Feature absence regularization for domain adaptive learning

Wouter Kouw, Laurens van der Maaten

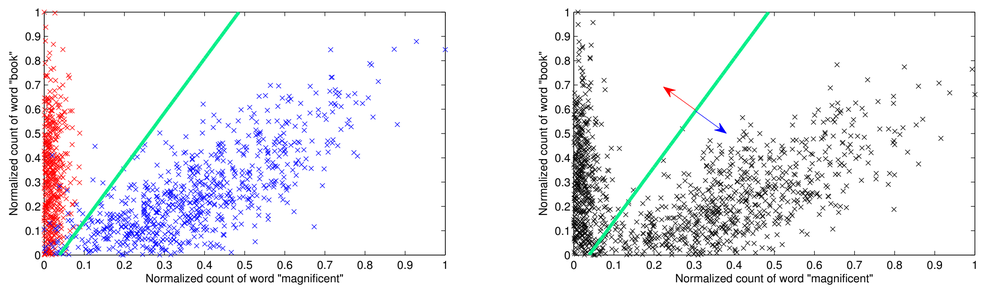
@W.M.Kouw@tudelft.nl, Pattern Recognition & Bioinformatics, TU Delft

Pattern Recognition

In pattern recognition, a sample is a vector of features of a particular object with a discrete label. By studying the relationship between the sample $\mathbf{x}$ and the label y, we can build systems that assign labels to novel samples.

I tried reading this book but found it so turgid and poorly written that I put it down in frustration...

This book is magnificent. I found myself so encrossed in the story. I am just. amazed at how ...



Problem

Domain adaptation is a pattern recognition problem setting where the test dataset (target domain) comes from a different distribution than the training dataset (source domain). The goal in this setting is to minimize the classification error on the target domain by making use of the unlabeled target data to adapt the source domain classification function.

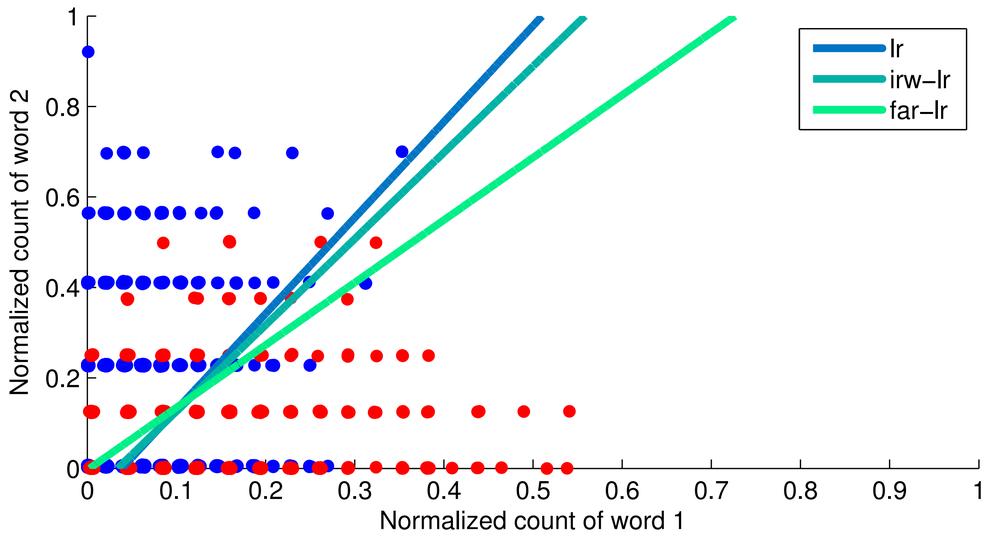


Figure 1: Scatterplot of two features of the Amazon sentiment dataset for the source domain. The three lines shown are the classification functions of a standard logistic regression, instance reweighted logistic regression and a feature regularized logistic regression function.

