Generating Learning Curves

Luis H. John, Xiaoyong Pan, Jenna Reps, Peter R. Rijnbeek

2018-10-07

Contents

1 Introduction

This vignette describes how you can use the Observational Health Data Sciences and Informatics (OHDSI) `PatientLevelPrediction` package to generate learning curves. This vignette assumes you have read and are comfortable with building patient level prediction models as described in the `BuildingPredictiveModels` vignette.

Prediction models will show overly-optimistic performance when predicting on the same data as used for training. Therefore, best-practice is to partition our data into a training set and testing set. We then train our prediction model on the training set portion and asses its ability to generalize to unseen data by measuring its performance on the testing set.

Learning curves assess the effect of training set size on model performance by training a sequence of prediction models on successively larger subsets of the training set. A learning curve plot can also help in diagnosing a bias or variance problem as explained below.

Figure 1, shows an example of learning curve plot in which the vertical axis represents the model performance and the horizontal axis the training set size. If training set size is small, the performance on the training set is high, because a model can often be fitted well to a limited number of training examples. At the same time, the performance on the testing set will be poor, because the model trained on such a limited number of training examples will not generalize well to unseen data in the testing set. As the training set size increases, the performance of the model on the training set will decrease. It becomes more difficult for the model to find a good fit through all the training examples. Also, the model will be trained on a more representative portion of training examples, making it generalize better to unseen data. This can be observed by the increasin testing set performance.

The learning curve can help us in diagnosing bias and variance problems with our classifier which will provide guidance on how to further improve our model. We can observe high variance (overfitting) in a prediction model if it performs well on the training set, but poorly on the testing set (Figure 2). Adding additional data is a common approach to counteract high variance. From the learning curve it becomes apparent, that adding additional data may improve performance on the testing set a little further, as the learning curve has not yet plateaued and, thus, the model is not saturated yet. Therefore, adding more data will decrease the gap between training set and testing set, which is the main indicator for a high variance problem.

Furthermore, we can observe high bias (underfitting) if a prediction model performs poorly on the training set as well as on the testing set (Figure 3). The learning curves of training set and testing set have flattened
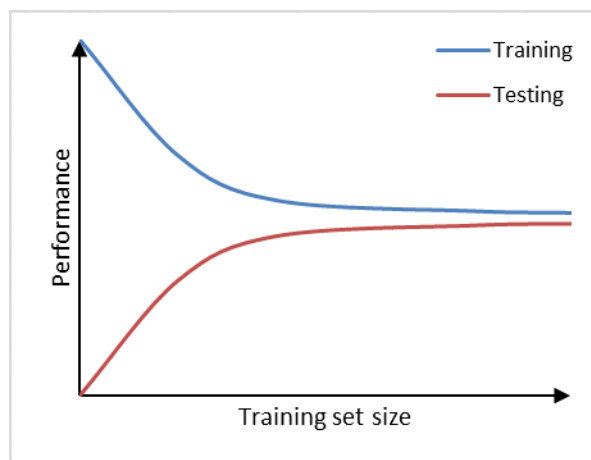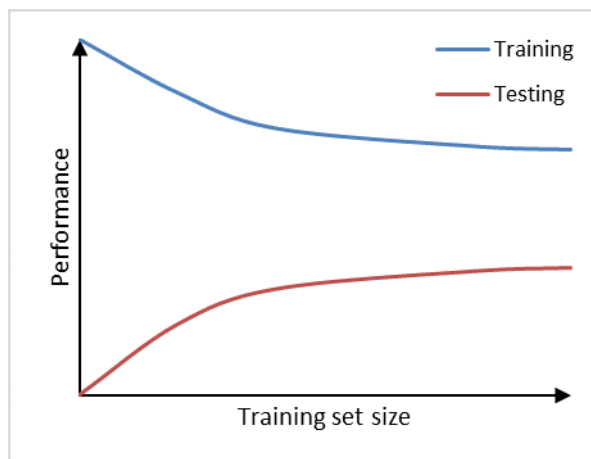
Figure 1: Learning curve example.

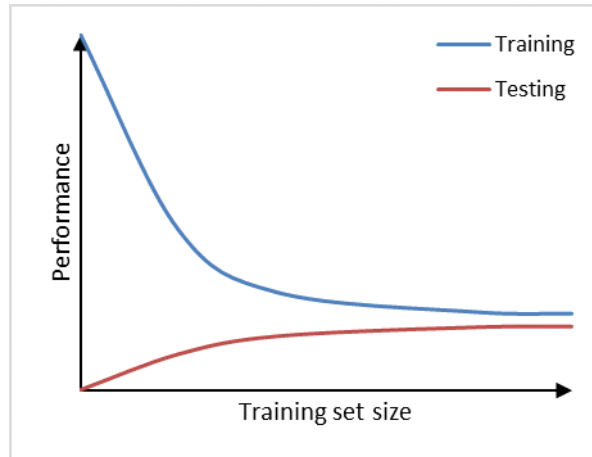Figure 2: Prediction model suffering from high variance.

Figure 3: Prediction model suffering from high bias.

on a low performance with only a small gap in between them. Adding additional data will in this case have little to no impact on the model performance. Choosing another prediction algorithm that can find more complex (for example non-linear) relationships in the data may be an alternative approach to consider in this high bias situation.

2 Generating the learning curve

Use the `PatientLevelPrediction` package to generate a `population` and `plpData` object. Alternatively, you can make use of the data simulator. The following code snippet creates a population of 12000 patients.

```
set.seed(1234)
data(plpDataSimulationProfile)
sampleSize <- 12000
plpData <- simulatePlpData(
  plpDataSimulationProfile,
  n = sampleSize
)

population <- createStudyPopulation(
  plpData,
  outcomeId = 2,
  binary = TRUE,
  firstExposureOnly = FALSE,
  washoutPeriod = 0,
  removeSubjectsWithPriorOutcome = FALSE,
  priorOutcomeLookback = 99999,
  requireTimeAtRisk = FALSE,
  minTimeAtRisk = 0,
  riskWindowStart = 0,
  addExposureDaysToStart = FALSE,
  riskWindowEnd = 365,
  addExposureDaysToEnd = FALSE,
  verbosity = "INFO"
)
```

Specify the prediction algorithm to be used.

```
# Use LASSO logistic regression
modelSettings <- setLassoLogisticRegression()
```

Specify a test fraction and a sequence of training set fractions.

```
testFraction <- 0.2
trainFractions <- seq(0.1, 0.8, 0.1) # Create eight training set fractions
```

Specify the test split to be used.

```
# Use a split by person, alternatively a time split is possible
testSplit <- 'person'
```

Create the learning curve object.

```
learningCurve <- createLearningCurve(population,
                                     plpData = plpData,
                                     modelSettings = modelSettings,
                                     testFraction = 0.2,
                                     verbosity = "TRACE",
                                     trainFractions = trainFractions,
                                     splitSeed = 1000,
                                     saveModel = TRUE)
```

Plot the learning curve object (Figure 4). Specify one of the available metrics: `AUROC`, `AUPRC`, `sBrier`.

```
plotLearningCurve(
  learningCurve,
  metric='AUROC',
  plotTitle = 'Learning Curve',
  plotSubtitle = 'AUROC performance'
)
```

3   Parallel processing

The learning curve object can be created in parallel, which can reduce computation time significantly. Currently this functionality is only available for LASSO logistic regression. Depending on the number of parallel workers it may require a significant amount of memory. We advise to use the parallelized learning curve function for parameter search and exploratory data analysis. Logging and saving functionality is unavailable.

Use the parallelized version of the learning curve function to create the learning curve object in parallel. R will find the number of available processing cores automatically and register the required parallel backend.

```
learningCurvePar <- createLearningCurvePar(
  population,
  plpData =  plpData,
  modelSettings = modelSettings,
  testSplit = testSplit,
  testFraction = testFraction,
  trainFractions = trainFractions,
  splitSeed = 1000
)
```
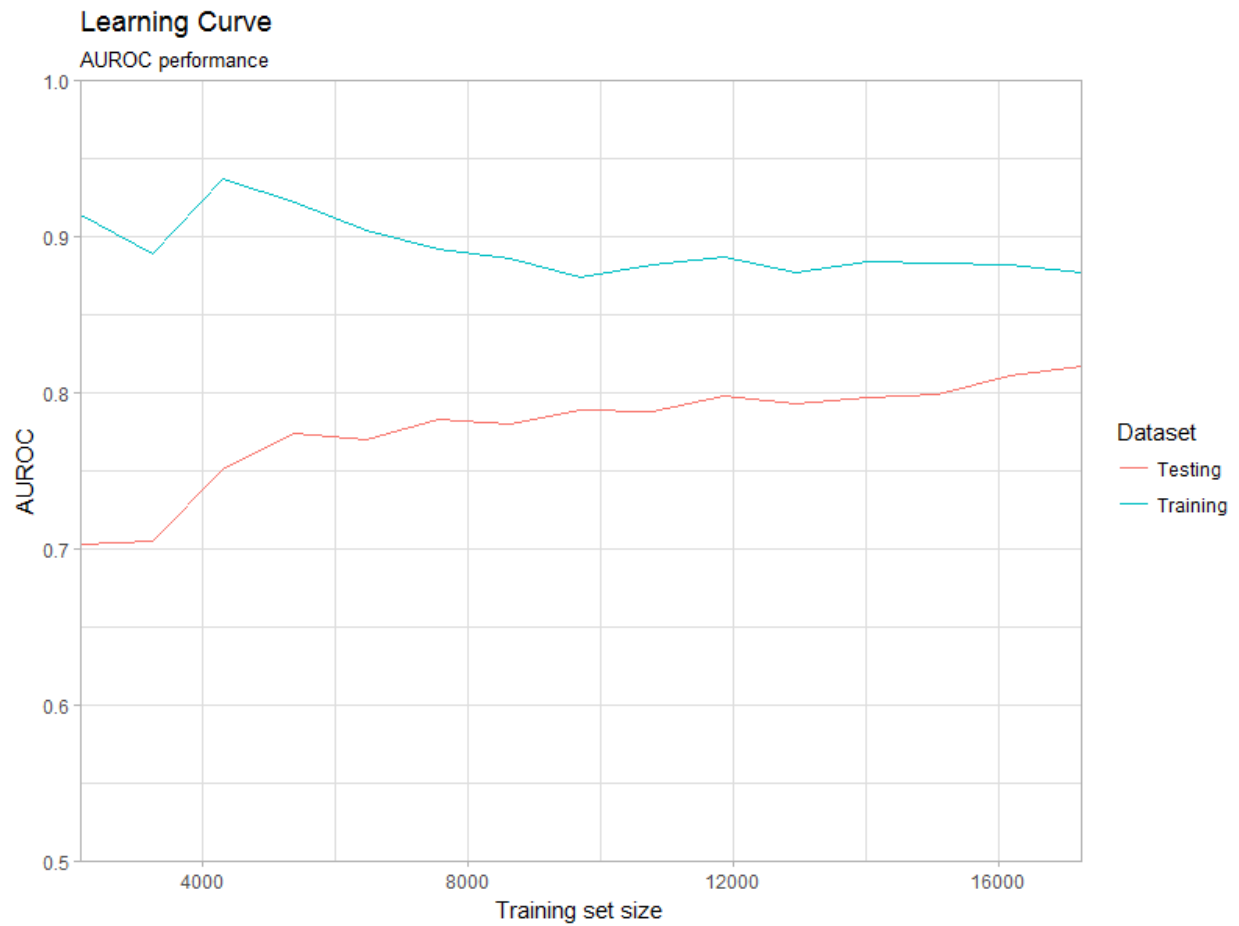
Figure 4: Learning curve plot.

4 Demo

We have added a demo of the learningcurve:

```r
# Show all demos in our package:
 demo(package = "PatientLevelPrediction")

# Run the learning curve
 demo("LearningCurveDemo", package = "PatientLevelPrediction")
```

Do note that running this demo can take a considerable amount of time (15 min on Quad core running in parallel)!

5 Acknowledgments

Considerable work has been dedicated to provide the `PatientLevelPrediction` package.

```r
citation("PatientLevelPrediction")
```

```
##
##   Jenna Reps, Martijn J. Schuemie, Marc A. Suchard, Patrick B.
##   Ryan and Peter R. Rijnbeek (2018). PatientLevelPrediction:
##   Package for patient level prediction using data in the OMOP
##   Common Data Model. R package version 3.0.0.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {PatientLevelPrediction: Package for patient level prediction using data in the OMOP Comm
## Model},
##     author = {Jenna Reps and Martijn J. Schuemie and Marc A. Suchard and Patrick B. Ryan and Peter R
##     year = {2018},
##     note = {R package version 3.0.0},
##   }
```

Please reference this paper if you use the PLP Package in your work:

Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-975.