CS5691: PRML Assignment 2

V.S.S.Anirudh Sharma (EE18B036), Hema Landa (EE19B036)

March 3, 2022

1 Abstract

This report works of Least Square Regression, Ridge Regression, and 5 different assumed cases of Bayesian classification, over different synthetic data sets. The reader will be taken through the process of tuning the hyper-parameters of models and drawing inferences from results.

2 Regression

2.1 Introduction

The idea is to implement polynomial curve fitting with Least Square and Ridge Regressions. Analysing the errors obtained for different orders of the polynomial and different regularization parameters, the aim is to find the best model fit for the data. Two types of data sets have been considered to perform the above actions over: univariate and bivariate data.Each type has training and development datasets. We will be training the models with the training data and optimize the hyperparameters, M (order of the fit polynomial) and λ (regularization parameter) to minimize the development data prediction errors.

2.2 Univariate Data

Fig 1 shows the scatterplot of training and development data. 200 points per each have been given. The x-axis is the 1-D input (variable) and the y-axis is the output (response).

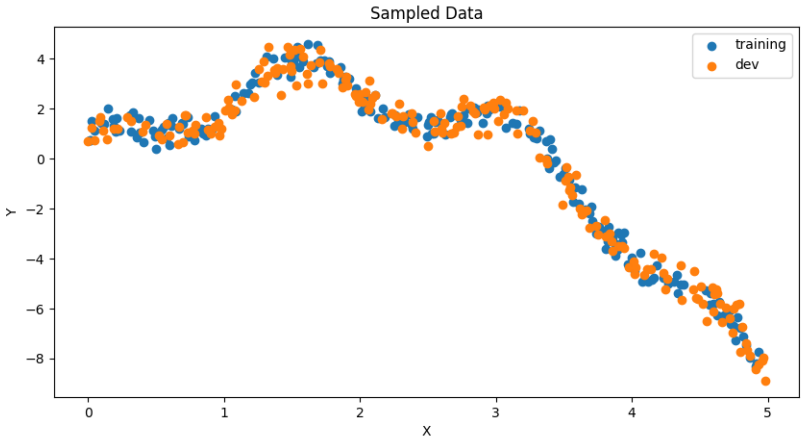

Figure 1: Univariate dataset

We now try fitting different orders of polynomials over the data, see how the fit-error changes, as shown in Fig 2. We observe that the errors for both training and development data hit a minima at M = 10. Thus, our optimal fit is a polynomial of order 10.

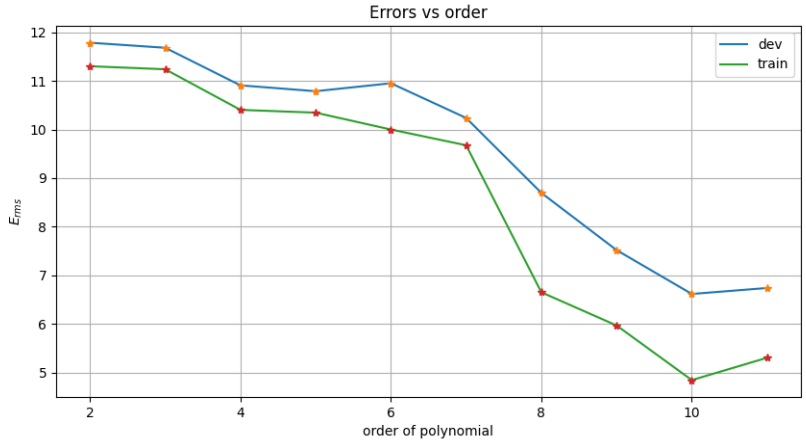

Figure 2: Error vs Order

With this M = 10, we perform Ridge regression. Similar to the previous case, we perform fitting over different λ values. From Fig 3, we see $\lambda = 10^{-5}$ as a regularization parameter with minimal errors for training data and development data. Though any value lesser than 10^{-5}, including 0 might work, we used $\lambda = 10^{-5}$ just for the joy of using Ridge Regression with the highest possible λ. The fitting errors can be seen in Table 1.

