Logistic regression with pairwise constraints

Jacek Tabor, Marek Śmieja, Oleksandr Myronov

Jagiellonian University

February 14, 2017

Motivation

Not all datasets are perfect. Typical additional information in an imperfect dataset

- soft assignment - class probabilities
- pairwise constraints
- unlabeled data

Examples

- Human labeling can introduce both the soft assignment (inter-annotator disagreement)
- Frame sequence from the person tracker in human classification task. People in subsequent frames are probably the same person, 2 people in the same frame are probably different.

Demonstration

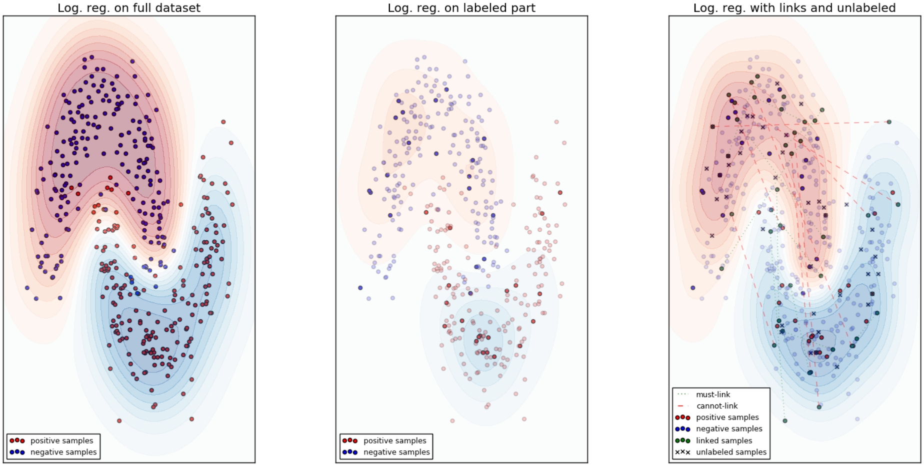


Figure 1: Logistic regression: vanilla vs links

Classification cost function - logistic regression

We have n classes which are categorical, and we want to estimate the probability that the point x belongs to a class. To do so, we search for vectors $v_1, \dots, v_{n-1} \in R^N$ (v_n can be reduced) so that

$$p_k(x) = p(k|x) = \frac{\exp(v_k \cdot x)}{1 + \sum_{l=1}^{n-1} \exp(v_l \cdot x)}, \text{ for } k = 1, \dots, n-1,$$

$$p_n(x) = p(n|x) = \frac{1}{1 + \sum_{l=1}^{n-1} \exp(v_l \cdot x)}.$$

For n classes $X_1, \dots, X_n \subset X$ the goal is to fix vectors v_k so that $p_k(x) = 1$ for all $x \in X_k$, $k = 1, \dots, n$. We maximize the log-likelihood, equivalent of maximizing

$$\prod_{k=1}^{n} \prod_{x \in X_k} p_k(x).$$

- pro: concave
- con: strong misclassification of a single point is not acceptable
- con: pairwise constraints may disrupt concavity

Classification cost function - labeled data

Expected probability of correct answers is given by:

$$\sum_{k=1}^{n} \sum_{x \in X_k} p_k(x). \tag{1}$$

Soft priors memberships, i.e. instead of giving hard class label $y_i \in \{1, \dots, k\}$ for a training point x_i, for $i = 1, \dots, t$, we assume that the point is assigned to classes according the probabilities $p_i = (p_i^{(1)}, \dots, p_i^{(n)})$, where $\sum_{k=1}^{n} p_i^{(k)} = 1$ and $p_i^{(k)}$ quantifies our believes (probability) that x_i belongs to k-th class. Objective function is the regularized expected misclassification probability,

$$\sum_{i=1}^{t} \|p_i - p(x_i)\|^2 = \sum_{i=1}^{t} \sum_{k} (p_i^{(k)} - p_k(x_i))^2$$

Minimization of the above function without the square is equivalent to maximization of (1). The square regularization was added to put even lower penalty for misclassification examples.

Classification cost function - pairwise constraints

Pairwise constrains define pairs $(x_i, x_j) \in X \times X$ that originate from the same class (must-links) or different classes (cannot-links).

