6 Conclusion

Here we can see that newton's method is converge within one iteration while gradient descent takes some time to converge. Thus we can say that newton's method is faster than gradient descent. Even newton's method is fast we mostly use grediant descent method because in newton method we use Hessian metrix which could not be invertable at some time or become sibgular matrix. When feature and parameter is less than we can use closed from solution.