# RSSL: Semi-Supervised Learning in R

Jesse H. Krijthe

Pattern Recognition Laboratory
Delft University of Technology, The Netherlands
Department of Molecular Epidemiology
Leiden University Medical Center, The Netherlands

## What is Semi-Supervised Learning?

In some prediction tasks, labeling data to be used for training a model is a relatively expensive process. Unlabeled data, on the other hand, may be easy to obtain. Semi-supervised learning is about using these unlabeled examples to improve supervised learning methods, which generally require labeled examples for training. Applications of semi-supervised learning include document and image classification, where documents and images are easy to obtain online in large volumes, while labeling all of them would be time-consuming. In other applications in, for instance, biology, ground truth labels may require expensive wet-lab experiments. In these and other applications, it would be very useful if unlabeled data could improve model estimation or selection.

## Approaches

The goal of the R Semi-Supervised Learning (RSSL) package is to provide implementations of different approaches to semi-supervised learning. Its primary goal is to facilitate research into these methods by making it easy to visualize and test the behaviour of these approaches on benchmark datasets. Additionally, it aims to provide user friendly implementations of these methods for practitioners.

### *Self-Learning*

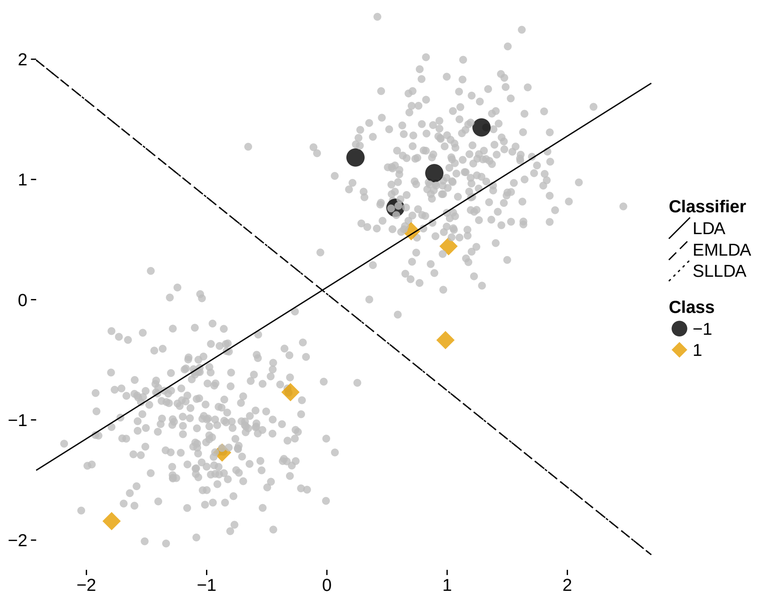


Figure: Self-Learning and the related Expectation Maximization type approaches work by using model predictions to impute the missing labels, after which the model is updated using these estimated labels. In this artificial dataset, self-learning and expectation-maximization versions of the linear discriminant classifier *reduce* performance compared to the supervised learner.

### *Low Density Separation*

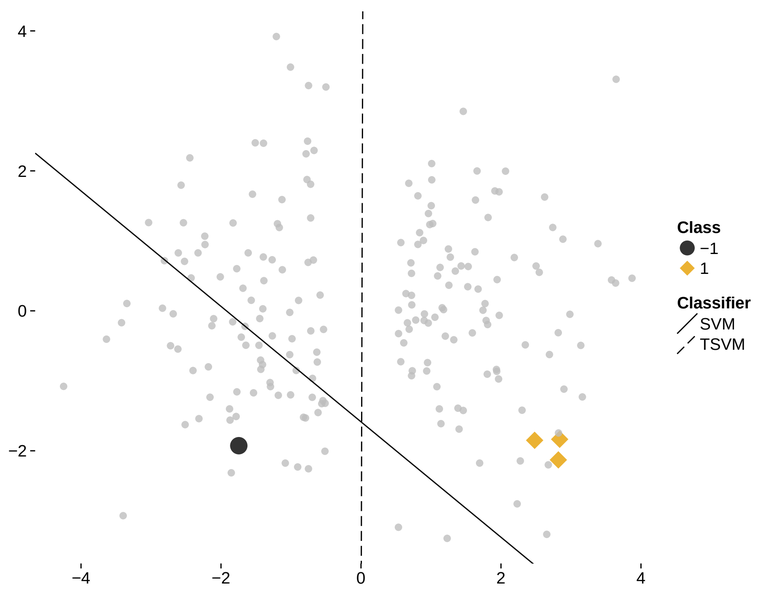


Figure: The Transductive SVM and related approaches nudge the decision boundary towards regions of low data density. In this artificial dataset, this assumption is useful and Transductive SVM leads to improved performance over the supervised SVM.