Let p_{ij} be a prior probability that a pair (x_i, x_j) belongs to the same class. In consequence, x_i, x_j belong to different class with probability $1 - p_{ij}$, which represents the probability of cannot-link. The probability that the points x_i, x_j are in the same class is given by

$$must(x_i, x_j) = \sum_k p_k(x_i) p_k(x_j).$$

$$cannot(x_i, x_j) = 1 - must(x_i, x_j).$$

Learning model tries to find such vectors v_i, for $i = 1, \dots, n$, that the pairwise constraints are satisfied. To maximize the expected probability of correct answers, we arrive at the regularized expected probability loss function, given by:

$$\sum_{(i,j) \in C} (p_{ij} - \sum_{k=1}^{n} p_k(x_i) p_k(x_j))^2,$$

that have to be minimized.

Classification cost function - unlabeled

We are often given a lot of data points with no class information. In a typical semi-supervised setting, we take into account a cluster assumption, which states that the class labels do not change much in dense regions. We realize this assumption by assigning the most probable label to every unlabeled data point.
Given a set of unlabeled data x_i, for $i = t+1, \ldots, T$, we would like to maximize:

$$\sum_{i=t+1}^{T} \sum_{k=1}^{n} p_k(x_i)^2.$$

Equivalently, we focus on minimizing

$$\sum_{i=t+1}^{T} -\|p(x_i)\|^2 = \sum_{i=t+1}^{T} \sum_{k=1}^{n} -p_k(x_i)^2.$$

Classification cost function

The final cost function:

$$\begin{aligned} L = \; & \sum_{i=1}^{t} \sum_{k=1}^{n} (p_i^{(k)} - p_k(x_i))^2 \\ & + \beta \sum_{(i,j) \in C} (p_{ij} - \sum_{k=1}^{n} p_k(x_i) p_k(x_j))^2 \\ & + \delta \sum_{i=t+1}^{T} \sum_{k=1}^{n} - p_k(x_i))^2 \\ & + \alpha \|v\|^2, \end{aligned}$$

where β, δ, α are hyperparameters. To prevent from model overfitting, the objective function is supplied with the regularization term, which is a squared norm of vector $v = (v_1, \ldots, v_{n-1}))$

Experiments - starting weights

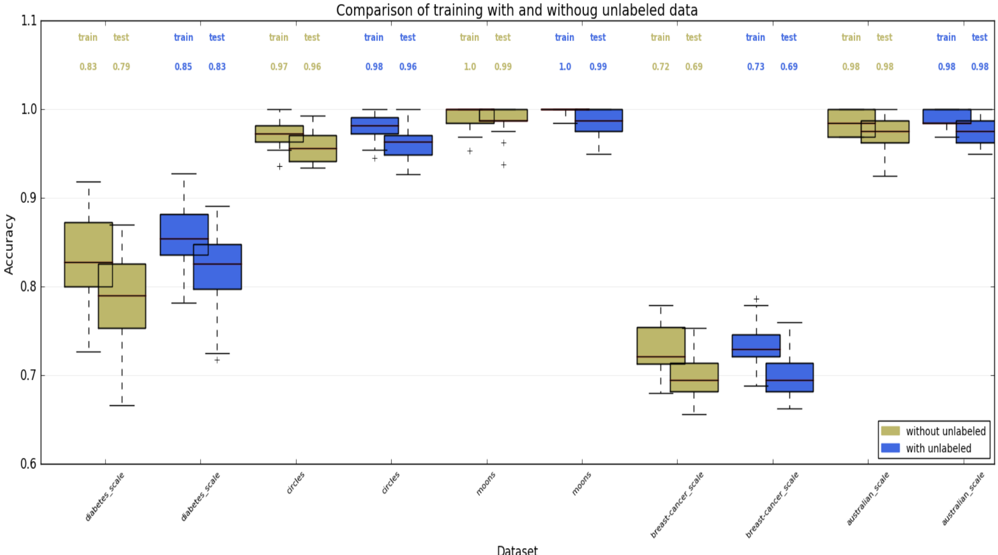
We use truncated Newton's algorithm for optimization.
The function is not convex.
Do starting weights matter?
We used some synthetic datasets and tested different initialization methods. We restarted the method 50 times and measured final obj. function loss. The table below shows the standard deviation of the loss for each method

dataset	circles	diabetes_scale	moons
method			
normal	4.548139e-08	6.684074e-09	7.496055e-09
normal_multivariate	1.652018e-08	7.265744e-09	5.570900e-09
normal_univariate	3.075394e-08	7.646498e-09	2.445021e-09
random_labels	9.307598e-09	2.440747e-09	6.107176e-09
random_links_diff	2.116533e-08	3.410780e-09	7.626612e-09
zeros	0.000000e+00	5.206897e-10	0.000000e+00

Experiments - unlabeled data

In order to validate our assumption that using unlabeled data helps, we trained and tested in double cross-validation 2 models. One was given the 20% of labeled data and 20% of links, the other was given also 20% of unlabeled data. The first model tuned it's α and β, as well as kernel γ, the second additionally tuned δ.

