

Solution 1:

- The inner loss is the loss that is optimized directly by the machine learning model. The outer loss is the loss (or performance measurement) used to evaluate the model.
- Which model is more likely to overfit the training data:
 - k-NN with 1 or with 10 neighbors? **1 neighbor**, because it's an exact memorization of training data.
 - Logistic regression with 10 or 20 features? **20 features**, because the more features, the more coefficients the learner estimates. More coefficients mean more degrees of freedom, which make overfitting more likely.
 - LDA or QDA? **QDA**, because it has more parameters to possibly overfit the data. LDA is more likely to underfit more complex relationships.
- Which of the following methods yield an unbiased generalization error estimate? Performance estimation ...
 - on training data: **Biased, too optimistic**
 - on test data: **Unbiased / Biased, too pessimistic** (Test data is *not included* / *included* in the final model)
 - on training and test data combined: **Biased, too optimistic** (But a little bit less than only using training data).
 - using cross validation: **Biased, too pessimistic** (The higher the ratio of folds / number of observation, the smaller the pessimistic bias)
 - using subsampling: **Biased, too pessimistic** (The smaller the subsampling rate, the larger the pessimistic bias)
- Resampling strategies solve the problem that comes from the randomness of the training and test data split: Error estimation using a single split has a high variance. Resampling estimates are more robust because they average over different splits.
- Nested resampling solves the problem of simultaneously conducting tuning/model selection and performance estimation. When we use the performance estimates from the same data that were used for model selection (as done in simple, not-nested resampling), the final error estimate is too optimistic.

Solution 2:

- The training performance is too optimistic (mmce of 0), because the mmce is higher on new data.
- The test performance is unbiased (if the final model is only trained on the training data), but it depends on the split, as can be seen in the CV folds: Each CV fold represents a training test split and the mmce measure varies between folds.
- The CV estimate averages over the different splits and gives an slightly pessimistic, more robust estimate.
- The CV estimate is preferable over the other two, but more computationally expensive.

Solution 3:

- a) Each loss function we have learned so far to fit the model (inner loss) can also be used as performance measure (outer loss).

For classification:

- 0-1 loss (= mean misclassification error),
- Logistic loss (bernoulli loss), ...

For regression:

- L_2 -loss (= mean squared error),
- L_1 -loss (= mean absolute error), ...

To get a list of all measures you can use `mlr_measures`.

```
b) # look at the task
task <- tsk("boston_housing")
task

## <TaskRegr:boston_housing> (506 x 19)
## * Target: medv
## * Properties: -
## * Features (18):
##   - dbl (13): age, b, cmedv, crim, dis, indus, lat, lon, lstat, nox,
##     ptratio, rm, zn
##   - int (3): rad, tax, tract
##   - fct (2): chas, town

n <- task$nrow

# select index vectors to subset the data randomly
set.seed(123)
train_ind <- sample(seq_len(n), 0.5*n)
test_ind <- setdiff(seq_len(n), train_ind)

# specify learner
learner <- lrn("regr.kknn", k = 3)

## Warning: Package 'kknn' required but not installed for Learner 'regr.kknn'

# train model to the training set
learner$train(task, row_ids = train_ind)

## Error: The following packages could not be loaded: kknn

# predict on the test set
pred <- learner$predict(task, row_ids = test_ind)

## Error: Cannot predict, Learner 'regr.kknn' has not been trained yet

pred

## <PredictionClassif> for 75 observations:
##   row_ids  truth  response
##       2    setosa  setosa
##       4    setosa  setosa
##       6    setosa  setosa
## ---
##    146 virginica virginica
##    148 virginica virginica
##    150 virginica virginica
```

```
c) # predict on the train set
pred_train <- learner$predict(task, row_ids = train_ind)
## Error: Cannot predict, Learner 'regr.kknn' has not been trained yet

pred_train$score(list(msr("regr.mse"), msr("regr.mae")))

## Error in eval(expr, envir, enclos): object 'pred_train' not found

# predict on the test set
pred_test <- learner$predict(task, row_ids = test_ind)

## Error: Cannot predict, Learner 'regr.kknn' has not been trained yet

pred_test$score(list(msr("regr.mse"), msr("regr.mae")))

## Error in eval(expr, envir, enclos): object 'pred_test' not found
```

Unsurprisingly the model performs better on the training data (smaller loss) then on the test data.

```
d) # select different index vectors to subset the data randomly
set.seed(321)
train_ind <- sample(seq_len(n), 0.5*n)
test_ind <- setdiff(seq_len(n), train_ind)

# specify learner
learner <- lrn("regr.kknn", k = 3)
## Warning: Package 'kknn' required but not installed for Learner 'regr.kknn'

# train model to the training set
learner$train(task, row_ids = train_ind)

## Error: The following packages could not be loaded: kknn

# predict on the test set
pred_test <- learner$predict(task, row_ids = test_ind)

## Error: Cannot predict, Learner 'regr.kknn' has not been trained yet

pred_test

## Error in eval(expr, envir, enclos): object 'pred_test' not found

pred_test$score(list(msr("regr.mse"), msr("regr.mae")))

## Error in eval(expr, envir, enclos): object 'pred_test' not found
```

Effect: We will predict different observations since the test set is different. The same observations get a slightly different prediction (e.g. observation with id 2). This affects the final error estimation.

```
e) rdesc <- rsmp("cv", folds = 10)
r <- resample(task, learner, rdesc)

## INFO [17:54:38.184] [mlr3] Applying learner 'regr.kknn' on task 'boston-housing' (iter 9/10)

## Error: The following packages could not be loaded: kknn

r$aggregate(list(msr("regr.mse"), msr("regr.mae")))

## Error in eval(expr, envir, enclos): object 'r' not found
```