# Machine Learning Assessment

## Pre-Assessment Test:

This problem set uses the heights dataset from the **dslabs** package, which consists of actual heights (in inches) of students in 3 Harvard biostatistics courses.

Install the **dslabs** package from CRAN, then load the **dslabs** package into your workspace with the library() command.

After loading the package, load the dataset heights into your workspace:

data(heights)  
heights

### **Q1: Object Classes**

5.0/5.0 points (ungraded)

Match each object to its corresponding class.

heights dataset                                                                                  

correct

sex column



correct

height column



correct

"Male"                                                                                  

correct

75.00000                                                                                  

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q2: Object Dimensions**

1/1 point (ungraded)

How many rows are in this dataset?  correct

1050 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q3: Indexing - 1**

1/1 point (ungraded)

What is the height in row 777?  correct

61 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q4: Indexing - 2**

1/1 point (ungraded)

Which of these pieces of code returns the sex in row 777?

Check all correct answers.

heights$sex[777]

heights[1, 777]

heights[777,1]

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q5: Maximum and Minimum**

2/2 points (ungraded)

What is the maximum height in inches?  correct

82.67717 Loading

Which row has the minimum height?  correct

1032 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Q6: Summary Statistics**

2/2 points (ungraded)

What is the mean height in inches?  correct

68.32301 Loading

What is the median height in inches?  correct

68.5 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Q7: Conditional Statements- 1**

1/1 point (ungraded)

What proportion of individuals in the dataset are male?  correct

0.7733333 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q8: Conditional Statements - 2**

1/1 point (ungraded)

How many individuals are taller than 78 inches (roughly 2 meters)?  correct

9 Loading

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q9: Conditional Statements - 3**

1/1 point (ungraded)

How many females in the dataset are taller than 78 inches?  correct

2 Loading

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

## Section 1: Introduction to Machine Learning

## Comprehension Check: Introduction to Machine Learning

 Bookmark this page

### **Q1**

1/1 point (graded)

True or False: A key feature of machine learning is that the algorithms are built with **data**.

True

False

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q2**

1/1 point (graded)

True or False: In machine learning, we build algorithms that take feature values (X) and train a model using known outcomes (Y) that is then used to predict outcomes when presented with features without known outcomes.

True

False

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 2: Machine Learning Basics

## 2.1 Basics of Evaluating Machine Learning Basics

## Comprehension Check: Basics of Evaluating Machine Learning Algorithms

 Bookmark this page

### **Q1**

2/2 points (graded)

For each of the following, indicate whether the outcome is continuous or categorical.

Digit reader                           

correct

Height                           

correct

Spam filter                           

correct

Stock prices                           

correct

Sex                           

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (2/2 points)

Review

### **Q2**

0/1 point (graded)

How many features are available to us for prediction in the mnist digits dataset?

You can download the mnist dataset using the read\_mnist() function from the **dslabs** package.

  incorrect

784

70000 Loading

**Explanation**

One way to figure out the number of features available for prediction is to download mnist and then use the ncol function, like this:

mnist <- read\_mnist()

ncol(mnist$train$images)

Submit

You have used 5 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

Answers are displayed within the problem

## Comprehension Check: Practice with Machine Learning, Part 1

 Bookmark this page

The following questions all ask you to work with the dataset described below.

The reported\_heights and heights datasets were collected from three classes taught in the Departments of Computer Science and Biostatistics, as well as remotely through the Extension School. The Biostatistics class was taught in 2016 along with an online version offered by the Extension School. On 2016-01-25 at 8:15 AM, during one of the lectures, the instructors asked student to fill in the sex and height questionnaire that populated the reported\_heights dataset. The online students filled out the survey during the next few days, after the lecture was posted online. We can use this insight to define a variable which we will call type, to denote the type of student, inclass or online.

The code below sets up the dataset for you to analyze in the following exercises:

library(dslabs)

library(dplyr)

library(lubridate)

data(reported\_heights)

dat <- mutate(reported\_heights, date\_time = ymd\_hms(time\_stamp)) %>%

filter(date\_time >= make\_date(2016, 01, 25) & date\_time < make\_date(2016, 02, 1)) %>%

mutate(type = ifelse(day(date\_time) == 25 & hour(date\_time) == 8 & between(minute(date\_time), 15, 30), "inclass","online")) %>%

select(sex, type)

y <- factor(dat$sex, c("Female", "Male"))

x <- dat$type

### **Q1**

2/2 points (graded)

The type column of dat indicates whether students took classes in person ("inclass") or online ("online"). What proportion of the inclass group is female? What proportion of the online group is female?

Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.

In class  correct

66% Loading

Online  correct

37% Loading

Submit

You have used 2 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q2**

1/1 point (graded)

In the course videos, height cutoffs were used to predict sex. Instead of using height, use the type variable to predict sex. Use what you learned in Q1 to make an informed guess about sex based on the most prevalent sex for each type. Report the accuracy of your prediction of sex. You do not need to split the data into training and test sets.

Enter your accuracy as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.  correct

63% Loading

Submit

You have used 3 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

### **Q3**

1/1 point (graded)

Write a line of code using the table() function to show the confusion matrix between y\_hat and y. Use the **exact** format function(a, b) (note the spacing!) for your answer and do not name the columns and rows.

Type the line of code below:  correct

Submit

You have used 2 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q4**

1/1 point (graded)

What is the sensitivity of this prediction? You can use the sensitivity() function from the **caret** package. Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.  correct

38% Loading

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q5**

1/1 point (graded)

What is the specificity of this prediction? You can use the specificity() function from the **caret** package. Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.  correct

84% Loading

Submit

You have used 2 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q6**

1/1 point (graded)

What is the prevalence (% of females) in the dat dataset defined above? Enter your answer as a percentage or decimal (eg "50%" or "0.50") to at least the hundredths place.  correct

45% Loading

Submit

You have used 1 of 5 attempts

## Comprehension Check: Practice with Machine Learning, Part 2

 Bookmark this page

We will practice building a machine learning algorithm using a new dataset, iris, that provides multiple predictors for us to use to train. To start, we will remove the setosa species and we will focus on the versicolor and virginica iris species using the following code:

library(caret)

data(iris)

iris <- iris[-which(iris$Species=='setosa'),]

y <- iris$Species

The following questions all involve work with this dataset.

### **Q7**

1/1 point (graded)

First let us create an even split of the data into train and test partitions using createDataPartition() from the **caret** package. The code with a missing line is given below:

set.seed(2) # if using R 3.6 or later, use set.seed(2, sample.kind="Rounding")

# line of code

test <- iris[test\_index,]

train <- iris[-test\_index,]

Which code should be used in place of # line of code above?

test\_index <- createDataPartition(y,times=1,p=0.5)

test\_index <- sample(2,length(y),replace=FALSE)

test\_index <- createDataPartition(y,times=1,p=0.5,list=FALSE)

test\_index <- rep(1,length(y))

correct

**Answer**

Correct:

Good choice! The createDataPartition function has a number of parameters that allow the user to specify a test/training partition by the percentage of data that goes to training. Indexes should be created on the outcome and not a predictor. See the associated help file.

Note: for this question, you may ignore any warning message generated by the code. If you have R 3.6 or later, you should always use the sample.kind argument in set.seed for this course.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

### **Q8**

1/1 point (graded)

Next we will figure out the singular feature in the dataset that yields the greatest overall accuracy when predicting species. You can use the code from the introduction and from Q7 to start your analysis.

Using only the train iris dataset, for each feature, perform a simple search to find the cutoff that produces the highest accuracy, predicting virginica if greater than the cutoff and versicolor otherwise. Use the seq function over the range of each feature by intervals of 0.1 for this search.

Which feature produces the highest accuracy?

Sepal.Length

Sepal.Width

Petal.Length

Petal.Width

correct

Note: if there are multiple cutoffs that produce the highest accuracy, please use the first/smallest one. Make sure you are setting the seed correctly for your R version in Q7.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Q9**

1/1 point (graded)

For the feature selected in Q8, use the smart cutoff value from the training data to calculate overall accuracy in the test data. What is the overall accuracy?  correct

0.9 Loading

Submit

You have used 2 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q10**

1/1 point (graded)

Notice that we had an overall accuracy greater than 96% in the training data, but the overall accuracy was lower in the test data. This can happen often if we overtrain. In fact, it could be the case that a single feature is not the best choice. For example, a combination of features might be optimal. Using a single feature and optimizing the cutoff as we did on our training data can lead to overfitting.

