STAT3006 Assignment 4

Ruyun Qi (44506065)

November 16, 2020

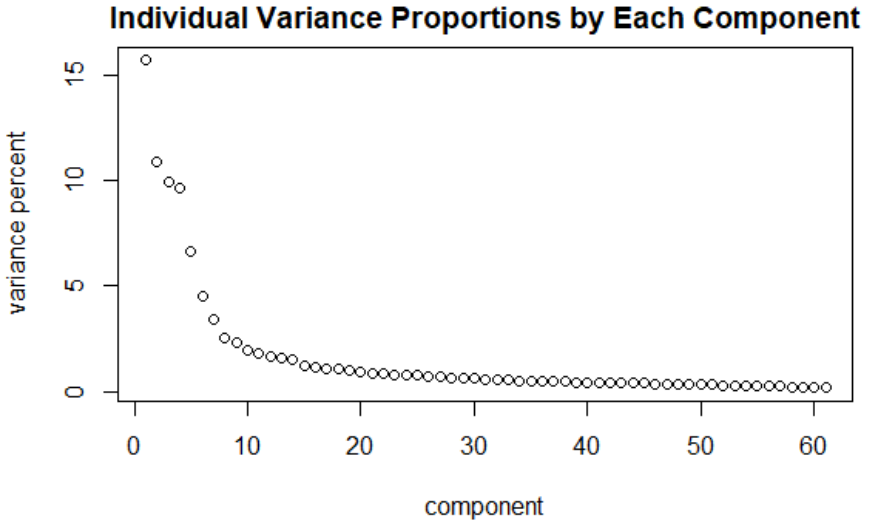
Task 1: PCA

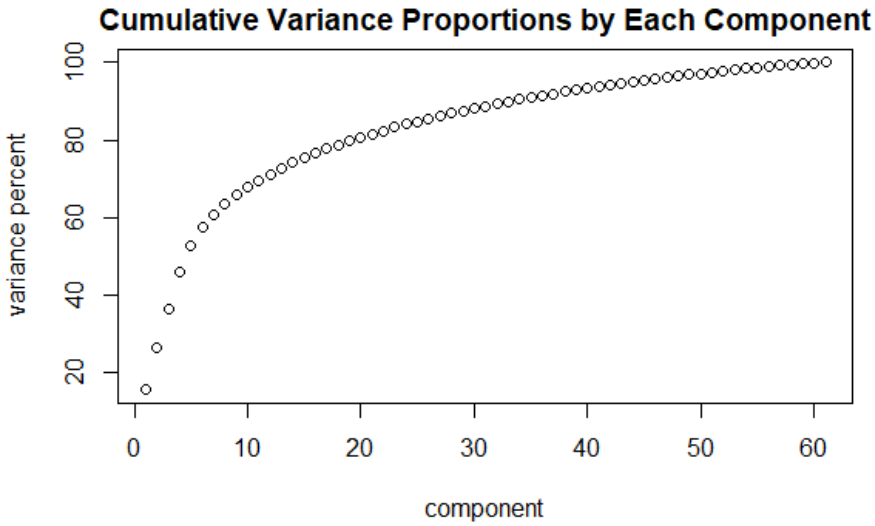
	A	B	C	D	E	F	G	H	I	J	K
1		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
2	1	0.007245	0.089207	0.035765	0.057615	0.159007	0.093555	0.01654	0.01747	0.030543	0.032212
3	2	0.173007	0.033466	0.000172	0.016281	0.00252	0.016163	0.029504	0.001541	0.092675	0.19729
4	3	0.155214	0.039655	0.005691	0.013946	0.008796	0.019313	0.049288	0.002655	0.034328	0.293501
5	4	0.007199	0.009774	0.073045	0.255574	0.024693	0.046198	0.012897	0.008378	5.14E-07	0.104832
6	5	0.05982	0.152434	0.00702	0.042586	0.007746	0.060673	0.076413	0.062074	0.088969	0.006279
7	6	0.014684	0.017075	0.004954	0.000405	0.051832	0.234225	0.350611	0.032022	0.002846	0.000561
8	7	0.001484	0.083002	1.06E-05	0.001434	0.128669	0.154345	0.207857	0.180843	0.078944	0.017522
9	8	0.04561	0.104398	0.012838	0.0616	0.001406	0.060795	0.069519	0.006276	0.027026	0.259811
10	9	0.010528	0.039683	0.002671	0.253454	0.039948	0.011268	0.044916	0.012849	0.0378	0.08164
11	10	0.16506	0.052778	0.01797	0.016247	0.013038	0.048906	0.041081	0.000316	9.12E-05	0.040989
12	11	0.037579	0.155357	0.000348	0.084918	1.02E-06	0.012229	0.004037	0.166677	0.002463	0.060206
13	12	0.0587	0.095897	0.021589	0.070497	0.000437	0.003877	0.052505	0.000473	0.139126	0.1937
14	13	0.010147	0.055348	0.235763	0.000261	0.052371	0.071808	0.008543	0.146296	0.087111	0.00102
15	14	0.068822	0.066659	0.101461	0.048912	0.024118	0.102598	0.051205	0.001135	9.60E-05	0.05997
16	15	0.085948	0.099437	0.028511	0.069955	0.004571	0.055569	0.000685	0.022534	0.0488	0.025571
17	16	0.002514	0.031246	0.173637	0.000176	0.002362	0.165735	0.068845	0.018475	0.098649	0.039187
18	17	0.094605	0.054745	0.070986	0.016673	0.025425	0.02952	0.06011	0.000163	7.97E-05	0.050844
19	18	0.062278	0.112633	0.05226	0.014876	0.027094	0.022532	0.005162	0.102519	0.007984	0.042421
20	19	0.002487	0.078521	0.00162	0.007915	0.140168	0.077607	0.258509	0.289914	0.001926	0.000142
21	20	0.000702	0.039214	1.15E-05	0.021098	0.236253	0.034654	0.226504	0.076624	0.096631	0.01208

