

A Short Introduction to the **caret** Package

Max Kuhn
max.kuhn@pfizer.com

September 1, 2017

The **caret** package (short for **c**lassification and **r**egression **t**raining) contains functions to streamline the model training process for complex regression and classification problems. The package utilizes a number of R packages but tries not to load them all at package start-up¹. The package “suggests” field includes 27 packages. **caret** loads packages as needed and assumes that they are installed. Install **caret** using

```
> install.packages("caret", dependencies = c("Depends", "Suggests"))
```

to ensure that all the needed packages are installed.

The **main help pages** for the package are at:

<http://caret.r-forge.r-project.org/>

Here, there are extended examples and a large amount of information that previously found in the package vignettes.

caret has several functions that attempt to streamline the model building and evaluation process, as well as feature selection and other techniques.

One of the primary tools in the package is the **train** function which can be used to

- evaluate, using resampling, the effect of model tuning parameters on performance
- choose the “optimal” model across these parameters
- estimate model performance from a training set

¹By adding formal package dependencies, the package startup time can be greatly decreased

More formally:

```

1 Define sets of model parameter values to evaluate
2 for each parameter set do
3   for each resampling iteration do
4     Hold-out specific samples
5     [Optional] Pre-process the data
6     Fit the model on the remainder
7     Predict the hold-out samples
8   end
9   Calculate the average performance across hold-out predictions
10 end
11 Determine the optimal parameter set
12 Fit the final model to all the training data using the optimal parameter set

```

There are options for customizing almost every step of this process (e.g. resampling technique, choosing the optimal parameters etc). To demonstrate this function, the Sonar data from the `mlbench` package will be used.

The Sonar data consist of 208 data points collected on 60 predictors. The goal is to predict the two classes (M for metal cylinder or R for rock).

First, we split the data into two groups: a training set and a test set. To do this, the `createDataPartition` function is used:

```

> library(caret)
> library(mlbench)
> data(Sonar)
> set.seed(107)
> inTrain <- createDataPartition(y = Sonar$Class,
+                               ## the outcome data are needed
+                               p = .75,
+                               ## The percentage of data in the
+                               ## training set
+                               list = FALSE)
>                               ## The format of the results
>
> ## The output is a set of integers for the rows of Sonar
> ## that belong in the training set.
> str(inTrain)

int [1:157, 1] 1 2 3 6 7 9 10 11 12 13 ...
- attr(*, "dimnames")=List of 2
 ..$ : NULL
 ..$ : chr "Resample1"

```

By default, `createDataPartition` does a stratified random split of the data. To partition the data:

```
> training <- Sonar[ inTrain,]
> testing  <- Sonar[-inTrain,]
> nrow(training)
```

```
[1] 157
```

```
> nrow(testing)
```

```
[1] 51
```

To tune a model using Algorithm , the `train` function can be used. More details on this function can be found at:

<http://caret.r-forge.r-project.org/training.html>

Here, a partial least squares discriminant analysis (PLSDA) model will be tuned over the number of PLS components that should be retained. The most basic syntax to do this is:

```
> plsFit <- train(Class ~ .,
+               data = training,
+               method = "pls",
+               ## Center and scale the predictors for the training
+               ## set and all future samples.
+               preProc = c("center", "scale"))
```

However, we would probably like to customize it in a few ways:

- expand the set of PLS models that the function evaluates. By default, the function will tune over three values of each tuning parameter.
- the type of resampling used. The simple bootstrap is used by default. We will have the function use three repeats of 10-fold cross-validation.
- the methods for measuring performance. If unspecified, overall accuracy and the Kappa statistic are computed. For regression models, root mean squared error and R^2 are computed. Here, the function will be altered to estimate the area under the ROC curve, the sensitivity and specificity

To change the candidate values of the tuning parameter, either of the `tuneLength` or `tuneGrid` arguments can be used. The `train` function can generate a candidate set of parameter values and the `tuneLength` argument controls how many are evaluated. In the case of PLS, the function uses a sequence of integers from 1 to `tuneLength`. If we want to evaluate all integers between 1 and 15, setting `tuneLength = 15` would achieve this. The `tuneGrid` argument is used when specific values are desired. A data frame is used where each row is a tuning parameter setting and each column is a tuning parameter. An example is used below to illustrate this.

