ML and PR Project Report – Part 10

The final step to conclude the analysis of our classifiers of the fingerprints dataset is considering a pristine evaluation dataset, on which to do the final performance analysis of our choices.

# Evaluating the best model, i.e. uncalibrated 8-component diagonal GMM

The system we chose for delivery was the 8-component diagonal GMM model, without calibration, as the evaluation of the calibrated scores yielded results that were just slightly worse than the non-calibrated ones for both the target application and other log-odds.

The evaluation on this final evaluation dataset yields the following results.

Immagine che contiene testo, Carattere, linea, numero

Descrizione generata automaticamenteImmagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

The application does not perform as well as it did on the validation set, but still yields good results. The calibration for the target application is not bad, but the difference between minimum and actual DCF for the validation set was of just 0.005 points, while here we get to almost 0.020, which is a value similar to what we had for the calibrated version of GMM, in the previous report.

Scores seem to remain well calibrated for lower-magnitude log-odds, but exhibit a not so ideal behaviour at higher-magnitude points.

At this point, I think it could prove useful seeing what would have happened if we chose the calibrated GMM model instead of the raw one, as their performance didn’t differ all that much.

Immagine che contiene testo, Carattere, schermata, linea

Descrizione generata automaticamenteImmagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

As we can see, the target application is almost perfectly calibrated, and this quality seems to hold even for growing values in magnitude of the log-odds.

Thus, the calibrated 8-component diagonal GMM would have been a better choice.

# Fusion of the three best systems

The fusion of the systems has been performed starting from the uncalibrated scores, in order not to calibrate the data twice, and the systems used are the three that exhibited better performance among all considered, namely Quadratic Logistic Regression, RBF Kernel SVM, and diagonal GMM. More details about each model will be added in the next paragraph.

In the following, we will compare the evaluation of the best models seen so far and of the aforementioned fusion. The models used are the following:

* **Quadratic Logistic Regression** with , calibrated using prior-weighted Logistic Regression with and K-fold using
* **RBF Kernel SVM** with and , calibrated using prior-weighted Logistic Regression with and K-fold using
* **Uncalibrated 8-component diagonal GMM** - uncalibrated

The results are the following.

Immagine che contiene testo, Carattere, schermata, numero

Descrizione generata automaticamente

Immagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamente

When looking at the actual DCFs our target application, we can see that the Fusion model actually performs slightly better than the delivered one, while the performance of the other models ranks as we expected from what has been seen in the previous report.

In terms of other applications, there is not much difference between GMM and Fusion, except they slightly outperform each other at times.

During the previous report, we actually had analyzed a similar thing, except the one model that was just slightly better was the uncalibrated GMM, while now it’s the Fusion that wins, although just barely.

If we get back to the observations done in the first section of this Report, we may notice that the calibrated GMM is actually better for the evaluation dataset, and it would have beaten the Fusion model, although, again, just barely.

Talking about LR and SVM instead, we can see that the scores seem to be well calibrated – although LR in the previous report exhibited a gap between minimum and actual DCF that was exactly 0, while now we have a 0.014 – but the overall performance has drastically worsened, when tested on this pristine evaluation set, as both had an increase, in both minimum and actual DCFs, of about 0.100 points.

# Going back to other models (only for LR)

Let us now see if the choices we have done for a single one of the three best models may have actually proven detrimental, and maybe some other version altogether, or some other hyper-parameter tuning choice, may result in better performance for the current evaluation set.

We are only going to tackle Logistic Regression in order to save time, and for that same reason, we are not going to calibrate the scores, meaning that we will only compare the performances based on the minimum DCF.

The results of the **Quadratic Logistic Regression** with , calibrated using prior-weighted Logistic Regression with and K-fold using , can be seen in the previous images: for general applications we can look at the green curves in the Bayes Error Plot, while for our target application, we have calculated a minimum DCF of 0.353.

The Logistic Regression models that we have analyzed during the whole project are mainly:

* prior-weighted Logistic Regression
* non-weighted Logistic Regression with pre-centering
* non-weighted Quadratic Logistic Regression

Immagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamenteImmagine che contiene testo, linea, diagramma, Diagramma

Descrizione generata automaticamenteImmagine che contiene testo, linea, Diagramma, diagramma

Descrizione generata automaticamente

We can easily see that the linear models perform horribly, with minimum DCFs above the 0.5 threshold.

Slightly different choices for would have instead given slightly better results, as we can see from the curve decreasing, although just a little, after the middle point between and on the x-axis (i.e. the point at which ). Anyway, we would have obtained, at best, minimum DCFs of 0.347 and 0.351 for and respectively, meaning a increase in performance of just 0.006 DCF for the first .

# Conclusion

The performance on the evaluation set is worse than that expected from the validations, although the rankings don’t differ much, i.e., during validation, the choices we had taken weren’t all that resolute, as the performance of Fusion, GMM with calibration, and GMM without calibration, and now, on the evaluation set, we observe the same.

As for the choices of the single systems, at least for Logistic Regression, we can assert that we created a mostly optimal model, at least if we account for minimum DCF.