The contributions of each gene to each component are included in the file "var_contr.csv". (61 components × 2000 genes)

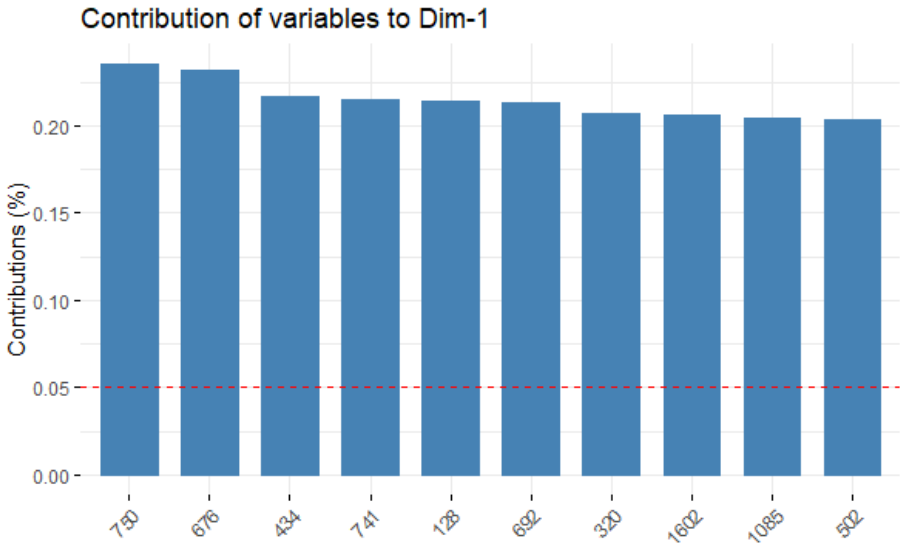
	A	B	C	D	E	F	G	H	I	J	K
1		Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
2	V1	2.096467	0.009995	0.061189	2.683103	1.957627	0.384712	0.606738	2.721802	0.316889	0.680864
3	V2	7.422377	0.23298	3.681456	2.394002	0.614144	13.73386	3.443456	0.977093	3.993375	1.79764
4	V3	0.819786	0.74959	0.84761	0.239811	3.877964	0.842206	0.027142	0.004972	2.14384	1.209749
5	V4	0.572378	0.458195	0.013266	5.344556	0.262002	0.448988	0.055942	1.746388	1.407473	0.003188
6	V5	0.086729	3.588952	3.434593	5.525093	0.979565	1.920198	17.20454	0.666675	0.945946	1.383569
7	V6	5.926367	8.500032	0.78939	2.335319	0.000537	3.135905	0.806701	1.229826	3.028146	1.286079
8	V7	0.06432	0.262318	0.205263	0.502688	0.27137	0.783444	3.811885	0.007297	0.105743	4.783635
9	V8	0.006957	3.677939	0.195648	0.166913	0.400103	0.012058	2.136422	0.294422	5.544938	0.160893
10	V9	0.90208	0.26449	0.206606	3.333831	5.698557	0.008247	1.734373	10.01334	3.459329	0.125521
11	V10	0.178951	3.403097	1.233842	3.266111	1.115183	0.174762	2.087605	1.545592	1.24542	1.0493