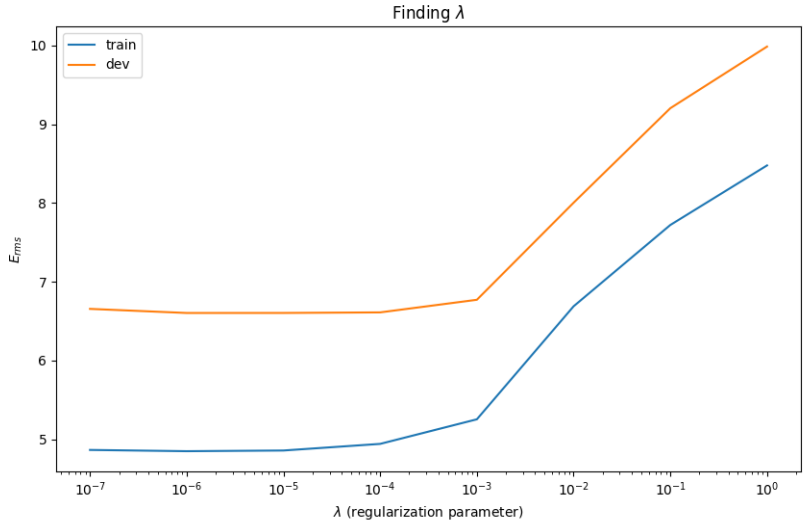

Figure 3: Error vs λ

Figures 4, 5, 6 represent the changes in fit for varying order, sample size and regularization parameters, respectively. The final model can be seen in Fig 7.

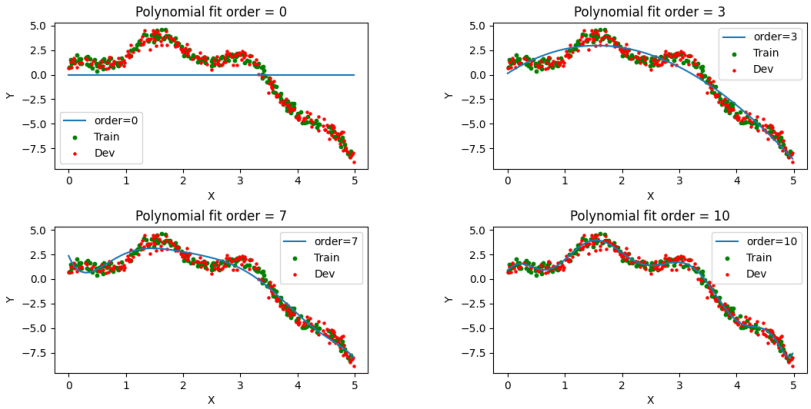

Figure 4: Curve fit with varying order of polynomial (M)

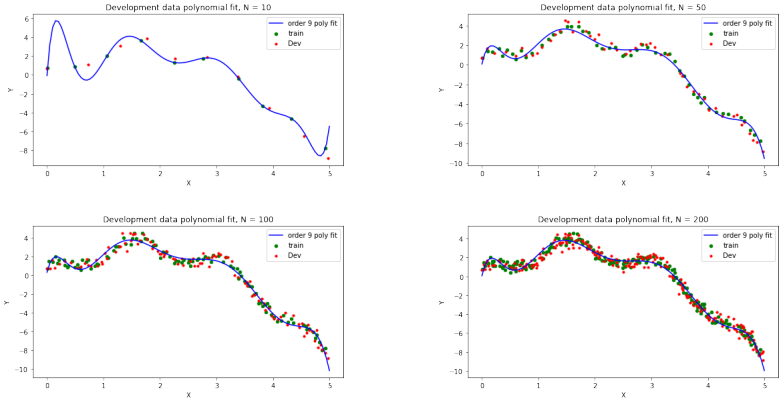

Figure 5: order 9 Curve fit with varying sample size (N)