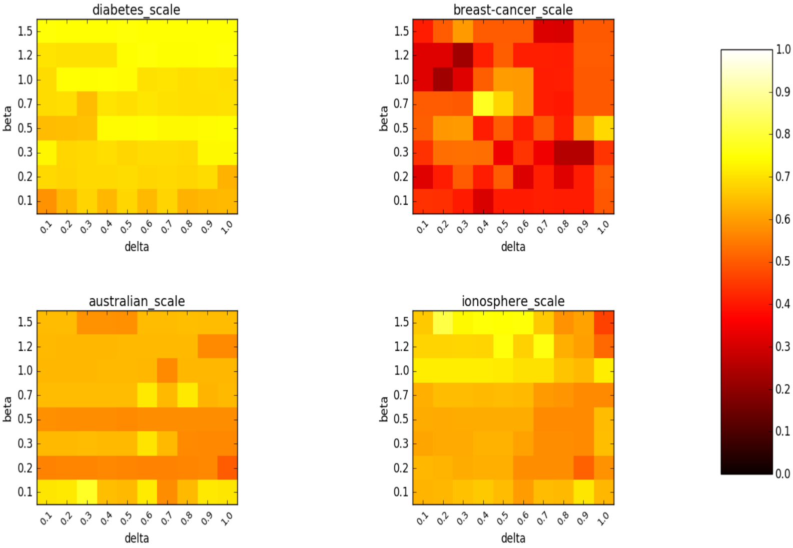
Experiments - unlabeled data



Experiments - hyperparameters

In order to see how much β (links weight) and δ (unlabeled weight) affects the model's performance, we did the following experiment. For several datasets, we fixed β and δ and performed a double cross-validation on the dataset, optimizing kernel γ and α hyperparameters. For the sake of simplicity, we assume breast cancer performance as the ground truth: β=0.7, δ=0.4 and use these values in further experiments.

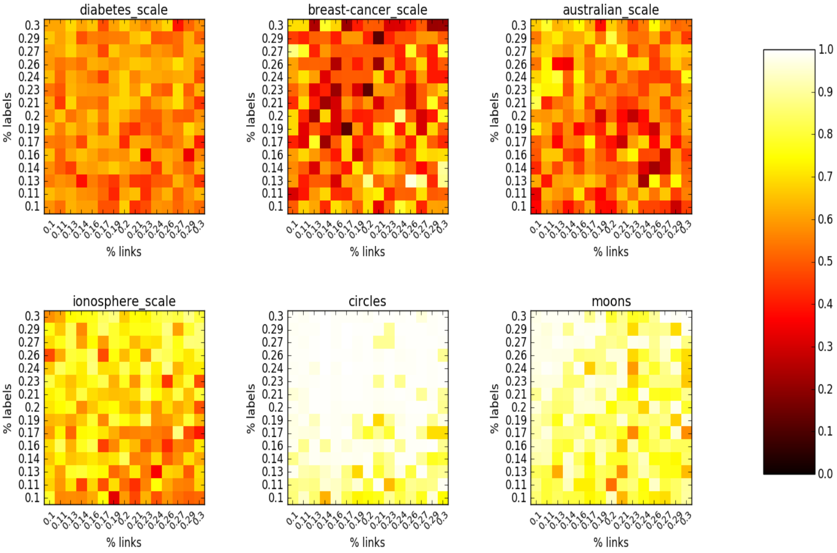
Experiments - hyperparameters



Experiments - percent of labels and links

In order to see how much information the models gets from labels and links, we concluded another experiment. For several datasets, we sampled different % of links and labels, performed grid search (α, γ) with cross-validation over the reduced dataset - labels, links and unlabeled data - then tested the model on a separate test set. The results below show different mean test scores for different %.

Experiments - percent of labels and links


diabetes_scale
breast-cancer_scale
australian_scale
ionosphere_scale
circles
moons
% labels
% links