12	V11	1.381758	0.930606	0.424397	2.049212	0.002777	2.462687	0.740698	2.584371	0.356673	0.046182
13	V12	3.90672	2.97896	0.036079	0.127264	2.290218	0.032828	2.282363	5.782177	4.762641	3.952296
14	V13	0.155009	1.325796	0.595236	0.191401	0.926686	0.439774	0.045151	0.527648	0.33105	0.066207
15	V14	0.845212	2.809067	2.017651	0.247049	2.108838	0.009973	0.030446	0.36331	0.524403	0.412733
16	V15	0.717137	2.717246	0.539922	0.027274	2.651722	1.121294	0.024538	1.412657	0.011312	0.43237
17	V16	0.768575	0.001815	2.992119	0.271709	2.566494	0.479306	0.000494	0.015839	0.655999	1.11761
18	V17	2.943031	5.097103	0.143276	1.010661	4.261976	0.000924	0.000843	0.175175	0.00914	0.023789
19	V18	0.115729	3.367615	0.661023	1.431722	0.975363	0.036607	0.024575	1.344354	0.293946	0.00946
20	V19	1.388513	0.05169	0.012066	0.004104	0.769371	1.386029	0.24368	10.38144	3.084521	5.119818
21	V20	5.434918	4.957775	0.045878	0.135272	0.042529	1.161435	0.076167	4.232442	1.358901	1.071206

The contributions of each sample to each component are included in the file "ind_contr.csv". (61 components × 62 samples)

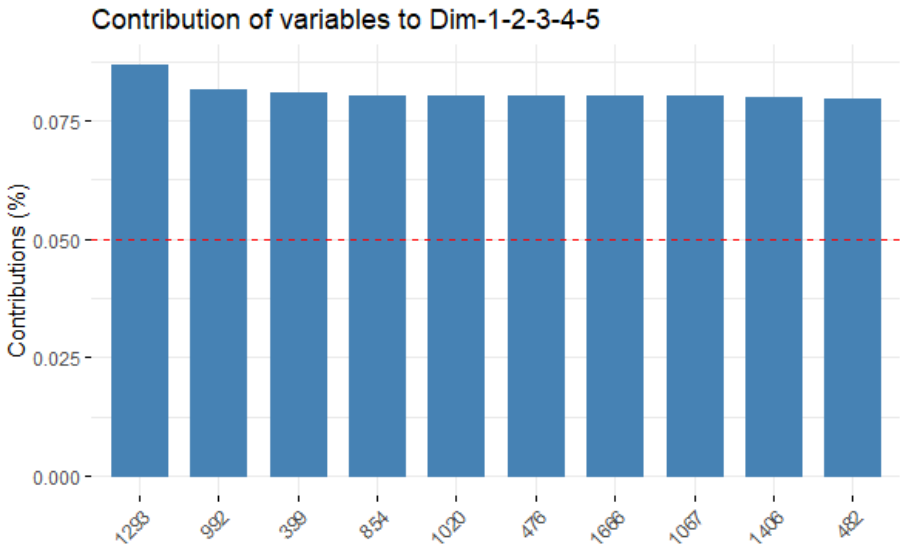




We can see that 90% of the variance is covered by 40-50 principal components.



