Basket Classifier: Fast and Optimal Restructuring of the Classifier for Differing Train and Target Samples

Anton Philippov, Fedor Ratnikov

HSE University, Moscow, Russia

May 19, 2021

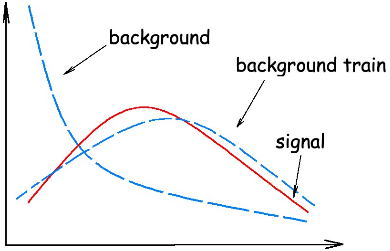
Overview

1. Motivation
2. Possible solution
3. Problem-1
4. Problem-2
5. Procedure
6. Toy example
7. Distributions
8. Classifiers
9. Scores
10. Conclusions

Motivation

Consider a problem of separating π^0 from photons using machine learning algorithms. There is a complication in the problem: the calibration sample is quite different from the real background. Let's formulate it in a more general way: there are two obvious problems of constructing classifier in the case of continuous spectrum:

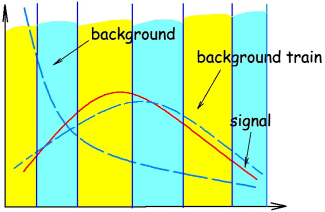
- We want to avoid dependence on the training sample
- We want to train a classifier on the training sample only once, avoiding this procedure in the future when changing spectra



Possible solution

What can we do to avoid these difficulties?

- Let's divide the spectrum into a number of baskets, for each of which we build its own classifier, that maximizes the area under the ROC curve.
- We obtain tolerance to changes in the distribution due to a) their narrowness, b) tolerance of ROC AUC to imbalance classes.
- Now we can solve the problem of maximizing the signal level for a given background level; to do this, we need to select a cut-off threshold in each basket so that for a given amount of background events across all baskets, the sum of signal events is maximum, i.e. solve the optimization problem.



Problem-1

- m_i - number of noise events for the i-th basket,
- n_i - number of signal events for the i-th basket,
- ξ_i - fraction of noise events for the i-th basket,
- α - target signal efficiency,
- f_i - ROC-curve for the i-th basket
- N - number of baskets.

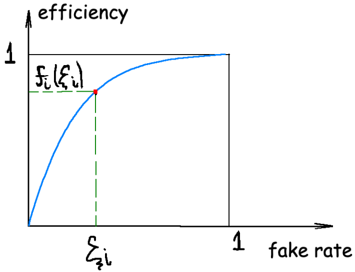
$$\min_{\xi} \quad \sum_{i=1}^{N} m_i \xi_i$$

$$\text{s.t.} \quad \frac{\sum_{i=1}^{N} f_i(\xi_i) n_i}{\sum_{i=1}^{N} n_i} = \alpha$$

Problem-2

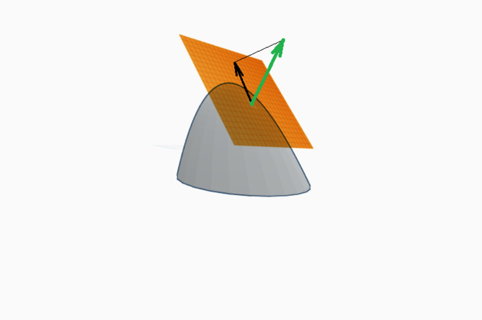
$$\min_{\xi} \quad \sum_{i=1}^{N} m_i \xi_i$$

$$\text{s.t.} \quad \frac{\sum_{i=1}^{N} f_i(\xi_i) n_i}{\sum_{i=1}^{N} n_i} = \alpha$$



Optimization procedure includes 2 steps:

- the projection of the gradient onto the tangent plane
- lowering the vector to the surface of constraints
- repeating the first 2 steps until convergence



Toy example

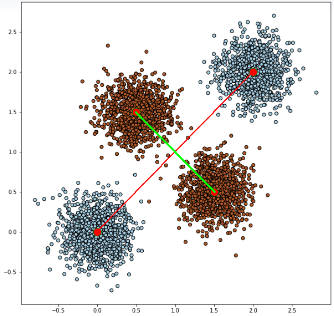
Parameterize 2 families of distributions:

- Distribution 1: two Gaussians with centers at $(-X, X)$ and $(X, -X)$ and fixed variance.
- Distribution 2: two Gaussians with centers at (Y, Y) and $(-Y, -Y)$ and fixed variance.

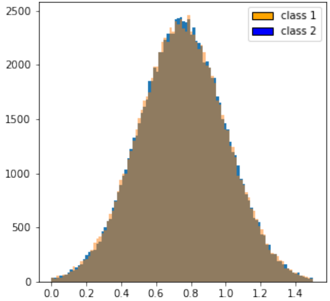
Our goal is to train the procedure on a given sample with parameters X_1 and Y_1 (sample A), and then apply the classifier to another sample with parameters X_2 and Y_2 (sample B), measuring score of the result.

Distributions

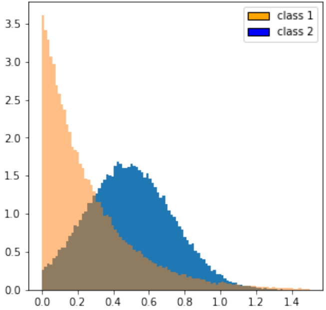
- ...



(a) Toy sample. X and Y define distances represented by red and green lines



(b) X and Y distributions for sample A



(c) X and Y distributions for sample B

Figure: Toy example distributions

Classifiers

- 4 classifier were built in the experiment. The first one was trained on a test sample (sample A), and tested on a target sample (sample B).
- The second one was trained on a target sample and tested on a target sample.
- The third and fourth (with 3 and 7 baskets, respectively) are basket classifiers, which were trained on a test sample and tested on a target sample.
- It is reasonable to expect that we will get the following ranking in terms of quality: the first classifier will demonstrate the worst quality, the second one - the best, and the basket classifiers will be located between them.

Scores

- Let's take a look at the results:

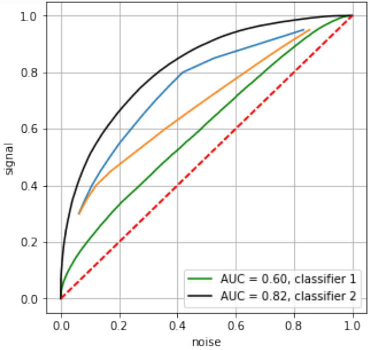


Figure: green line - efficiency for case (1), black line - efficiency for case (2), orange line - basket classifier with 3 baskets, blue - basket classifier with 7 baskets

The graph clearly shows that our assumptions were fully justified. The classifiers are indeed clearly ranked in terms of quality.

Conclusions

- In this paper, we present a concept of basket-based dynamic classifier.
- Such classifier demonstrates a tolerance to significant variations of the spectrum of the target analyzed data from data used for training.
- The procedure of fast adjustment of the basket-based classifier for a given analysis performance is also shown.
- In case of real experiments, such a basket-based classifier may be trained and validated only once in advance of data analyses. Further adjustments to real spectra of particular data analyses does not require re-training if using *a priori* knowledge of shapes of the target data sets.

The End