CoConut: Co-Classification with Output Space Regularization

Sameh Khamis
sameh@umiacs.umd.edu
University of Maryland
College Park, MD 20740

Christoph H. Lampert
chl@ist.ac.at
IST Austria
Am Campus 1, 3400 Klosterneuburg

Classification is one of the most fundamental and best understood machine learning problems. Different scenarios differ strongly in their training procedure, but agree fundamentally in their prediction step at test time: each test sample is assigned a label individually. However, in many real-world the samples to be classified occur in batches, such as words in a document, images in a photo collection, or stocks in a portfolio, and exploiting this fact should make it possible to achieve increased classification accuracy.

To motivate our framework, consider the situation of a linear classifier, which is efficiently trainable and exhibits good generalization capabilities but has a decision hypersurface that might not perfectly reflect the class boundaries in feature space. Given sufficiently many test samples it should be possible to modulate the classifier's decision boundary, for example, based on the cluster assumption, which states that class decision boundaries typically do not cross high density regions (see Figure 1 for an illustration).

Despite its potential, the task of co-classification, *i.e.* classifying a set of points jointly, has received little attention in the literature. In this work, we introduce CoConut, a method for co-classification based on the established principle of regularized risk minimization. It jointly labels all test points by minimizing a regularized risk functional that that incorporates additional information in the output (label) space. CoConut only requires the output of a set of classifiers as input, but makes no assumption on how they were trained. It is also efficient, as it requires no additional training step but only solves a regularized risk functional using efficient energy minimization techniques.

We formalize the co-classification scenario in the following way. We are given a set of (test) examples, $X = \{x_1, \ldots, x_n\}$ from an input space $\mathcal{X}$, and we want to predict labels $Y = \{y_1, \ldots, y_n\}$ from a label set $\mathcal{Y} = \{1, \ldots, L\}$. For this task we have access to L fixed *base classifiers* with prediction functions, $f_1, \ldots, f_L : \mathcal{X} \to \mathbb{R}$, where for any $x \in \mathcal{X}$ and $l \in \mathcal{Y}$ the value $f_l(x)$ reflects a confidence that the sample x belongs to class l. The straight-forward choice for labeling the test points is then to predict (greedily) the most confident label for each sample, $y_i = \operatorname{argmax}_{l=1,\ldots,L} f_l(x_i)$.

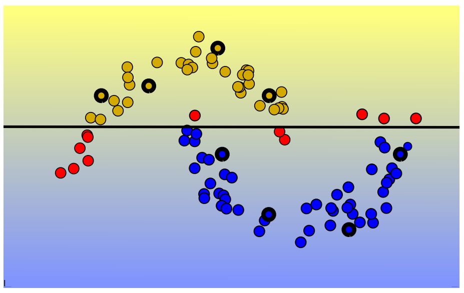
We propose to compute a joint labeling $y^* = (y_1^*, \ldots, y_n^*) \in \mathcal{Y}^n$ of the test points by solving the following optimization problem:

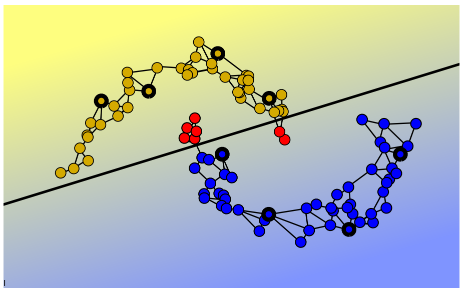
$$y^* = \operatorname*{argmin}_{y \in \mathcal{Y}^n} \; -\sum_{l=1}^{L} [\![y_i = l]\!] f_l(x_i) + \lambda \Omega(y), \tag{1}$$

where Ω is a regularizer that penalizes undesirable label combinations and $\lambda \in \mathbb{R}^+$ is a constant that controls the regularization strength. Note that for $\lambda \to 0$ we recover independent per-sample predictions, showing that per-example label selection can be thought of as a special case of this framework. Equation (1) resembles the expressions occurring in the classical framework of *regularized risk minimization* [1]. The difference lies in the fact that we regularize in the output space (the space of all labelings), not in the space of classifier parameters. Therefore, we call the resulting approach *Co-Classification with output space regularization* (CoConut).

In our choice of regularizer we encode the *inductive bias* we have about the problem. Often this would be an assumption that the true labels vary smoothly with respect to the inputs. For any point x_i, let $N_i \subset X$ be the set of neighbors that are similar to x_i. Let w_{ij} denote the a measure of the similarity between two neighbors x_i and x_j. For any $x_j \in N_i$ the slope of g between x_i and x_j is $w_{ij}\delta_{ij}(g)$, where $\delta_{ij}(g) := [\![g(x_i) \neq g(x_j)]\!]$ indicates whether g changes value between x_i and x_j. Averaging this quantity across all neighbors and all points, we obtain a measure for the average discontinuity (lack of smoothness) of any labeling function $g \in \mathcal{G}$:

$$\Omega_S(g) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|N_i|} \sum_{x_j \in N_i} w_{ij} \delta_{ij}(g). \tag{2}$$


(a) no cluster assumption (CA)


(b) CA at training time

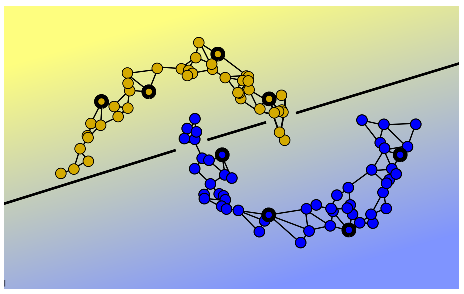

(c) CA at test time

Figure 1: Schematic illustration of the effect of the cluster assumption. Left: supervised training of a linear classifier with few training examples (bold circles): many mistakes occur at test time (red dots). Middle: cluster assumption during training reduces errors. Right: cluster assumption at test time reduces errors even further.

A regularizer can also encode a preference for a certain class label distribution at test time. This can counter the effect of the bias introduced by training with imbalanced class distributions. We assume that the target (expected) class label proportion for class l is Q_l, where $\sum_{l=1}^{L} Q_l = 1$. We define a measure for the disparity between the class label proportion $p_l(g)$ induced by labeling function g and the target proportion for each class l:

$$\Omega_D(g) = \sum_{l=1}^{L} |p_l(g) - Q_l| \tag{3}$$

where $p_l(g) = \frac{1}{n}\sum_{i=1}^{n} [\![g(x_i) = l]\!]$ are the label proportions the hypothesis g. The regularizer Ω_D penalizes the deviation from the target distribution, and this penalty is linear in the amount of deviation.

In the paper we discuss these regularizer choices, what information they can incorporate at test time, how they can be efficiently optimized, and their effect on the classification performance theoretically and empirically. We report our results using each regularizer on six different datasets, reporting consistent improvements over baselines.

[1] V. N. Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.