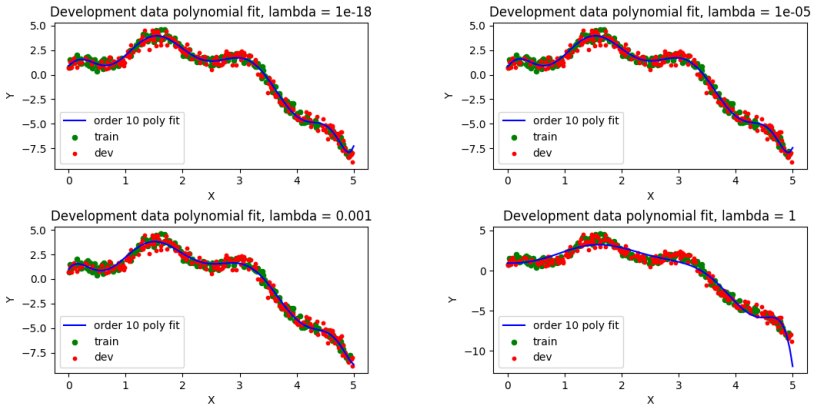

Figure 6: Order 10 Curve fit with varying regularization parameter (λ)

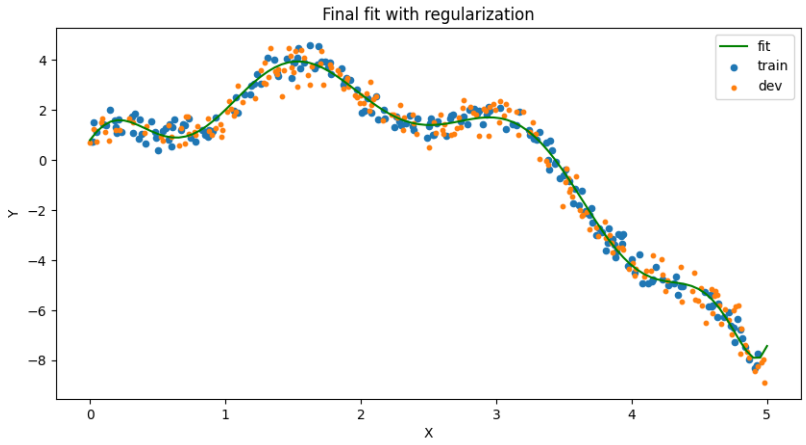

Figure 7: Best fitting model, with M = 10 and $\lambda = 10^{-5}$

Univariate Data	
M = 10	$\lambda = 10^{-5}$
Training data fit Error	4.858247346091573
Development data fir Error	6.604624193040589

Table 1: 1D data: error values for train and development data

2.3 Bivariate Data

1000 points of training and development data each have been given. The x-axis and y-axis are the 2-D inputs (variables) and the z-axis is the output (response).

We now try fitting different orders of polynomials over the data, see how the fit-error changes, as shown in Fig 9. We observe that the errors for development data hit a minima at M = 11. Thus, our optimal fit is a polynomial of order 11. With this M = 11, we perform Ridge regression. Similar to the previous case, we perform fitting over different λ values. From Fig 10, we see $\lambda = 10^{-6}$ as a regularization parameter with minimal errors for training data and development data. The fitting errors can be seen in Table 2. Thus, the final model can be seen in Fig 8

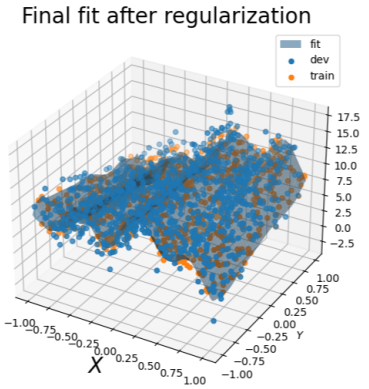

Figure 8: Bivariate dataset and Best fitting model, with M = 11 and $\lambda = 10^{-6}$

Bivariate Data	
M = 11	$\lambda = 10^{-6}$
Training data fit Error	26.494398653857456
Development data fir Error	59.62581849110795

