Assignment 2
ELL409 : Machine Intelligence and Learning

Navreet Kaur
Indian Institute of Technology, Delhi
2015TT10917
tt1150917@iitd.ac.in

Siddharth Khera
Indian Institute of Technology, Delhi
2015MT60567
mt6150567@iitd.ac.in

Utsav Sen
Indian Institute of Technology, Delhi
2015MT60569
mt6150569@iitd.ac.in

I. Introduction

We have used different models like Linear, Logistic, Perceptron, Support Vector Machines and Discriminative for classification on three different datasets(F-MNIST, Medical and Railway), and Linear models for regression on the River dataset.

Accuracies reported in the tables are the highest ones obtained after tuning all the hyperparameters:

- Linear: Degree of polynomial kernel, regularization parameter
- Logistic: Degree of polynomial kernel, regularization parameter, learning rate
- LDA: the number of dimensions of the projection was varied to obtain best results
- SVM: Different kernels were used
    - Linear - constant term *b* in $(b + X^TX)$ was tuned
    - Polynomial - *a, b, d* were tuned in $(a + b^* X^TX)^d$
    - RBF - *a, s* were tuned for $\exp(a^*|X_i-X_j|^2 / s)$

II. Classification

A. Fashion MNIST Dataset

Different classification schemes used for this task and their accuracies are shown in Table 1.
These accuracies correspond to the highest accuracies obtained by varying number of components and trying out different hyperparameters.

Model	Parameters	Train Accuracy	Test Accuracy
Bayesian	#PC* = 80	71.89	70.51
Naive Bayes	#PC = 80	71.26	69.51
K - Means	PCs = 10, n** = 2	48.57	46.90
Linear	#PC = 80 ***	81.02	80.46
Logistic	#PC = 80 ***	80.96	80.45
	#PC = 5 ****	68.02	67.5
Perceptron	#PCs = 80	69.47	70.42
SVM	#PCs = 80, Linear Kernel	87.26	86.03
LDA	#PCs = 80, Projected Dim = 40	79.49	82.09

Table 1. Classification Schemes and their Accuracies
* #PC = no. of principal components
** n-number of initialisations
*** with regularisation
**** without regularisation

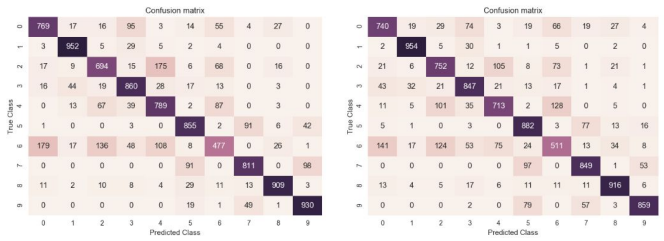


Figure 1. Confusion Matrix on test data (Linear and Logistic)

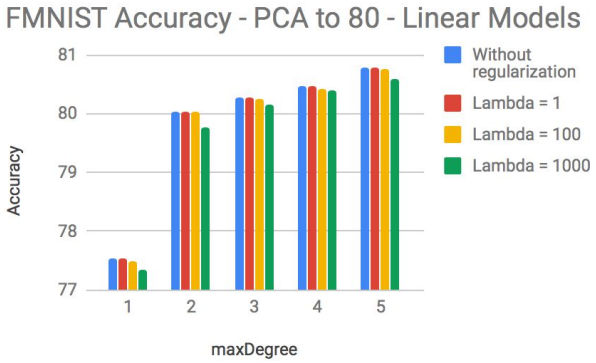
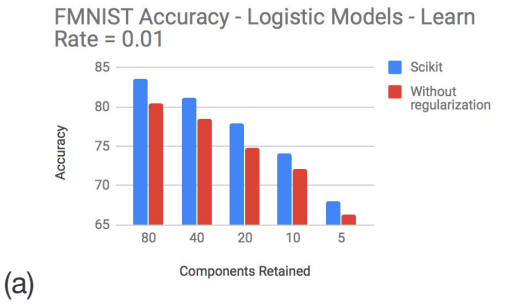
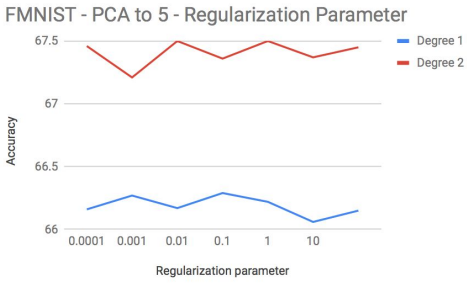


Figure 2. Variation of Accuracy by varying regularization parameter in linear models