The syntax for the model would then be:

```
> plsFit <- train(Class ~ .,
+               data = training,
+               method = "pls",
+               tuneLength = 15,
+               preProc = c("center", "scale"))
```

To modify the resampling method, a `trainControl` function is used. The option `method` controls the type of resampling and defaults to `"boot"`. Another method, `"repeatedcv"`, is used to specify repeated K -fold cross-validation (and the argument `repeats` controls the number of repetitions). K is controlled by the `number` argument and defaults to 10. The new syntax is then:

```
> ctrl <- trainControl(method = "repeatedcv",
+                      repeats = 3)
> plsFit <- train(Class ~ .,
+               data = training,
+               method = "pls",
+               tuneLength = 15,
+               trControl = ctrl,
+               preProc = c("center", "scale"))
```

Finally, to choose different measures of performance, additional arguments are given to `trainControl`. The `summaryFunction` argument is used to pass in a function that takes the observed and predicted values and estimate some measure of performance. Two such functions are already included in the package: `defaultSummary` and `twoClassSummary`. The latter will compute measures specific to two-class problems, such as the area under the ROC curve, the sensitivity and specificity. Since the ROC curve is based on the predicted class probabilities (which are not computed automatically), another option is required. The `classProbs = TRUE` option is used to include these calculations.

Lastly, the function will pick the tuning parameters associated with the best results. Since we are using custom performance measures, the criterion that should be optimized must also be specified. In the call to `train`, we can use `metric = "ROC"` to do this.

The final model fit would then be:

```
> set.seed(123)
> ctrl <- trainControl(method = "repeatedcv",
+                      repeats = 3,
+                      classProbs = TRUE,
+                      summaryFunction = twoClassSummary)
> plsFit <- train(Class ~ .,
+               data = training,
+               method = "pls",
+               tuneLength = 15,
+               trControl = ctrl,
+               metric = "ROC",
+               preProc = c("center", "scale"))

> plsFit
```

Partial Least Squares

```
157 samples
60 predictor
2 classes: 'M', 'R'
```

```
Pre-processing: centered (60), scaled (60)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 141, 141, 142, 141, 142, 141, ...
Resampling results across tuning parameters:
```

| ncomp | ROC | Sens | Spec |
|-------|-----------|-----------|-----------|
| 1 | 0.8104663 | 0.7157407 | 0.7297619 |
| 2 | 0.8601025 | 0.7703704 | 0.8000000 |
| 3 | 0.8630208 | 0.7620370 | 0.8196429 |
| 4 | 0.8649802 | 0.7657407 | 0.7821429 |
| 5 | 0.8393105 | 0.7282407 | 0.7744048 |
| 6 | 0.8139964 | 0.7523148 | 0.7880952 |
| 7 | 0.8003555 | 0.7083333 | 0.7589286 |
| 8 | 0.8100446 | 0.7069444 | 0.7589286 |
| 9 | 0.8079034 | 0.7148148 | 0.7642857 |
| 10 | 0.8136326 | 0.7120370 | 0.7720238 |
| 11 | 0.8146991 | 0.7120370 | 0.7589286 |
| 12 | 0.8203538 | 0.7240741 | 0.7625000 |
| 13 | 0.8212632 | 0.7333333 | 0.7583333 |
| 14 | 0.8158730 | 0.7407407 | 0.7583333 |
| 15 | 0.8195685 | 0.7245370 | 0.7488095 |

```
ROC was used to select the optimal model using the largest value.
The final value used for the model was ncomp = 4.
```

In this output the grid of results are the average resampled estimates of performance. The note at the bottom tells the user that PLS components were found to be optimal. Based on this value, a final PLS model is fit to the whole data set using this specification and this is the model that is used to predict future samples.

The package has several functions for visualizing the results. One method for doing this is the `plot` function for `train` objects. The command `plot(plsFit)` produced the results seen in Figure 1 and shows the relationship between the resampled performance values and the number of PLS components.

To predict new samples, `predict.train` can be used. For classification models, the default behavior is to calculate the predicted class. Using the option `type = "prob"` can be used to compute class probabilities from the model. For example:

```
> plsClasses <- predict(plsFit, newdata = testing)
> str(plsClasses)

Factor w/ 2 levels "M","R": 2 1 1 2 1 2 2 2 2 2 ...
```

