example-self-learning-rssl

February 21, 2018

1 A Failure of Self-Learning

Tomado de RSSL: Semi-supervised Learning in R, Jesse H. Krijthe.

While semi-supervised learning may seem to be obviously helpful, the fact that semi-supervised methods can actually lead to worse performance than their supervised counterparts has been both widely observed and described [4]. We will generate an example where unlabeled data is helpful (using the 2ClassGaussian problem from Figure 1) and one where unlabeled data actually leads to an increase in the classification error (2ClassGaussian (alt) in Figure 1), for the least squares classifier and self-learning as the semi-supervised learner. This can be done using the following code.

En la Figura 2 se ven los resultados del error de la clasificación, en azul la supervisada (usando solo los datos etiquetados) y en rojo la semi-supervisado (usando todos los datos).

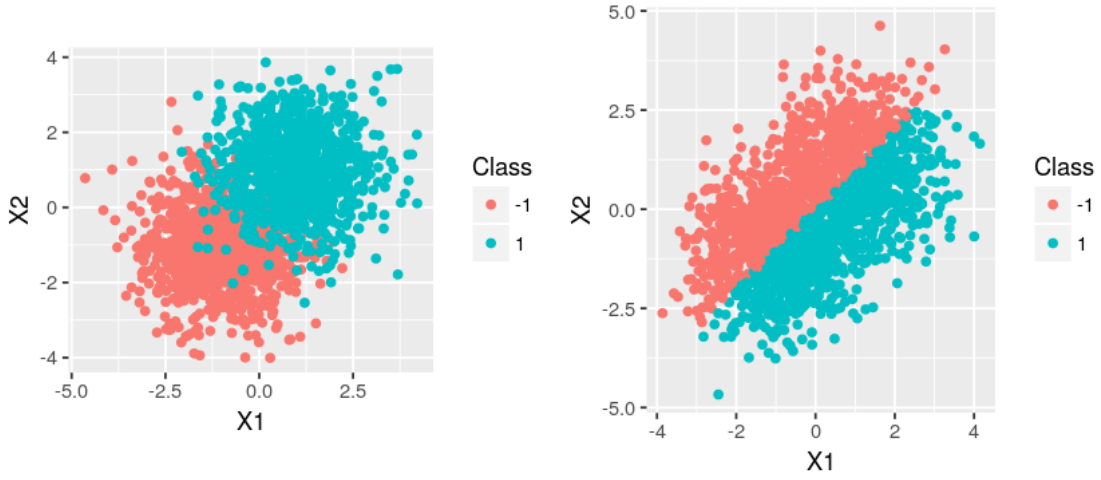
[4] Cozman, F.G., Cohen, I., Cirelo, M.C.: Semi-Supervised Learning of Mixture Models.

```
In [110]: library(RSSL)
          library(ggplot2)
          require(gridExtra)
```

```
In [111]: set.seed(1)
          # Set the datasets and corresponding formula objects
          datasets <- list("2 Gaussian Expected"=generate2ClassGaussian(n=2000,d=2, var = 1,ex
              "2 Gaussian Non-Expected"=generate2ClassGaussian(n=2000,d=2, var = 1,expected=FA
          formulae <- list("2 Gaussian Expected"=formula(Class~.),
              "2 Gaussian Non-Expected"=formula(Class~.))
```

2 Figura 1

```
In [112]: gp1<-ggplot(datasets[[1]],aes(x=X1,y=X2,color=Class)) + geom_point() + coord_equal()
          gp2<-ggplot(datasets[[2]],aes(x=X1,y=X2,color=Class)) + geom_point() + coord_equal()
          grid.arrange(gp1, gp2, ncol=2) # para subplots
```



In [113]:
```
# Define the classifiers to be used
classifiers <- list("Supervised" = function(X,y,X_u,y_u) {LeastSquaresClassifier(X,y
    "Self-learning" = function(X,y,X_u,y_u) { SelfLearning(X,y,X_u,method = LeastSqua
```