(a)



(b)

Figure 3. Variation of Accuracy by varying regularization parameter in logistic models

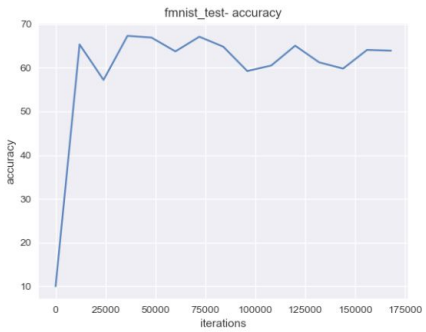


Figure 4. Accuracy with increasing iterations of Perceptron

Performance of logistic for $\phi(x)$ of degree 2 is better.

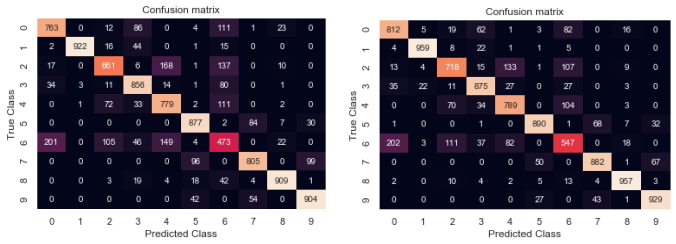


Figure 5: Confusion Matrix on test data (LDA and SVM)

Applying PCA before LDA

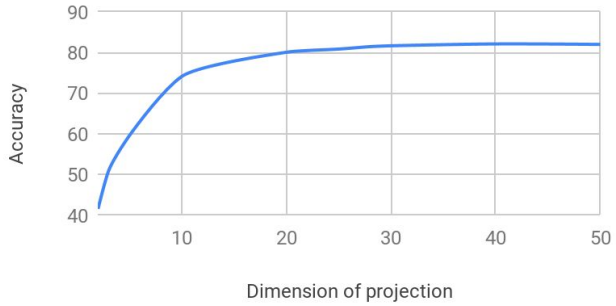


Figure 6: Accuracy of LDA v/s dimension of projection

For very high dimensions, inverting S_w is a very sensitive operation that can only be done if we have a precise measure of the quantity and it is difficult to obtain a precise estimate of S_w for such high dimensions. Therefore, training data must be significantly larger than the size of feature space for a good estimate, otherwise, it will be almost singular and cause overfitting instead (although not the case for F-MNIST). Therefore, we kind of regularise the problem by applying PCA to reduce the dimensionality and that improves generalisation.

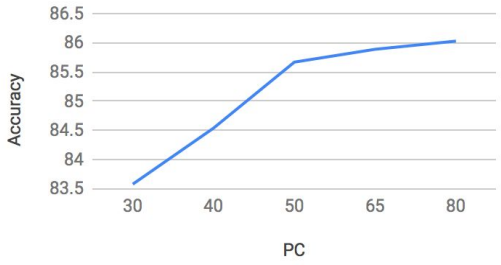


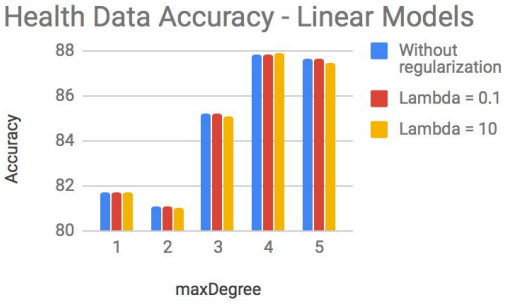
Figure 7: Performance of SVM with number of principal component with a linear kernel

OVO(One-vs-One) trains K(number of classes) classifiers. For each class *i* it will assume *i*-labels as positive and negative. This might lead to problems if the dataset is imbalanced i.e. there are not similar number of instances for each class. In OVR(One-vs-Rest), K(K-1)/2 classifiers have to be trained and it is less sensitive to imbalanced dataset but is computationally expensive.