## Code Example

```
library(RSSL)
library(magrittr)
library(ggplot2)

# Plotting 2D classifiers
data_2gauss <- generate2ClassGaussian(n=500,d=2,var=0.2,expected=FALSE) %>%
  add_missinglabels_mar(formula=Class~.,prob=0.98)
problem <- data_2gauss %>% df_to_matrices(Class~.)

g_emlda <- EMLinearDiscriminantClassifier(problem$X,problem$y,problem$X_u)
ggplot(data_2gauss,aes(x=X1,y=X2,shape=Class,color=Class)) +
  geom_point() +
  geom_classifier("EMLDA"=g_emlda)

# Generate Learning Curve
datasets <- list("2 Gaussian Expected" =
                   generate2ClassGaussian(n=1000,d=2,expected=TRUE),
                 "2 Gaussian Non-Expected" =
                   generate2ClassGaussian(n=1000,d=2,expected=FALSE))
formulae <- list("2 Gaussian Expected" = formula(Class~.),
                 "2 Gaussian Non-Expected" = formula(Class~.))

classifiers <- list("LS" = function(X,y,X_u,y_u) {
                      LeastSquaresClassifier(X,y)},
                    "ICLS" = function(X,y,X_u,y_u) {
                      ICLeastSquaresClassifier(X,y,X_u)},
                    "EMLS" = function(X,y,X_u,y_u) {
                      EMLeastSquaresClassifier(X,y,X_u)},
                    "SLLS" = function(X,y,X_u,y_u) {
                      SelfLearning(X,y,X_u,
                                   method = LeastSquaresClassifier)})

measures = list("Error" = measure_error,
                "Loss test" = measure_losstest)

curve <- LearningCurveSSL(formulae, datasets, classifiers, measures,
                          type ="unlabeled", mc.cores=1,
                          n_l=10,sizes = 2^(0:10),repeats=200)
plot(curve)
```

### *Manifold Assumption*

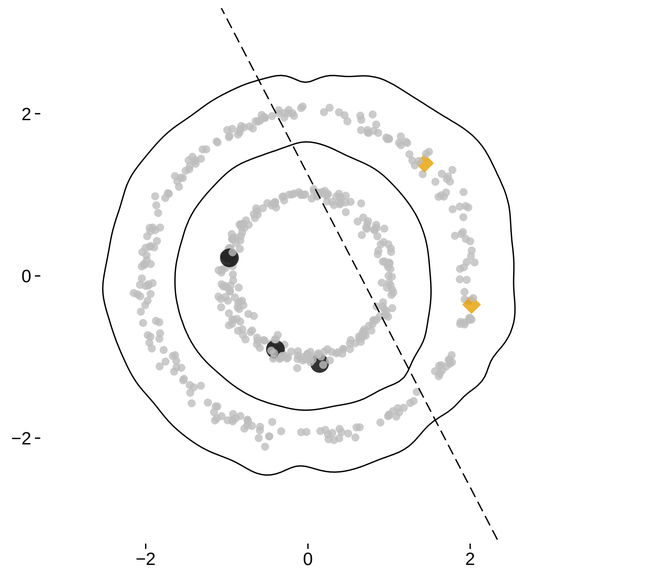


Figure: Assuming the labels change smoothly over the data manifolds makes it possible to "propagate" labels over the unlabeled data by ensuring its imputed label has to be similar to those objects that are close by. The Laplacian Regularized Least Squares Classifier with an RBF kernel is able to leverage this assumption to improve over the linear Least Squares Classifier.

### *Robust Estimation*

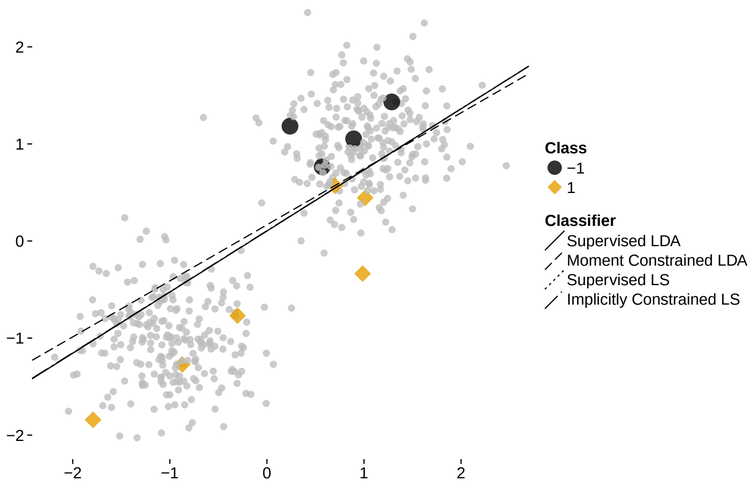


Figure: Compared to the first example, the goal of robust or "safe" methods is to ensure we at least do not reduce performance compared to the supervised alternative. Compare this to the self-learning methods, which are less conservative, leading to degraded performance on this dataset.