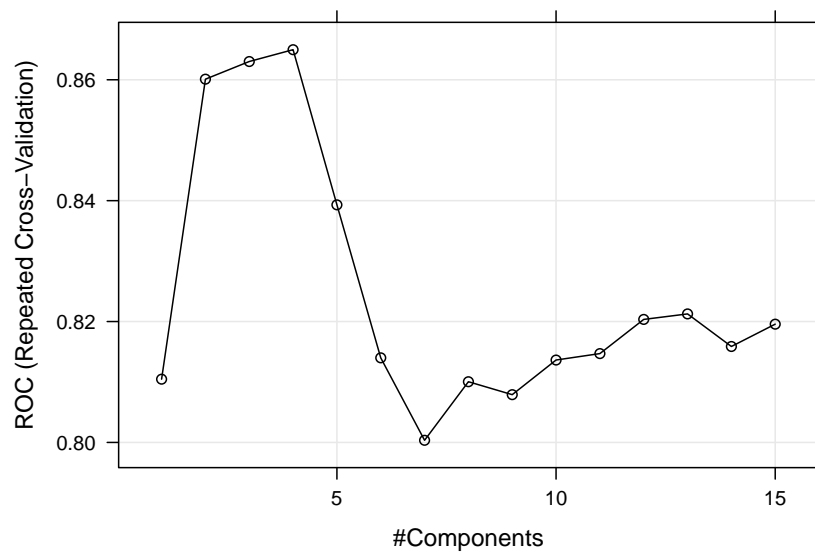


Figure 1: `plot(plsFit)` shows the relationship between the number of PLS components and the resampled estimate of the area under the ROC curve.

```
> plsProbs <- predict(plsFit, newdata = testing, type = "prob")
> head(plsProbs)
```

| | M | R |
|----|-----------|-----------|
| 4 | 0.4408455 | 0.5591545 |
| 5 | 0.5446774 | 0.4553226 |
| 8 | 0.6310261 | 0.3689739 |
| 16 | 0.3478948 | 0.6521052 |
| 20 | 0.7830037 | 0.2169963 |
| 25 | 0.2486379 | 0.7513621 |

caret contains a function to compute the confusion matrix and associated statistics for the model fit:

```
> confusionMatrix(data = plsClasses, testing$Class)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | M | R |
| M | 17 | 6 |
| R | 10 | 18 |

Accuracy : 0.6863

95% CI : (0.5411, 0.8089)
No Information Rate : 0.5294
P-Value [Acc > NIR] : 0.01674

Kappa : 0.3761
McNemar's Test P-Value : 0.45325

Sensitivity : 0.6296
Specificity : 0.7500
Pos Pred Value : 0.7391
Neg Pred Value : 0.6429
Prevalence : 0.5294
Detection Rate : 0.3333
Detection Prevalence : 0.4510
Balanced Accuracy : 0.6898

'Positive' Class : M

To fit another model to the data, `train` can be invoked with minimal changes. Lists of models available can be found at:

<http://caret.r-forge.r-project.org/modelList.html>

<http://caret.r-forge.r-project.org/bytag.html>

For example, to fit a regularized discriminant model to these data, the following syntax can be used:

```
> ## To illustrate, a custom grid is used
> rdaGrid = data.frame(gamma = (0:4)/4, lambda = 3/4)
> set.seed(123)
> rdaFit <- train(Class ~ .,
+               data = training,
+               method = "rda",
+               tuneGrid = rdaGrid,
+               trControl = ctrl,
+               metric = "ROC")
> rdaFit
```

Regularized Discriminant Analysis

157 samples
60 predictor
2 classes: 'M', 'R'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 141, 141, 142, 141, 142, 141, ...
Resampling results across tuning parameters:

| gamma | ROC | Sens | Spec |
|-------|-----------|-----------|-----------|
| 0.00 | 0.8257440 | 0.7726852 | 0.8107143 |
| 0.25 | 0.8877397 | 0.7884259 | 0.8000000 |
| 0.50 | 0.8812831 | 0.7805556 | 0.7821429 |
| 0.75 | 0.8707176 | 0.7611111 | 0.7369048 |
| 1.00 | 0.7560599 | 0.7018519 | 0.6607143 |

Tuning parameter 'lambda' was held constant at a value of 0.75
ROC was used to select the optimal model using the largest value.
The final values used for the model were gamma = 0.25 and lambda = 0.75.