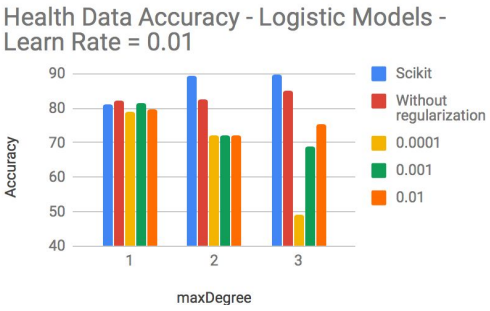
B. Blood Test Dataset

Model	Train Accuracy	Test Accuracy
Bayesian	90.47	89.81
Naive Bayes	90.66	89.86
K - Means	77.84	77.1
K- NN*	89.90	89.433
Linear****	88.04	87.90
Logistic****	89.98	89.66
Perceptron	82.02	81.86
SVM**	83.46	83.43
LDA***	86.2	84.4

Table 2. Classification Schemes and their Accuracies
*with k = 9
**with linear kernel
*** dimension of projection = 2
**** without regularisation



(a)



(b)

Figure 8. Variation of accuracy with regularisation (Linear and Logistic)

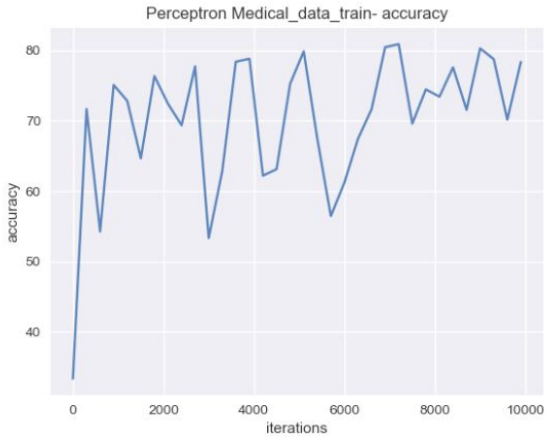


Figure 9. Accuracy with increasing iterations of Perceptron

Significant difference between accuracies on decreasing the number of iterations is observed for Perceptron which is also expected as it is essentially an error correcting algorithm, hence, more passes on data or seeing more training examples should intuitively generalise more.

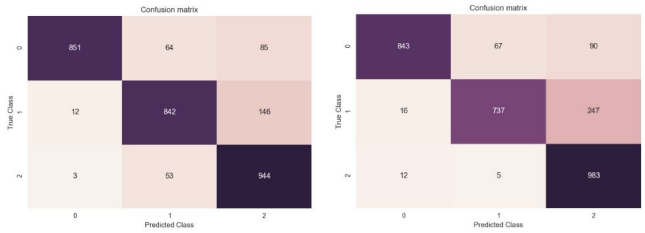
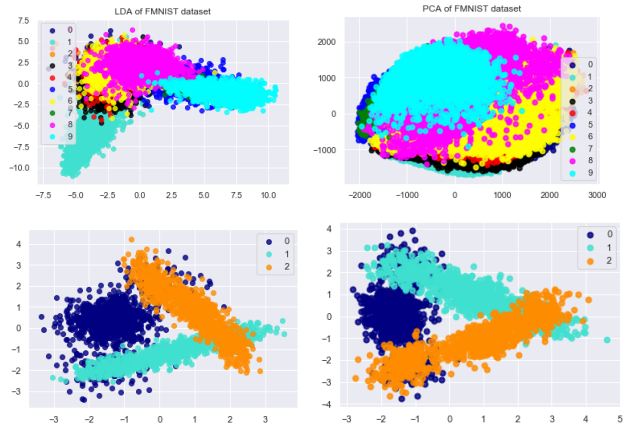


Figure 10. Confusion Matrix on test data (Linear and Logistic)

Comparison of LDA and PCA



(A)

(B)

Figure 11. LDA and PCA of data in 2-D space(F-MNIST and Medical data)

Both project data on a lower dimension. PCA maximises the variance of data while LDA finds a

feature subspace that maximises the class separability. In the above Figure, it can be noticed that LDA tries to seperate the overlap between data points of different classes, however, in PCA projection, the data points of either two classes overlap considerably and the overall data is also spread out (doesn't lie very close together).
LDA makes assumption of normal distribution and of equal class covariance. QDA can perform better in this case since covariance of class 0 differs from class 1 and 2 while that of 1 and 2 is similar.

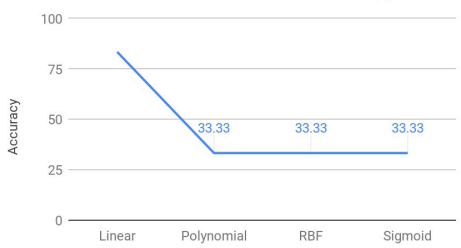


Figure 11. Performance of SVM for different kernel types

Kernels other than linear in this case classify all the data points to the same class thus giving an accuracy of 33.33 for polynomial, rbf and sigmoid.