In [114]:
```
# Define the performance measures to be used and run experiment
rep = 100
n_sizes = 10
measures <- list("Error" = measure_error, "Loss" = measure_losstest)
results_lc <- LearningCurveSSL(formulae,datasets,
    classifiers=classifiers,measures=measures,verbose=FALSE,
    repeats=rep,n_l=10,sizes = 2^(1:n_sizes))
```

In [115]:
```
# Una muestra de cómo es la estructura de results_lc y cómo acceder a sus partes
str(results_lc)
```

```
          results_lc[[2]]
          results_lc[[2]]$`Number of unlabeled objects`[1:results_lc[[1]]]
```

```
List of 3
 $ n_l    : num 10
 $ results:'data.frame':        12000 obs. of  6 variables:
  ..$ repeats                    : int [1:12000] 1 1 1 1 1 1 1 1 1 1 ...
  ..$ Number of unlabeled objects: int [1:12000] 2 4 8 16 32 64 128 256 512 1024 ...
  ..$ Classifier                 : Factor w/ 2 levels "Supervised","Self-learning": 1 1 1 1 1 1
  ..$ Measure                    : Factor w/ 3 levels "Error","Loss",..: 1 1 1 1 1 1 1 1 1 1 .
  ..$ value                      : num [1:12000] 0.113 0.113 0.113 0.113 0.112 ...
  ..$ Dataset                    : chr [1:12000] "2 Gaussian Expected" "2 Gaussian Expected" "2
 $ n_test : num 1000
 - attr(*, "class")= chr "LearningCurve"
```

repeats	Number of unlabeled objects	Classifier	Measure	value	Dataset
1	2	Supervised	Error	0.11267606	2 Gaussian Expected
1	4	Supervised	Error	0.11278953	2 Gaussian Expected
1	8	Supervised	Error	0.11301715	2 Gaussian Expected
1	16	Supervised	Error	0.11296859	2 Gaussian Expected
1	32	Supervised	Error	0.11235955	2 Gaussian Expected
1	64	Supervised	Error	0.11266874	2 Gaussian Expected
1	128	Supervised	Error	0.11278195	2 Gaussian Expected
1	256	Supervised	Error	0.11245675	2 Gaussian Expected
1	512	Supervised	Error	0.11502030	2 Gaussian Expected
1	1024	Supervised	Error	0.10766046	2 Gaussian Expected
1	2	Self-learning	Error	0.09406439	2 Gaussian Expected
1	4	Self-learning	Error	0.09063444	2 Gaussian Expected
1	8	Self-learning	Error	0.09233098	2 Gaussian Expected
1	16	Self-learning	Error	0.09371834	2 Gaussian Expected
1	32	Self-learning	Error	0.08631256	2 Gaussian Expected
1	64	Self-learning	Error	0.09293873	2 Gaussian Expected
1	128	Self-learning	Error	0.09398496	2 Gaussian Expected
1	256	Self-learning	Error	0.08996540	2 Gaussian Expected
1	512	Self-learning	Error	0.09066306	2 Gaussian Expected
1	1024	Self-learning	Error	0.09006211	2 Gaussian Expected
1	2	Supervised	Loss	0.14466129	2 Gaussian Expected
1	4	Supervised	Loss	0.14463581	2 Gaussian Expected
1	8	Supervised	Loss	0.14483941	2 Gaussian Expected
1	16	Supervised	Loss	0.14460208	2 Gaussian Expected
1	32	Supervised	Loss	0.14420181	2 Gaussian Expected
1	64	Supervised	Loss	0.14507539	2 Gaussian Expected
1	128	Supervised	Loss	0.14430407	2 Gaussian Expected
1	256	Supervised	Loss	0.14360941	2 Gaussian Expected
1	512	Supervised	Loss	0.14353100	2 Gaussian Expected
1	1024	Supervised	Loss	0.14231100	2 Gaussian Expected
100	2	Self-learning	Loss	0.11210856	2 Gaussian Non-Expected
100	4	Self-learning	Loss	0.12552091	2 Gaussian Non-Expected
100	8	Self-learning	Loss	0.11377464	2 Gaussian Non-Expected
100	16	Self-learning	Loss	0.13434285	2 Gaussian Non-Expected
100	32	Self-learning	Loss	0.11135146	2 Gaussian Non-Expected
100	64	Self-learning	Loss	0.09743428	2 Gaussian Non-Expected
100	128	Self-learning	Loss	0.09626715	2 Gaussian Non-Expected
100	256	Self-learning	Loss	0.09292103	2 Gaussian Non-Expected
100	512	Self-learning	Loss	0.09516855	2 Gaussian Non-Expected
100	1024	Self-learning	Loss	0.10735350	2 Gaussian Non-Expected
100	2	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	4	Supervised	Time	0.00000000	2 Gaussian Non-Expected
100	8	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	16	Supervised	Time	0.00000000	2 Gaussian Non-Expected
100	32	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	64	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	128	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	256	Supervised	Time	0.00100000	2 Gaussian Non-Expected
100	512	Supervised	Time	0.00000000	2 Gaussian Non-Expected
100	1024	Supervised	Time	0.00000000	2 Gaussian Non-Expected
100	2	Self-learning	Time	0.00200000	2 Gaussian Non-Expected