This figure shows the top 10 significant genes for PC1. However, the first principal component only covers approximately 15% of the variance.



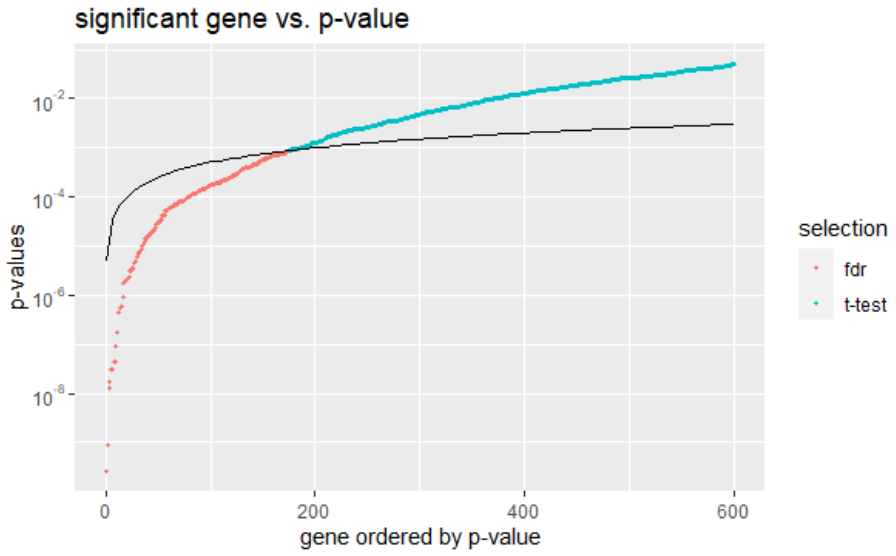
Instead of just PC1, this figures shows the top 10 significant genes for PC1-5. We can see that the top significant genes and contributions are totally different from those of PC1. Therefore, the judgement of cancer or non-cancer cannot be decided by the first few principal component.

Task 2: Single Variable Analysis

I used Welch's t-test to test the p-values of genes divided by the cancer and non-cancer samples with the confidence level 95%. Therefore, the p-value < 0.05 can be considered as significant.

	A	B	C	D	E
1		gene	p_value	fdr	judgement
2	1	1635	2.66E-10	5.31E-07	significant
3	2	493	8.74E-10	8.74E-07	significant
4	3	1843	1.32E-08	8.39E-06	significant
5	4	377	1.68E-08	8.39E-06	significant
6	5	1042	3.09E-08	9.10E-06	significant
7	6	1423	3.13E-08	9.10E-06	significant
8	7	513	3.18E-08	9.10E-06	significant
9	8	897	4.39E-08	9.78E-06	significant
10	9	249	4.40E-08	9.78E-06	significant
11	10	625	9.37E-08	1.87E-05	significant

173 genes are labelled as 'significant' with the control of fdr ≤ 0.01 in the file "significant_genes.csv".



The function of the line indicating the FDR control is defined as:

$$P(i) = \frac{i}{2000} \times 0.01$$

Task 3: Binary Lasso Logistic Regression

In binary logistic regression, the response variable is a Bernoulli random variable, so given any new data $\boldsymbol{x}$, we can obtain the probability of $\boldsymbol{x}$ belonging to either class 0 or 1 as

$$\Pr(y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}) = \frac{e^{\beta \boldsymbol{x}}}{1 + e^{\beta \boldsymbol{x}}}.$$

The probability density function is defined as

$$f(y) = \pi^y (1 - \pi)^{1-y}.$$

The likelihood is

$$L(\beta, y_i) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Thus, the log-likelihood is

$$\ell(\beta, y_i) = \sum_{i=1}^{n} \{y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\}.$$

The penalized logistic regression (PLR) is defined as

$$PLR = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{e^{\beta x_i}}{1 + e^{\beta x_i}}\right) + (1 - y_i) \log\left(1 - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}\right) \right\} + \lambda P(\beta),$$

where λ is a tuning parameter and $P(\beta)$ is the penalty term.
Optimization function of LASSO is

$$\hat{\beta_{LASSO}} = \underset{\beta}{\operatorname{argmin}} \left\{ \ell(\beta, y_i) + \lambda \sum_{i=1}^{p} |\beta_j| \right\},$$

