How to be wrong the right amount of time

Miloš Simić

University of Belgrade

Problem: Uncontrollable error rates

Traditional binary classifiers cannot control the class-conditional error rates. Sometimes, one of the errors is much more severe. For example, in medicine:

- A false negative misses a condition and puts health to risk.
- A false positive induces stress and requires further testing.

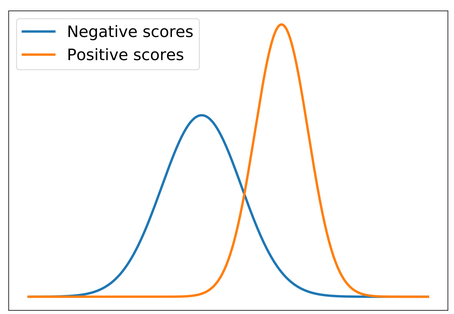
We see that a false negative is more severe. So, we need classifiers that give a false negative less than, say, 5% of time.

Solution: Treat the score as a statistic!

There is a way to control the already trained classifier's error rate over the chosen class (positive or negative).

- The idea is to **treat the classifier's score function as a statistic** S and determine its distributions over negative and positive objects.

Usually, the classifiers are such that the scores of positive objects tend to take higher values, and the scores of negative objects tend to take lower values.



Statistical tests of classification

We can create two statistical tests this way to test two hypotheses for each object we want to classify:

- H_0: The object is negative
- H_1: The object is positive

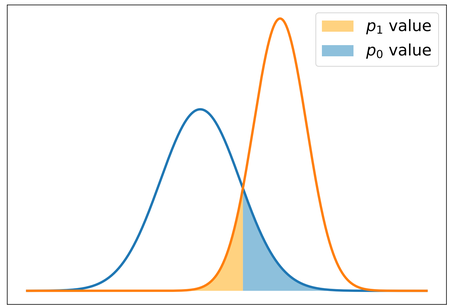
We can calculate the **p values** for both hypotheses.
By definition, the p value of a score is the probability that the statistic takes a value that is less compatible with the hypothesis.

The classification p values

Let s be the score of the object we want to classify.

- p_0: The p value that corresponds to H_0 is equal to $P(S \geq s \mid H_0)$ because higher scores are less compatible with the negative class.
- p_1: The p value that corresponds to H_1 is equal to $P(S \leq s \mid H_1)$ because lower scores are less compatible with the positive class.

The example in the figure below shows that the p values correspond to the areas under the score densities.
For better clarity, the score s for which the p values are calculated is taken to be the score of the intersection.



The approach in practice

- In practice, we determine which error rate we want to control. For example, the rate of false negatives.
- Then, we take a classifier that has already been trained the usual way and a number of scores of negative objects that weren't used during training $(s_1, \ldots, s_n)$.
- When we get a new object x, we calculate its score s using the classifier, and approximate its p_0 value as follows:

$$p_0(x) = P(S \geq s) \approx \frac{1}{n} \sum_{i=1}^{n} 1_{s_i \geq s} \qquad (1)$$

- If the p_0 value is lower than, say, 5%, then we reject H_0. Otherwise, we accept H_0.
- The long-run frequency of false negatives will be 5% if n is large enough.
- The same goes for false positives.

Conclusion

The classification p values and their estimators (such as (1)) are mathematically sound and have some nice properties.

- The p values are uniform over $[0, 1]$: this ensures that the chosen error rate can be set in advance to any value from $[0, 1]$.
- The estimators, such as the one in Equation (1), can very precisely estimate the true p value.
- The underlying classifier needs to be trained only once.
- The approach effectively recalibrates the underlying classifier without retraining it.
- However, the approach requires a possibly larger set of the objects belonging to the target class that were not used during training.
- A way to circumvent even this limitation is presented in the paper. See the reference below.

Contact and additional Information

- Author: Miloš Simić, milos.simic.csci@gmail.com
- Check out the paper to see the math and how the method was applied to neural networks and normality testing. **The obtained neural tests proved superior to the standard tests of normality.**
- Paper "**How to Control the Error Rates of Binary Classifiers**" is available at `https://arxiv.org/abs/2010.11039`.

University of Belgrade