1. 2 2. 4 3. 8 4. 16 5. 32 6. 64 7. 128 8. 256 9. 512 10. 1024

In [116]:
```
# Medias de errores en expected
df <- results_lc[[2]]

error_s_exp <- array(dim=c(rep,n_sizes))
for(r in 1:rep) {
    error_s_exp[r,] = df[df$repeats == r & df$Classifier == 'Supervised' & df$Measure
        & df$Dataset == '2 Gaussian Expected', "value"]
}
error_s_exp_mean <- apply(error_s_exp, 2, mean, na.rm=TRUE)

error_sl_exp <- array(dim=c(rep,n_sizes))
for(r in 1:rep) {
    error_sl_exp[r,] = df[df$repeats == r & df$Classifier == 'Self-learning' & df$Mea
        & df$Dataset == '2 Gaussian Expected', "value"]
}
error_sl_exp_mean <- apply(error_sl_exp, 2, mean, na.rm=TRUE)
```

In [117]:
```
# Medias de errores en non-expected
error_s_noexp <- array(dim=c(rep,n_sizes))
for(r in 1:rep) {
    error_s_noexp[r,] = df[df$repeats == r & df$Classifier == 'Supervised' & df$Measu
        & df$Dataset == '2 Gaussian Non-Expected', "value"]
}
error_s_noexp_mean <- apply(error_s_noexp, 2, mean, na.rm=TRUE)

error_sl_noexp <- array(dim=c(rep,n_sizes))
for(r in 1:rep) {
    error_sl_noexp[r,] = df[df$repeats == r & df$Classifier == 'Self-learning' & df$
        & df$Dataset == '2 Gaussian Non-Expected', "value"]
}
error_sl_noexp_mean <- apply(error_sl_noexp, 2, mean, na.rm=TRUE)
```

3 Figura 2

In [118]:
```
par(mfrow=c(1,2), pty="s") # Para subplots cuadrados
xticks <- df$`Number of unlabeled objects`[1:results_lc[[1]]]

p1 <- plot(xticks, error_s_exp_mean, type="o", ylim=c(0, 0.25), log="x", xaxt="n",
           xlab="num. unlabeled", ylab="Error rate",
           pch = 21, col='blue', asp = 0.1)
lines(xticks, error_sl_exp_mean, type="o", pch = 21, col='red')
axis(1, at=xticks,labels=xticks)
title(main="2 Gaussian Expected")

p2 <- plot(xticks, error_s_noexp_mean, type="o", ylim=c(0, 0.25), log="x", xaxt="n",
           xlab="num. unlabeled", ylab="Error rate",
```

```
            pch = 21, col='blue', asp = 0.1)
lines(xticks, error_sl_noexp_mean, type="o", pch = 21, col='red')
axis(1, at=xticks,labels=xticks)
title(main="2 Gaussian non-Expected")
```