Given that we know the test data, we can treat it like we did our training data to see if the same feature with a different cutoff will optimize our predictions. Redo the analysis from Q8, this time using the test set.

Which feature best optimizes our overall accuracy?

Sepal.Length

Sepal.Width

Petal.Length

Petal.Width

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q11**

0/1 point (graded)

Now we will perform some exploratory data analysis on the data.

plot(iris,pch=21,bg=iris$Species)

Notice that Petal.Length and Petal.Width in combination could potentially be more information than either feature alone.

Optimize the the cutoffs for Petal.Length and Petal.Width separately in the train dataset by using the seq function with increments of 0.1. Then, report the overall accuracy when applied to the test dataset by creating a rule that predicts virginica if Petal.Length is greater than the length cutoff OR Petal.Width is greater than the width cutoff, and versicolor otherwise.

What is the overall accuracy for the test data now?  incorrect

0.94 Loading

Submit

You have used 5 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 2.2 Conditional Probability

## Comprehension Check: Conditional Probabilities Part 1

 Bookmark this page

### **Q1**

1/1 point (graded)

In a previous module, we covered Bayes' theorem and the Bayesian paradigm. Conditional probabilities are a fundamental part of this previous covered rule.

P(A|B)=P(B|A)P(A)P(B)

We first review a simple example to go over conditional probabilities.

Assume a patient comes into the doctor’s office to test whether they have a particular disease.

* The test is positive 85% of the time when tested on a patient with the disease (high sensitivity): P(test+|disease)=0.85
* The test is negative 90% of the time when tested on a healthy patient (high specificity): P(test−|heathy)=0.90
* The disease is prevalent in about 2% of the community: P(disease)=0.02

Using Bayes' theorem, calculate the probability that you have the disease if the test is positive.

Enter your answer as a percentage or decimal (eg "50%" or "0.50").

  correct

14% Loading

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

Correct (1/1 point)

Review

The following 4 questions (Q2-Q5) all relate to implementing this calculation using R.

We have a hypothetical population of 1 million individuals with the following conditional probabilities as described below:

* + The test is positive 85% of the time when tested on a patient with the disease (high sensitivity): P(test+|disease)=0.85
  + The test is negative 90% of the time when tested on a healthy patient (high specificity): P(test−|heathy)=0.90
  + The disease is prevalent in about 2% of the community: P(disease)=0.02

Here is some sample code to get you started:

set.seed(1) # set.seed(1, sample.kind="Rounding") if using R 3.6 or later

disease <- sample(c(0,1), size=1e6, replace=TRUE, prob=c(0.98,0.02))

test <- rep(NA, 1e6)

test[disease==0] <- sample(c(0,1), size=sum(disease==0), replace=TRUE, prob=c(0.90,0.10))

test[disease==1] <- sample(c(0,1), size=sum(disease==1), replace=TRUE, prob=c(0.15, 0.85))

### **Q2**

1/1 point (graded)

What is the probability that a test is positive?  correct

0.114509

0.115 Loading

**Explanation**

The probability of a positive test can be calculated using mean(test).

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

What is the probability that an individual has the disease if the test is negative?  correct

0.003461356

0.003065 Loading

**Explanation**

The probability of having the disease given a negative test can be calculated using mean(disease[test==0]).

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q4**

1/1 point (graded)

What is the probability that you have the disease if the test is positive?

Remember: calculate the conditional probability the disease is positive assuming a positive test.

  correct

0.1471762

0.1471762 Loading

**Explanation**

The probability of having the disease given a positive test can be calculated using mean(disease[test==1]==1).

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q5**

1/1 point (graded)

Compare the prevalence of disease in people who test positive to the overall prevalence of disease.

If a patient's test is positive, how much does that increase their risk of having the disease?

First calculate the probability of having the disease given a positive test, then divide by the probability of having the disease.

  correct

7.389106

7.389106 Loading

**Explanation**

The increase in risk can be calculated using mean(disease[test==1]==1)/mean(disease==1).