Table 2: 2D data: error values for train and development data

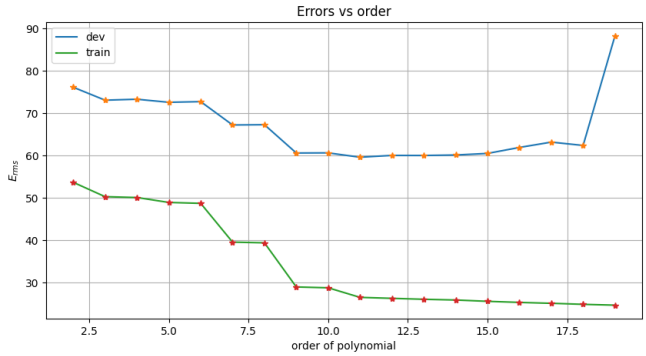

Figure 9: Error vs Order

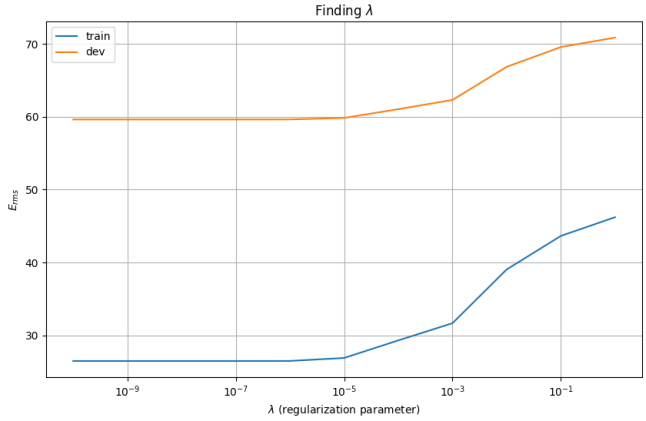

Figure 10: Error vs λ

3 Classification

3.1 Introduction

We are given 3 different 2D data sets each with training and development data:

1. Linear seperable data
2. Linear inseperable data
3. Real data

Each data set has 3 classes, over which a predictor is to be modelled. In this report, we build Bayesian classifiers with multivariate uni-modal Gaussian distributions for each data-set, with 5 different assumptions over the Gaussian covariance parameters:

1. Bayes with Σ same for all classesc
2. Bayes with Σ different for all classes
3. Naive Bayes with $\Sigma = \sigma^2 * I$.
4. Naive Bayes with Σ same for all classes.
5. Naive Bayes with Σ different for all classes.

The corresponding parameters for each of these 5 cases has been estimated using Maximum Likelihood Estimation, with the results as follows:

1. $\Sigma_i = \sum_{j=1}^{N_{class}} Cov(X_{train}[class = j]) * P(\omega_j) \forall i \in \{1, 2...N_{class}\}$
2. $\Sigma_i = Cov(X_{train}[class = i])$
3. $\Sigma_i = \sigma^2 I \forall i \in \{1, 2...N_{class}\} : \sigma^2 = mean(\sum_{i=1}^{3} Var(X_{train,j^{th}dim}[class = i]) * P(\omega_i))$
4. $\Sigma_i = diag(\sigma_1^2, \sigma_2^2 ... \sigma_d^2) \forall i \in \{1, 2..N_{class}\}$
   $: \sigma_j^2 = \sum_{i=1}^{N_class} Var(X_{train,j^{th}dim}[class = i]) * P(\omega_i) \forall j \in \{1, 2, ...d\}$
5. $\Sigma_i = diag(\sigma_{i1}^2, \sigma_{i2}^2 ... \sigma_{id}^2) : \sigma_{ij}^2 = Var(X_{train,j^{th}dim})[class = i] \forall j \in \{1, 2, ..d\}, i \in \{1, 2...N_{class}\}$

It can be seen graphically and shown analytically that the decision boundaries for cases 1, 3 and 4, where all the classes are modelled with same co-variance, give linear boundaries while cases 2, 5 give non-linear decision boundaries (conic sections) as in 12 and 11 respectively.