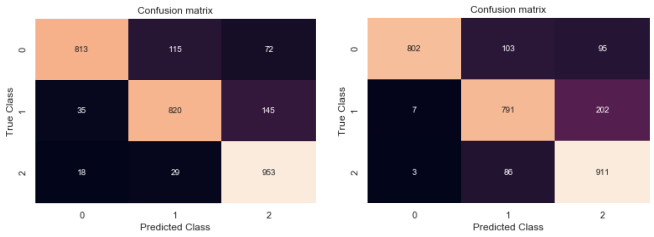
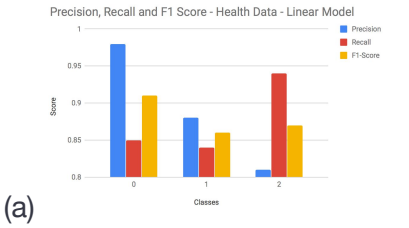
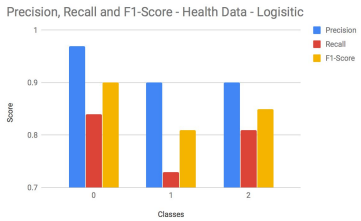


Figure 12. Confusion matrix on test data (LDA and SVM)



(a)



(b)

Figure 13. Performance Parameters

C. Train Selection Dataset

The dataset contains all the information about the person booking the train along with whether the person has boarded the train or not. Classifiers are built to predict whether a person will board the train or not if provided with information such as age, fare paid, number of members traveling with etc.

Model	Train Accuracy	Test Accuracy
Bayesian	78.51	77.10
Naive Bayes*	75.95	75.54
K - Means	64.46	64.00
K- NN**	80.60	80.1
Linear^{+}	82.45	82.06
Logistic^{++}	82.01	81.64
Perceptron	82.89	82.53
SVM***	78.6	77.4
LDA****	71.34	70.6

Table 4. Classification Schemes and their Accuracies

* Assuming Gaussian CCDs on continuous variables and Multinomial on categorical variables {budget: Gaussian, number_count: Gaussian, sex: Multinomial, preffered_class: Multinomial, Age: Gaussian}
** with k = 15
*** linear kernel
**** dimension of projection = 4
$^{+}$ with 1 regularisation
$^{++}$ with 0.01 regularisation

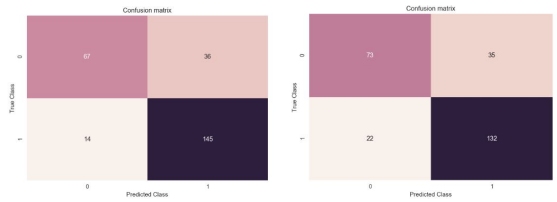
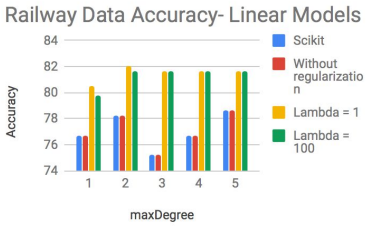
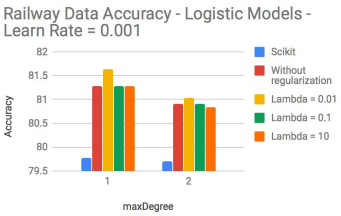


Figure 14. Confusion Matrix on test data (Linear, Logistic)



(A)



(B)

Figure 15. Variation of accuracy with regularisation

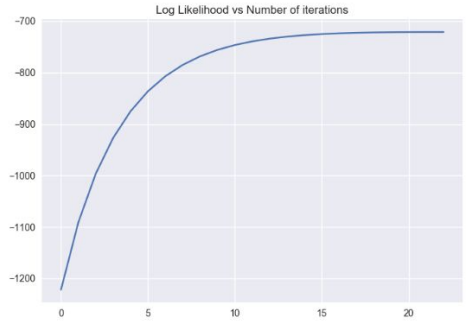


Figure 16. Change in log likelihood with number of iterations of Logistic Regression

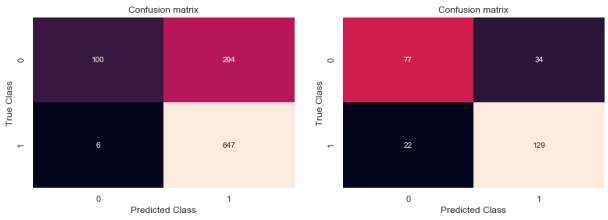


Figure 17. Confusion matrix on test data (LDA and SVM)

