GAUSSIAN CLASSIFIERS

Assuming we don’t normalize training and evaluation data, testing Gaussian, NaiveBayes and Tied Covariance Classifier it highlights Naive Bayes and Gaussian classifiers perform worst w.r.t. Tied Covariance Classifier on raw data. We used for our first evaluations a single fold dataset with different pre-processing:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Raw Data** | **Gaussianized Data** | **Z-Normalized Data** |
| ***Gaussian Classifier*** | 0.041 | 0.083 | 0.043 |
| ***Naive Bayes*** | 0.063 | 0.058 | 0.063 |
| ***Tied Covariance*** | 0.024 | 0.060 | 0.024 |

Evaluating the small (or absent) improvements given by Gaussianization and Z-normalization, we decided to skip these steps of pre-processing. Sensitivity and specificity were computed because after consulting previous confusion matrix we found our dataset was unbalanced. Then we evaluated DCF on different applications:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **(0.5, 1, 1)** | | **(0.1, 1, ,1)** | | **(0.9, 1, 1)** | |
| **Error | DCF**  **Sensitivity | Specificity** | | | | | |
| ***Gaussian Classifier*** | 0.041 | 0.165 | 0.033 | 0.342 | 0.0470 | 0.925 |
| 0.865 | 0.969 | 0.847 | 0.979 | 0.902 | 0.958 |
| ***Naive Bayes*** | 0.063 | 0.194 | 0.054 | 0.552 | 0.074 | 1.214 |
| 0.862 | 0.944 | 0.847 | 0.956 | 0.873 | 0.931 |
| ***Tied Covariance*** | 0.024 | 0.199 | 0.027 | 0.287 | 0.024 | 1.514 |
| 0.807 | 0.993 | 0.760 | 0.995 | 0.833 | 0.991 |

Evaluating results we can see our balanced application has a low TPR compared to the TNR and a low actDCF. Our unbalanced applications performed better increasing our TPR. We can see our Gaussian Classifier in an unbalanced application improved TPR without increasing too much the error rate but DCF is high. Application (0.1, 1, 1) has a low DCF but a lower sensitivity w.r.t balanced application.

LOGISTIC REGRESSION

We decided also to implement a Logistic Regression classification method. After we gave to the algorithm raw data, we obtained overflow errors which didn’t allow us to proceed with analysis. So we transformed our data using Z-normalization and gaussianization. After running the model on both transformed dataset, we deduced Z-normalized data give slightly better results. We used for our first evaluations a single fold dataset:

K-FOLD

We decided to implement the k-fold cross validation in order to have more evaluation metrics. We measured DCF, minDCF, error, sensitivity and specificity, all this to have a complete overview of the goodness of our model.

* Gaussian classifiers: we applied this approach only on the Gaussian Classifier and on the the Tied Covariance Classifier just because they give us apparently better results in terms of cost and error w.r.t. Gaussian Bayes Classifier (which give us worst results). From the results we found unbalanced applications give us better results overall because we think, in this case, a higher TPR is better than a low overall error.
* Logistic regression: