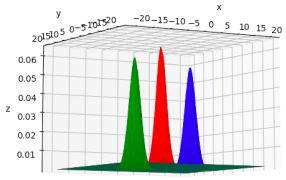
Bayesian Classifier

Contents

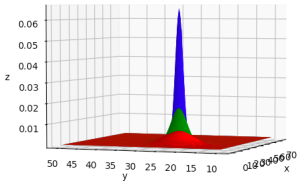
1 Bayesian Classifier

1.1 Probability Density and Constant Density Curves

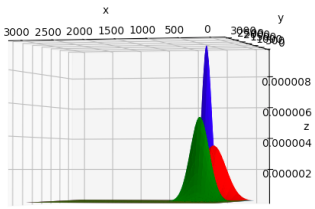
From the contour and probability density plots, fig 1, it is clear that whether the data is linearly separable or not. The axis of the contours plots are eigen vectors of the covariance matrix and they are orthogonal, because the covariance matrix is symmetric.



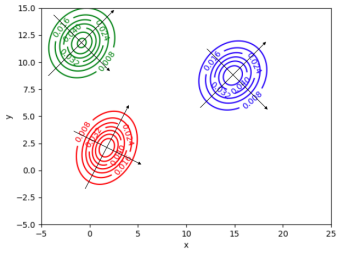
(a) Linearly Separable (pdf)



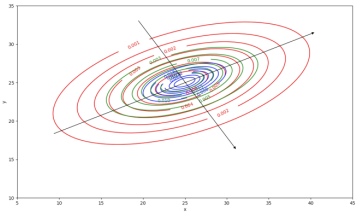
(b) Non Linearly Separable (pdf)



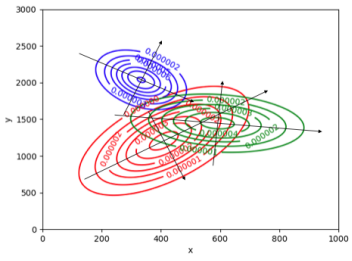
(c) Real Data (pdf)



(d) Linearly Separable

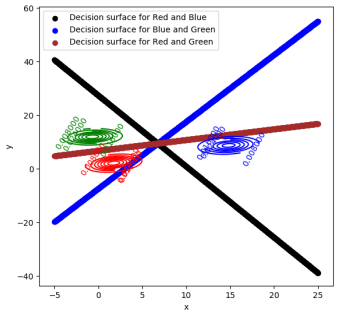


(e) Non Linearly Separable

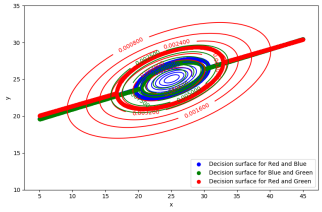


(f) Real Data

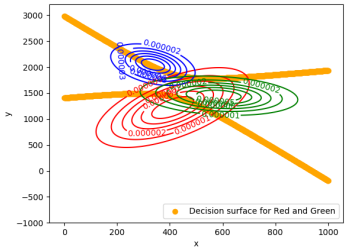
Figure 1: Probability Density and Constant Density curves



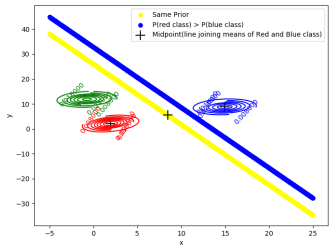
(a) Linearly Separable



(b) Non Linearly Separable



(c) Real Data



(d) Linerly separable - unequal prior

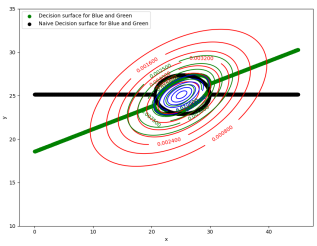
Figure 2: Boundaries with equal prior 2(a) 2(b) 2(c) and unequal prior 2(d)

.

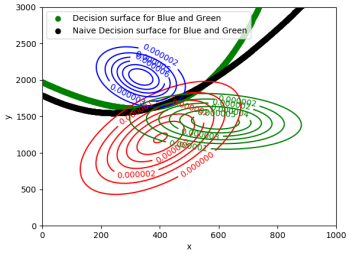
1.2 Decision Boundary and Surface

Decision boundary discriminates the data. Objective of the classifier is to find the best decision boundary. The decision boundary between two classes is calculated by equating the discriminant of the classes. As indicated in the plots, fig 2, using bayesian classifer with same full covariance or σ^2Icovariance matrix for all classes gives a linear decision boundary. Same covariance matrix is found by taking average of all the covariance matrices of the different classes. From the plots, we can infer that bayesian model with the mentioned covariance settings performs best on linear separable data and badly on real and non linearly separable data. Bayesian classifier with unequal covariance matrices gives non-linear decision boundaries.

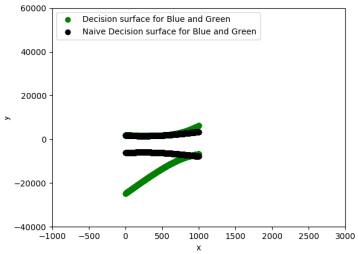
When the **priors are equal** the bayesian classifier with same covariance matrix passes through the mid point of the mean of the two classes but it is not perpendicular to the line joining priors due to rotation given by Σ^{-1}. When priors are different, decision boundary moves in the direction of less prior class, thereby increasing the space of region belonging to more prior class.