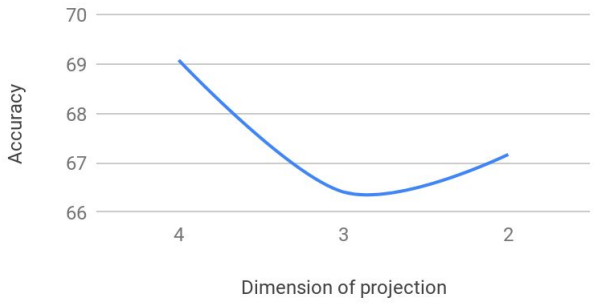


Figure 18. Accuracy of LDA v/s dimension of projection

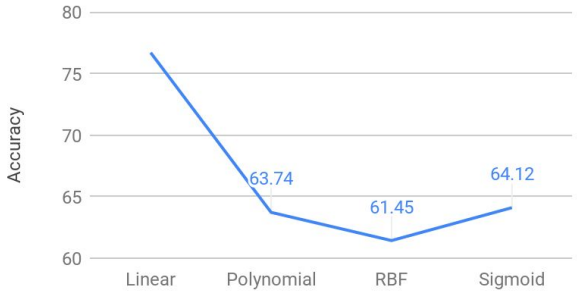


Figure 19: Performance of SVM for different kernel types

D. River Dataset

Figure shows the visualisation of data in 2-D. It seems to come from a 3-degree polynomial function.

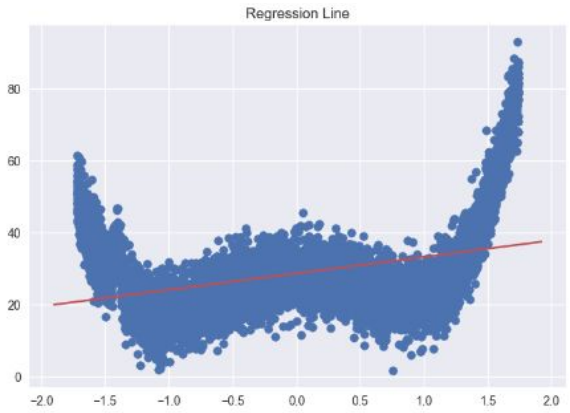


Figure 20. Visualisation of data in 2-D and predicted curve using Linear models with $\phi(X) = X$

Model	Train Accuracy(R^2)	Test Accuracy(R^2)
Linear	0.84	0.81
SVM	0.75	0.74

Table 5. Regression Models and their Accuracies

Highest accuracy (lowest RMSE) is obtained from linear model by keeping degree of X as 4, as is also evident from the scatter plot that the approximation function for this data would be >=3 degree polynomial.

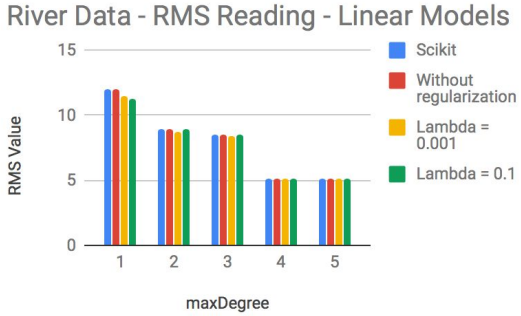


Figure 21. Performance of Linear Model

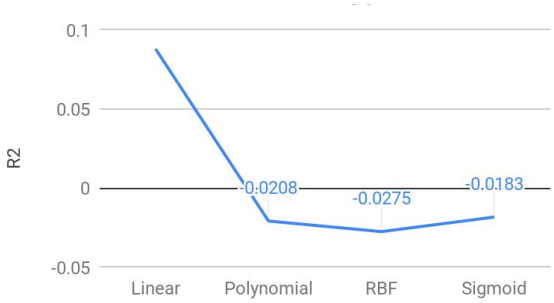


Figure 22. Performance of SVM(R^2) for different kernel types

III. Inferences

Performance of SVM and Linear Separability
On all datasets, It is observed that SVM has been the best performer except for regression in River dataset. Given the fact that SVM attempts to find the optimal separating hyperplane, we can infer that the data we have is linearly separable in the projected space.

Regularization for multi class Logistic
L2 regularization in multi-class regularization performs worse than with no regularization. The reason for the above can be stated as the fact that the ‘b’ (intercept) term is included in the weights and hence is also regularized. Thus, it leads to an additional constraint of keeping the intercept small which might not be the case with the data available.

Irregular accuracy vs iterations for Perceptron
The graph of Perceptron’s accuracy on the test dataset against the number of iterations reaches a max region and then oscillates irregularly about the same until the number of iterations are exhausted. This implies that at a certain number of iterations, the separating hyperplane is found and all further iterations keeps moving the hyperplane around, not necessarily improving it.