3.2 Linear Seperable Data

All the 5 cases give perfect predictions for the training and development data sets. The confusion matrices are perfectly diagonal, the ROC curves are a perfect Γ in shape with AUC = 1 and the DET curves are virtually invisible. The only interesting observation would be the decision boundaries for case 2 assumption where all the boundaries are hyperbolic, as shown in Figure 11. The PDFs modelled by case 3 can seen in Fig 20.

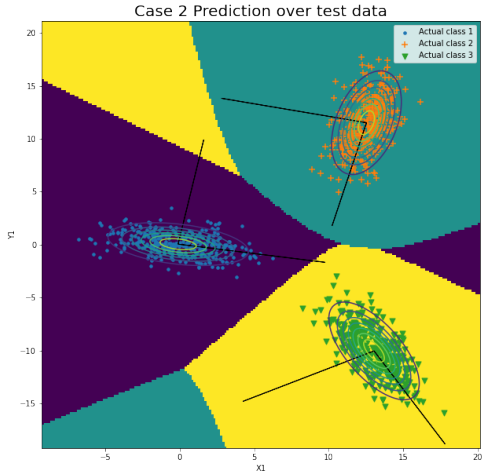

Figure 11: Decision boundaries, Case 2, Linear Seperable Data

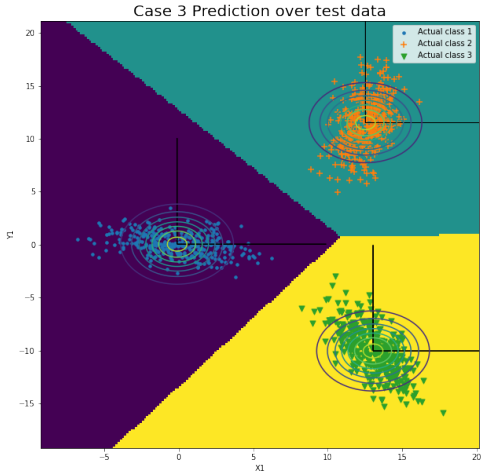

Figure 12: Decision boundaries, Case 3, Linear Seperable Data

3.3 Linear Inseperable Data

This is an interesting case wherein the data for class 1 and class 3 are scattered in 2 cross cutting clusters as shown in Figure 13 , none of which seem to be modelled with a uni-modal Gaussian distribution. The first idea would be either to involve bi-modal Gaussian distribution or use some transformation over the data space. But experimenting with cases 2 and 5 give a very high accuracy, with cross cutting boundaries, as seen in Figure 14. The decison boundaries for case 1 has been shown in Fig 15 for cross referencing. Note that cases 3 and 4 give boundaries very similar to it. The ROC curve (Fig 16) and DET curves (Fig 17) agree with this, showing case 2 and 5 performing much better than the rest.

Figure 13: Linear Inseperable Data

Here, as discussed earlier, we have hyperbolic/pair of lines as the decision boundary between classes 1 and 3. This shows that though the data is not distributed anything like a Gaussian, but if the viable decision boundary can be modelled as a conic section, then Gaussian modelling based Bayesian classifier can work. The PDFs modelled by case 2, which is the best case according to the ROC/DET, can seen in Fig 21.