Thus, the lasso logistic regression (LLR) is defined as

$$LLR = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{e^{\beta x_i}}{1 + e^{\beta x_i}}\right) + (1 - y_i) \log\left(1 - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}}\right) \right\} + \lambda \sum_{i=1}^{p} |\beta_j|.$$

Task 4: Benefits and Drawbacks of PCA

Benefits of PCA:

- Removing correlated features which reduces the number of features and providing the non-correlated principal components.
- Improving algorithm performance and reducing over-fitting by reducing the input dimensions.

Drawbacks of PCA:

- Might be hard to find the optimal principal components for abnormal data.
- Information loss is unavoidable during the reduction of data dimensions.

For Neural Network, I decided not to implement PCA since it has strong non-linear computing power . However, when I was working on the coding, errors occured due to the large size of the network. After applying PCA, the training process has been completed in a short time.

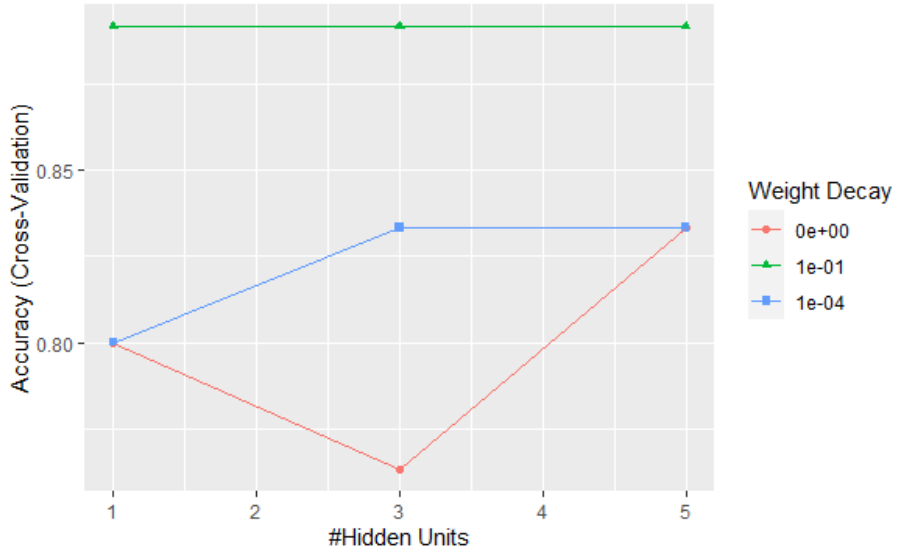
For Lasso Logistic Regression, I decided not to implement PCA to avoid information loss. Also, PCA takes 40-50 principal directions which means the reduction of dimensions does not make too much differences compared to the original 62 dimensions.

Task 5: Two Classification Methods

I spitted the data into the training set (60%) and the test set (40%).

1. Neural Network

(a & b) parameter estimates and CV-based error rates



The 10-fold CV was used to find the optimal hyper-parameter for NNet. The final values used for the model were size = 1 and weight decay = 0.01, which have the highest accuracy 0.8916667.