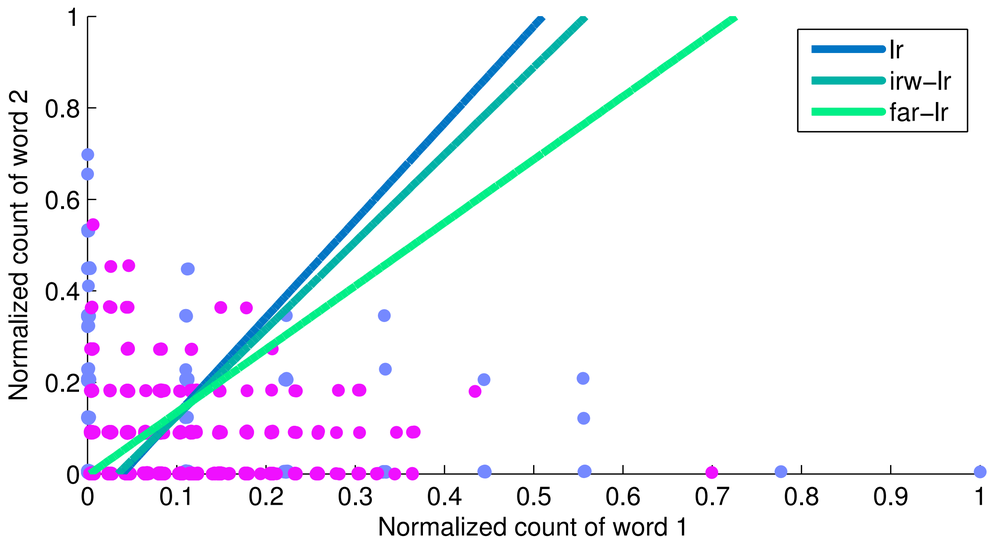


Figure 2: Scatterplot of the same two features of the Amazon sentiment dataset for the target domain. The three lines shown are the classification functions of a standard logistic regression, instance reweighted logistic regression and a feature regularized logistic regression function.

Classification errors	Source	Target
LR	0.3090	0.2014
Instance Reweighted LR	0.3104	0.2014
Feature Absence Regularized LR	0.4012	0.1597

Table 1: Table of classification errors of all 3 systems on source and target domains. Note that FAR makes less errors on the target domain at the cost of more errors on the source domain.

Data-dependent regularization

We propose a system that finds a solution adapted to the relative absence of features in the target domain. For M features of a sample, the value is either absent with probability q_m (here imputed as 0) or present with probability $(1 - q_m)$. The result of taking the expectation of this Bernouilli distribution over the samples is a data-dependent regularization term. This term ensures that the system finds a different minimal solution. The loss function we use is the likelihood of a logistic regression model:

$$L(\mathbf{X}, \mathbf{y} \,|\mathbf{w}, \mathbf{q}) = \sum_i^N \mathbb{E}_{\tilde{p}} \left[\frac{\exp(y_i \mathbf{w}^\top \tilde{\mathbf{x}}_i)}{\sum_j^K \exp(-y_j \mathbf{w}^\top \tilde{\mathbf{x}}_i)} \right]$$

where $\tilde{p}$ is the probability of feature absence:
$\Pr(\tilde{\mathbf{x}}_{im} = 0) = q_m$ and $\Pr(\tilde{x}_{im} = \frac{x_{im}}{1-q_m}) = 1 - q_m$.

Comparison with instance reweighting

In instance reweighting, the likelihood of each sample in the source domain is weighted according to its' probability under the target domain:

$$L(\mathbf{X}, \mathbf{y}|\mathbf{w}) = \sum_i^N \frac{p_T(\mathbf{x}_i)}{p_S(\mathbf{x}_i)} \left[\frac{\exp(y_i \mathbf{w}^\top \mathbf{x}_i)}{\sum_j^K \exp(-y_j \mathbf{w}^\top \mathbf{x}_i)} \right]$$

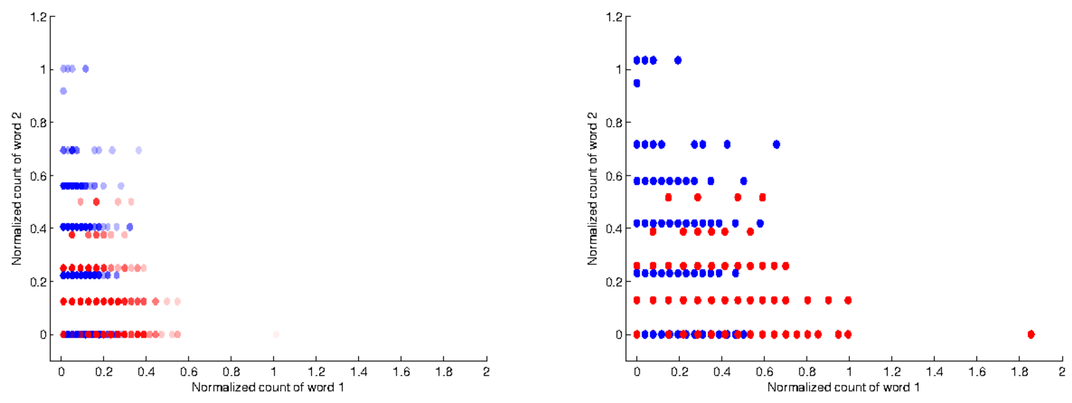


Figure 3: A) The likelihood of each sample in the source problem is reweighted based on its' probability under the target distribution, as illustrated here by the transparency of the samples. B) Features in the source problem are indirectly scaled according to their probability of being absent in the target problem: $q_1 = 0.4607$ and $q_2 = 0.3309$.

Discussion

We have shown a proof of concept for our method. The following open research questions relate to future work:

- Can we do the same for other transfer distributions (e.g. Gaussian, Laplacian, Poission, empirical)?
- Can we incorporate biased transfer distributions?
- Instead of estimating the transfer parameters, can we learn them instead?

TU Delft Delft University of Technology