## Evaluation

### *Learning Curves*

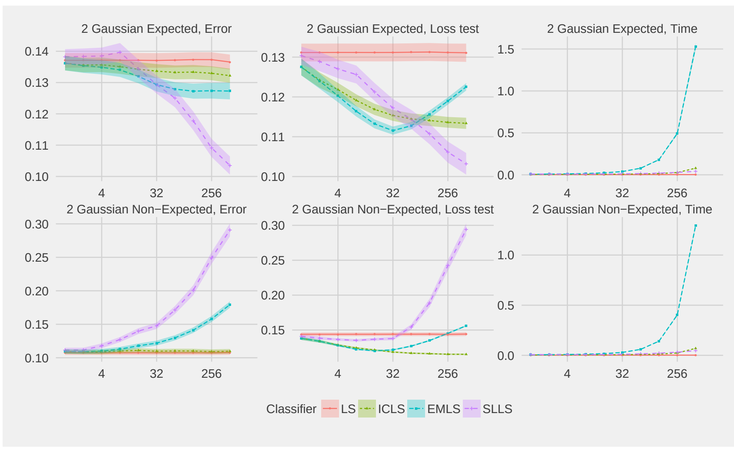


Figure: Example of a learning curve for the behaviour of different semi-supervised learning approaches as the number of unlabeled objects is increased, using 10 labeled objects. The datasets are the same as the two class gaussian dataset used before when either the classes correspond to the gaussian clusters (top) or the true decision boundary crosses the classes (bottom).

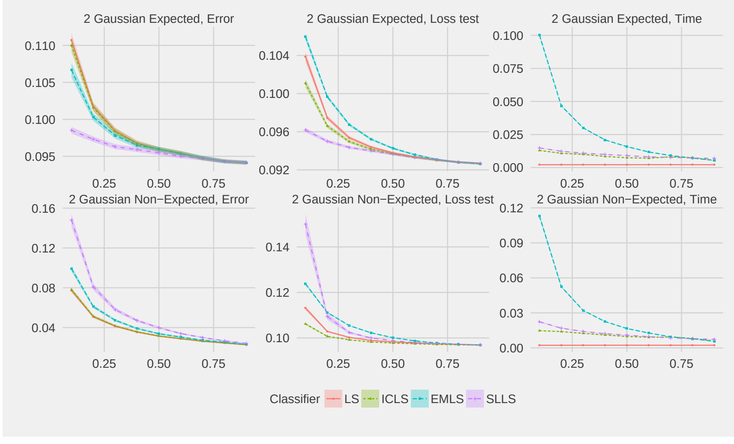


Figure: Similar setup to the previous figure, but now the *fraction* of labeled objects is varied. The total number of objects is 200, while 800 objects are used for evaluating the performance.

### *Cross-Validation*

| Dataset | Supervised | Self-Learning | ICLS | Oracle |
|---|---|---|---|---|
| Ionosphere | 0.29 | 0.24 (1) | **0.19 (0)** | 0.13 |
| Parkinsons | 0.33 | 0.29 (3) | 0.27 (0) | 0.11 |
| Diabetes | 0.32 | 0.33 (16) | **0.31 (2)** | 0.23 |
| Sonar | 0.42 | 0.37 (1) | **0.32 (1)** | 0.25 |
| SPECT | 0.42 | 0.40 (7) | **0.33 (0)** | 0.17 |
| WDBC | 0.27 | 0.17 (0) | **0.12 (0)** | 0.04 |
| Digit1 | 0.41 | 0.34 (0) | **0.20 (0)** | 0.06 |
| BCI | 0.40 | 0.35 (0) | **0.28 (0)** | 0.16 |
| g241d | 0.45 | 0.39 (0) | **0.29 (0)** | 0.13 |

Table: Example of a cross-validation experiment. Indicated in **bold** is when a semi-supervised classifier has significantly lower error than the other, using a Wilcoxon signed rank test at 0.01 significance level. A similar test is done to determine whether a semi-supervised classifier is significantly worse than the supervised classifier, indicated by underlined values.

## Discussion

- What interface for the classifiers best facilitates interaction with other packages providing hyperparameter search and model selection?
- Should we aim for replicability, calling the original implementations of the different methods, or reproducibility, by providing new implementations in R?

TU Delft

Leiden University Medical Center

COMMIT/