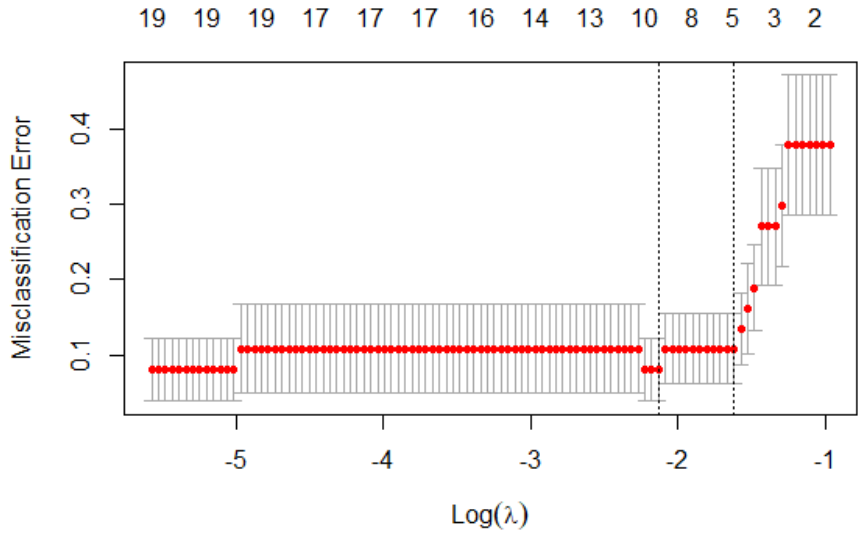
```
The final model:

a 28-1-1 network with 31 weights
options were - entropy fitting  decay=0.1
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1 i10->h1
   0.08    0.07   -0.32   -0.09   -0.19    0.37   -0.14   -0.11   -0.08   -0.17    0.05
i11->h1 i12->h1 i13->h1 i14->h1 i15->h1 i16->h1 i17->h1 i18->h1 i19->h1 i20->h1 i21->h1
   0.15    0.29   -0.09    0.19    0.13   -0.01    0.07    0.06   -0.03    0.14   -0.04
i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1
   0.19   -0.05   -0.30    0.06   -0.02   -0.10    0.04
 b->o h1->o
-2.14  5.21
```

CV-based error rates:

class <chr>	error_rate <dbl>
Normal	0.1250000
Tumour	0.3529412
Overall	0.2800000

2. Binary Lasso Logistic Regression

(c & a) hyper-parameter and parameter estimates



The figure plots the $\log(\lambda)$ vs. the misclassification error.

```
Measure: Misclassification Error

    Lambda Measure      SE Nonzero
min 0.1190 0.08108 0.04210      10
1se 0.1985 0.10811 0.04696       5
```

$\lambda = 0.1190$ gives the smallest mean cross-validated error.
$\lambda = 0.1985$ is chosen for the model, which is the largest value of lambda such that error is within 1 standard error of the minimum.

Predictor variables (gene index) with their coefficients with $\lambda = 0.1985$:

	coefficient <dbl>
377	-0.20150453
493	-0.38429152
571	0.06627460
601	0.00599072
1771	0.35171012

(b) CV-based error rates

class <chr>	error_rate <dbl>
Normal	0.3750000
Tumour	0.2352941
Overall	0.2800000

Task 6: Comparisons for 4 Approaches

1. PCA is a common approach to reduce the dimensions of data. It figures out the contributions of each variable to each sample and reconstructs the contributions into principal components. If the first 2 principal components take the most of the variances, then their 2-D plot is more understandable than high-dimensional visualization. However, in the case of Alon data, taking at least 40 principal components can roughly take only 90% variance. Therefore, it is difficult to visualize the most significant genes by all selected principal components. Also, the principal components are not interpretable biologically.

2. he two-sample t-test is testing for the differences between the means of 2 classes. The gene with a p-value less than 0.05 can be denoted as 'significant' genes. FWER might cause miss finding because of the large number of genes in the analysis, while FDR provides a less strict multiple-testing criterion. Applying FDR to the t-test gives us a straightforward view of the genes which contribute the most to distinguish cancer and non-cancer samples. It is more interpretable than PCA without constructing new components.

3. Neural Networks consider variable interactions and create a non-linear prediction model. In the case of Alon data, PCA helps the training of NNet in the dimensional reduction to make the networks less complicated. Since I did not get the prediction results without PCA as previously mentioned, I could not judge whether applying PCA to NNet helps improve the accuracy of predictions. The final accuracy which is above 80% seems that NNet is not a bad classifier in this case.

4. Binary Logistic Regression builds prediction models for the outcomes of cancer or non-cancer samples. Since the number of genes is 2000 but the number of samples is 62, a Lasso penalty function is added. It discovered 5 significant genes with the optimal λ obtained in a 10-fold CV. It can be considered as a one-layer Neural Network since logistic sigmoid functions are commonly used in the hidden layer of a neural network. It might cause similar accuracies of these two methods in this case.