# Comprehension Check: Ensembles

For these exercises we are going to build several machine learning models for the **`mnist_27`** dataset and then build an ensemble. Each of the exercises in this comprehension check builds on the last.

## Q1

1/1 point (graded)

Use the training set to build a model with several of the models available from the caret package. We will test out 10 of the most common machine learning models in this exercise:

```
models <- c("glm", "lda", "naive_bayes", "svmLinear", "knn", "gamLoess", "multinom", "qda", "rf", "
```

Apply all of these models using `train` with all the default parameters. You may need to install some packages. Keep in mind that you will probably get some warnings. Also, it will probably take a while to train all of the models - be patient!

Run the following code to train the various models:

```
library(caret)
library(dslabs)
set.seed(1) # use `set.seed(1, sample.kind = "Rounding")` in R 3.6 or later
data("mnist_27")

fits <- lapply(models, function(model){
        print(model)
        train(y ~ ., method = model, data = mnist_27$train)
})

names(fits) <- models
```

Did you train all of the models?

● Yes

○ No

✓

**Explanation**

Before proceeding, make sure you have trained all of the models - you will need the results for the following exercises.

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

## Q2

1/1 point (graded)

Now that you have all the trained models in a list, use `sapply` or `map` to create a matrix of predictions for the test set. You should end up with a matrix with `length(mnist_27$test$y)` rows and `length(models)` columns.

What are the dimensions of the matrix of predictions?

Number of rows:

200 ✓ **Answer:** 200

200

Number of columns:

10 ✓ **Answer:** 10

10

**Explanation**

You can generate the matrix of predictions for the test set and get its dimensions using the following code:

```
pred <- sapply(fits, function(object)
        predict(object, newdata = mnist_27$test))
dim(pred)
```

Submit You have used 1 of 10 attempts

ℹ Answers are displayed within the problem

## Q3

1/1 point (graded)

Now compute accuracy for each model on the test set.

Report the mean accuracy across all models.

0.789 ✓ **Answer:** 0.789

0.789

**Explanation**

Accuracy for each model in the test set and the mean accuracy across all models can be computed using the following code:

```
acc <- colMeans(pred == mnist_27$test$y)
acc
mean(acc)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q4

1/1 point (graded)

Next, build an ensemble prediction by majority vote and compute the accuracy of the ensemble.

What is the accuracy of the ensemble?

0.815 ✓ **Answer:** 0.815

0.815

**Explanation**

The ensemble prediction can be built using the following code:

```
votes <- rowMeans(pred == "7")
y_hat <- ifelse(votes > 0.5, "7", "2")
mean(y_hat == mnist_27$test$y)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q5

2/2 points (graded)

In Q3, we computed the accuracy of each method on the test set and noticed that the individual accuracies varied.

How many of the individual methods do better than the ensemble?

3 ✓ **Answer:** 3

3

Which individual methods perform better than the ensemble?
Select ALL that apply.

- ☐ glm
- ☐ lda
- ☐ naive_bayes
- ☐ svmLinear
- ☑ knn
- ☑ gamLoess
- ☐ multinom
- ☑ qda
- ☐ rf
- ☐ adaboost

✓

**Explanation**

The comparison of the individual methods to the ensemble can be done using the following code:

```
ind <- acc > mean(y_hat == mnist_27$test$y)
sum(ind)
models[ind]
```

Submit You have used 1 of 5 attempts

Answers are displayed within the problem

## Q6

1/1 point (graded)

It is tempting to remove the methods that do not perform well and re-do the ensemble. The problem with this approach is that we are using the test data to make a decision. However, we could use the minimum accuracy estimates obtained from cross validation with the training data for each model. Obtain these estimates and save them in an object. Report the mean of these training set accuracy estimates.

What is the mean of these training set accuracy estimates?

0.8085677 ✓ **Answer:** 0.809

0.8085677

**Explanation**

You can calculate the mean accuracy of the new estimates using the following code:

```
acc_hat <- sapply(fits, function(fit) min(fit$results$Accuracy))
mean(acc_hat)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q7

1/1 point (graded)

Now let's only consider the methods with an estimated accuracy of greater than or equal to 0.8 when constructing the ensemble.

What is the accuracy of the ensemble now?

0.83 ✓ **Answer:** 0.825

0.83

**Explanation**
The new ensemble prediction can be built using the following code:

```
ind <- acc_hat >= 0.8
votes <- rowMeans(pred[,ind] == "7")
y_hat <- ifelse(votes>=0.5, 7, 2)
mean(y_hat == mnist_27$test$y)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Ask your questions or make your comments about Ensembles here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

## Discussion: Ensembles

Show Discussion

**Topic:** Section 6: Model fitting and recommendation systems / 6.1.1: Ensembles