(a) Tilt not captured in Naive Bayes



(b) Smooth curve of Naive Bayes (Real data)



(c) Flatter Hyperbolas (Real Data)

Figure 3: Nature of curve in Naive Bayes (black) vs Bayesian with Full Σ (green)

.

1.3 Naive Bayes

Naive Bayes classifier considers covariances of different random variables to be zero. In σ^2I case, σ^2 is calculated by averaging over all diagonal elements of different classes. In same Σ case, Σ is calculated by taking average of Σ of different classes. In different Σ, we just use the given Σ for different classes. In all the 3 cases, we set non-diagonal elements to be zero.

Due to independence of features assumption, decision boundary generated will not capture the tilt in data points. Look at the plots of Bayesian and Naive Bayes different Σ classifier on Non Linearly separable data, fig 3(a). Naive Bayes classifier is not tilted. When tested on real data, naive bayes classifier with full Σ is much smoother than bayesian classifier with full Σ, fig 3(b). Zoomed out view shows smooth hyperbolas for naive bayes case as compared to bayesian case, fig 3(c).

One interesting observation for naive bayes classifier with σ^2I on linear data set was decision boundary moved towards perpendicular bisector of means of 2 classes from the position when we considered full Σ bayesian classifier. Full Σ

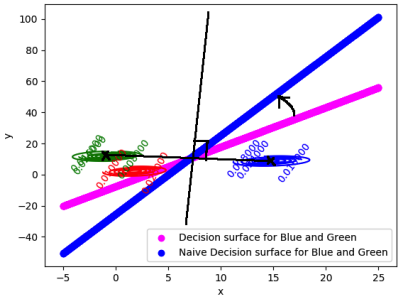


Figure 4: Naive Bayes σ^2 decision boundary rotated

passes through the midpoint of means but it might rotate line joining the means when there exists a λ different from 1. Ideal case of data having covariance matrix of σ^2I gives a perpendicular bisector as decision boundary. Naive bayes is expected to lie between above 2 decision boundaries which is confirmed by experimentation.

1.4 Confusion Matrix

Confusion matrix is used to visualize the performance of the classifier. The confusion matrices with different covariance settings are shown in Table 1, 2 and 3. Since the covariance matrices are same, the decision boundary will be a line. So for linearly separable data this setting will perform very well. But when we are making a classifer for non linearly separable data with same covariance matrix, performance will be poor in comparison with different covariance matrices. For linearly separable data, bayesian classifier with same covariance matrix gives perfect classification.

Input	**Prediction**		
	C1	**C2**	**C3**
C1	82	69	54
C2	62	80	64
C3	65	69	55

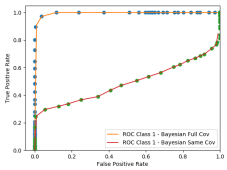
Table 1: Non Linearly Separable - same Σ

Input	**Prediction**		
	C1	**C2**	**C3**
C1	165	0	38
C2	0	172	33
C3	0	0	192

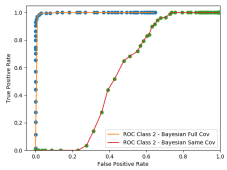
Table 2: Non Linearly Separable - Different Σ

Input	**Prediction**		
	C1	**C2**	**C3**
C1	96	0	0
C2	0	107	0
C3	0	0	97

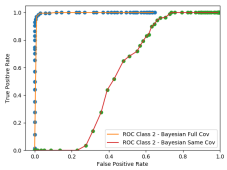
Table 3: Linearly Separable - Same Σ



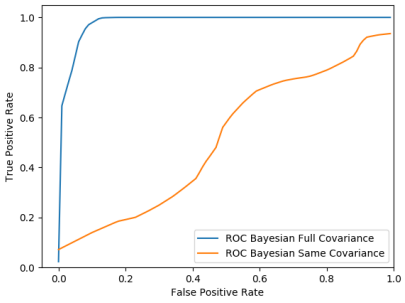
(a) 1 vs 2,3 ROC Curve



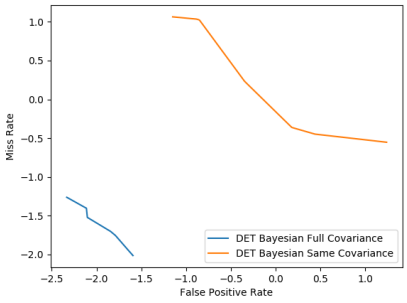
(b) 2 vs 1,3 ROC Curve



(c) 3 vs 1,2 ROC Curve



(d) Average ROC Curve



(e) Average DET Curve

Figure 5: ROC and DET

1.5 Receiver Opearting Characteristic and Detection Error Tradeoff

ROC and DET curves are used to compare performance of different classifiers.

ROC curve for non linearly separable data with same covariance and different covariance matrix are plotted by finding ROC for 1 vs rest and averaging true positive rate by interpolating along false positive, fig 5. As expected, having same Σ gives a linear decision boundary which performs poorly compared to having different Σ which gives non linear decision boundary performing well. ROC curve peaks up to 1 quickly for bayesian different Σ classifier.

DET curve is plotted after scaling miss rate ratio and false positive ratio by **inverse standard normal** to get a linear plot. Similar to ROC result, DET is averaged across all 1 vs rest.