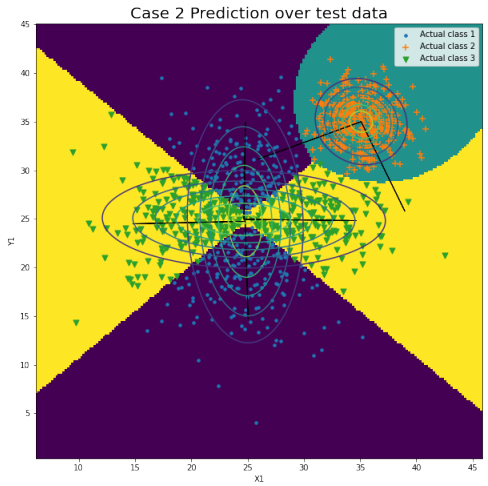

Figure 14: Linear Inseperable Data modelled by Case 2

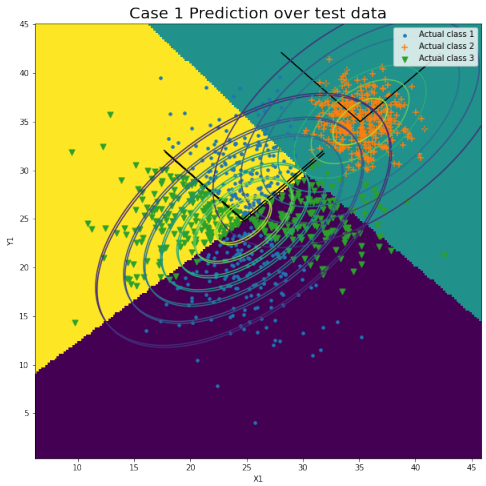

Figure 15: Linear Inseperable Data modelled by Case 1

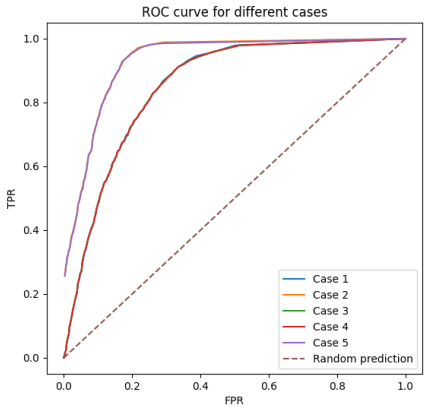

Figure 16: Linear Inseperable Data ROCs

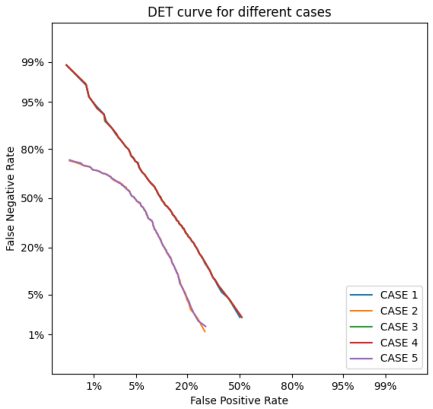

Figure 17: Linear Inseperable Data DETs

3.4 Real Data

The 3 cases in the real data-sets seem visibly divided into 3 slightly intermixing clusters with few points way away from the said clusters. These "leaks" from cluster decrease the accuracy to ∼80% as seen from the confusion matrices in figures 18 and 19. These confusion matrices are drawn from case 3 based model, which by a very narrow margin, happens to be the best model for the data, as interpreted form the ROC/DET curves. The PDFs modelled by case 3, which is the best case according to the ROC/DET (Fig 16,17), can seen in Fig 22.

Figure 18: Real dataset training confusion matrix for case 3

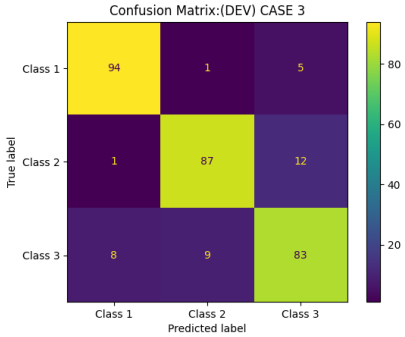

Figure 19: Real dataset training confusion matrix for case 3

3.5 The Best PDFs

Figure 20: PDF of case 3 model for linear seperable data

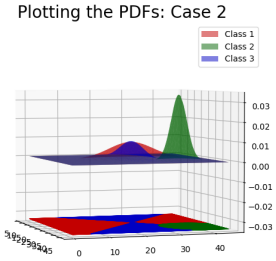

Figure 21: PDF of case 2 model for linear inseperable data

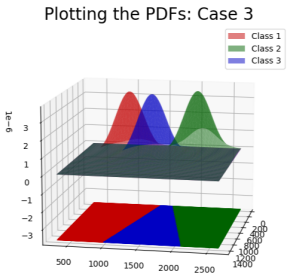

Figure 22: PDF of case 3 model for real data