```
> rdaClasses <- predict(rdaFit, newdata = testing)
> confusionMatrix(rdaClasses, testing$Class)
```

Confusion Matrix and Statistics

| | Reference | |
|------------|-----------|----|
| Prediction | M | R |
| M | 22 | 5 |
| R | 5 | 19 |

Accuracy : 0.8039
95% CI : (0.6688, 0.9018)
No Information Rate : 0.5294
P-Value [Acc > NIR] : 4.341e-05

Kappa : 0.6065
Mcnemar's Test P-Value : 1

Sensitivity : 0.8148
Specificity : 0.7917
Pos Pred Value : 0.8148
Neg Pred Value : 0.7917
Prevalence : 0.5294
Detection Rate : 0.4314
Detection Prevalence : 0.5294
Balanced Accuracy : 0.8032

'Positive' Class : M

How do these models compare in terms of their resampling results? The `resamples` function can be used to collect, summarize and contrast the resampling results. Since the random number seeds were initialized to the same value prior to calling `train`, the same folds were used for each model. To assemble them:

```
> resamps <- resamples(list(pls = plsFit, rda = rdaFit))
> summary(resamps)
```


Call:

```
summary.resamples(object = resamps)
```

Models: pls, rda

Number of resamples: 30

ROC

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|-----|-----------|-----------|-----------|-----------|-----------|------|------|
| pls | 0.5396825 | 0.8058036 | 0.8988095 | 0.8649802 | 0.9345238 | 1 | 0 |
| rda | 0.5873016 | 0.8577009 | 0.8988095 | 0.8877397 | 0.9642857 | 1 | 0 |

Sens

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|-----|-------|-----------|-----------|-----------|-----------|------|------|
| pls | 0.375 | 0.6666667 | 0.7638889 | 0.7657407 | 0.8854167 | 1 | 0 |
| rda | 0.375 | 0.6875000 | 0.8263889 | 0.7884259 | 0.8854167 | 1 | 0 |

Spec

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|-----|-----------|-----------|--------|-----------|-----------|------|------|
| pls | 0.2857143 | 0.7142857 | 0.75 | 0.7821429 | 0.8571429 | 1 | 0 |
| rda | 0.4285714 | 0.7142857 | 0.75 | 0.8000000 | 0.9687500 | 1 | 0 |

There are several functions to visualize these results. For example, a Bland–Altman type plot can be created using `xyplot(resamps, what = "BlandAltman")` (see Figure 2). The results look similar. Since, for each resample, there are paired results a paired t -test can be used to assess whether there is a difference in the average resampled area under the ROC curve. The `diff.resamples` function can be used to compute this:

```
> diffs <- diff(resamps)
> summary(diffs)
```

Call:

```
summary.diff.resamples(object = diffs)
```

p-value adjustment: bonferroni

Upper diagonal: estimates of the difference

Lower diagonal: p-value for H0: difference = 0

ROC

| | pls | rda |
|-----|----------|----------|
| pls | | -0.02276 |
| rda | 0.007757 | |

Sens

| | pls | rda |
|-----|--------|----------|
| pls | | -0.02269 |
| rda | 0.2357 | |

Spec

| | pls | rda |
|--|-----|-----|
|--|-----|-----|

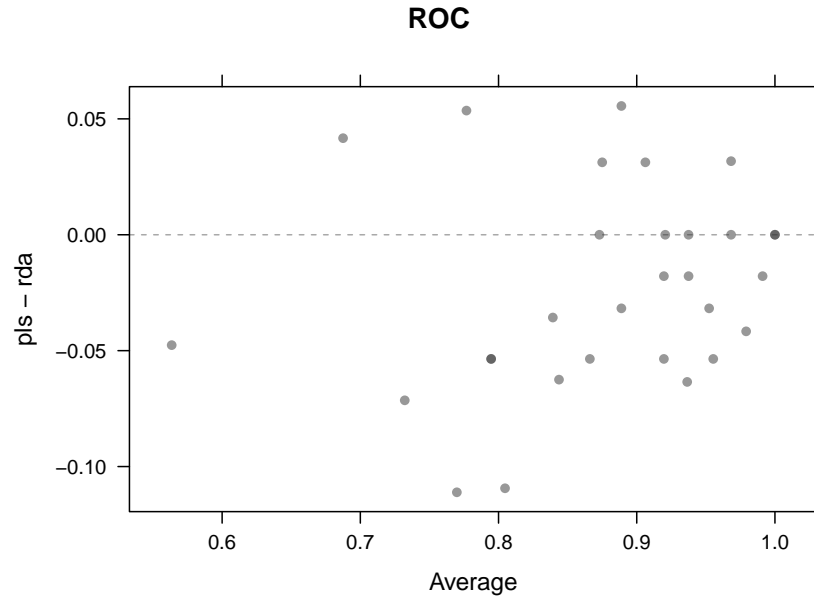


Figure 2: A Bland–Altman plot of the resampled ROC values produced using `xyplot(resamps, what = "BlandAltman")`.

```
pls      -0.01786
rda 0.4231
```

Based on this analysis, the difference between the models is -0.023 ROC units (the RDA model is slightly higher) and the two-sided p -value for this difference is 0.0077573.