Submit

You have used 1 of 5 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

## Comprehension Check: Conditional Probabilities Part 2

 Bookmark this page

### **Q6**

1/1 point (graded)

We are now going to write code to compute conditional probabilities for being male in the heights dataset. Round the heights to the closest inch. Plot the estimated conditional probability P(x)=Pr(Male|height=x) for each x.

Part of the code is provided here:

library(dslabs)

data("heights")

# MISSING CODE

qplot(height, p, data =.)

Which of the following blocks of code can be used to replace **# MISSING CODE** to make the correct plot?



heights %>%

group\_by(height) %>%

summarize(p = mean(sex == "Male")) %>%



heights %>%

mutate(height = round(height)) %>%

group\_by(height) %>%

summarize(p = mean(sex == "Female")) %>%



heights %>%

mutate(height = round(height)) %>%

summarize(p = mean(sex == "Male")) %>%



heights %>%

mutate(height = round(height)) %>%

group\_by(height) %>%

summarize(p = mean(sex == "Male")) %>%

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q7**

1/1 point (graded)

In the plot we just made in Q6 we see high variability for low values of height. This is because we have few data points. This time use the quantile 0.1,0.2,…,0.9 and the cut() function to assure each group has the same number of points. Note that for any numeric vector x, you can create groups based on quantiles like this: cut(x, quantile(x, seq(0, 1, 0.1)), include.lowest = TRUE).

Part of the code is provided here:

ps <- seq(0, 1, 0.1)

heights %>%

# MISSING CODE

group\_by(g) %>%

summarize(p = mean(sex == "Male"), height = mean(height)) %>%

qplot(height, p, data =.)

Which of the following lines of code can be used to replace **# MISSING CODE** to make the correct plot?



mutate(g = cut(male, quantile(height, ps), include.lowest = TRUE)) %>%



mutate(g = cut(height, quantile(height, ps), include.lowest = TRUE)) %>%



mutate(g = cut(female, quantile(height, ps), include.lowest = TRUE)) %>%



mutate(g = cut(height, quantile(height, ps))) %>%

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q8**

1/1 point (graded)

You can generate data from a bivariate normal distrubution using the **MASS** package using the following code:

Sigma <- 9\*matrix(c(1,0.5,0.5,1), 2, 2)

dat <- MASS::mvrnorm(n = 10000, c(69, 69), Sigma) %>%

data.frame() %>% setNames(c("x", "y"))

And you can make a quick plot using plot(dat).

Using an approach similar to that used in the previous exercise, let's estimate the conditional expectations and make a plot. Part of the code has again been provided for you:

ps <- seq(0, 1, 0.1)

dat %>%

# MISSING CODE

qplot(x, y, data =.)

Which of the following blocks of code can be used to replace **# MISSING CODE** to make the correct plot?



mutate(g = cut(x, quantile(x, ps), include.lowest = TRUE)) %>%

group\_by(g) %>%

summarize(y = mean(y), x = mean(x)) %>%



mutate(g = cut(x, quantile(x, ps))) %>%

group\_by(g) %>%

summarize(y = mean(y), x = mean(x)) %>%



mutate(g = cut(x, quantile(x, ps), include.lowest = TRUE)) %>%

summarize(y = mean(y), x = mean(x)) %>%



mutate(g = cut(x, quantile(x, ps), include.lowest = TRUE)) %>%

group\_by(g) %>%

summarize(y =(y), x =(x)) %>%

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## Section 3: Linear Regression for Prediction, Smoothing, and Working with Matrices

## 3.1 Linear Regression for Predictions

## Comprehension Check: Linear Regression

 Bookmark this page

### **Q1**

2/2 points (graded)

Create a data set using the following code:

set.seed(1) # set.seed(1, sample.kind="Rounding") if using R 3.6 or later

n <- 100

Sigma <- 9\*matrix(c(1.0, 0.5, 0.5, 1.0), 2, 2)

dat <- MASS::mvrnorm(n = 100, c(69, 69), Sigma) %>%

data.frame() %>% setNames(c("x", "y"))

We will build 100 linear models using the data above and calculate the mean and standard deviation of the combined models. First, **set the seed to 1 again** (make sure to use sample.kind="Rounding" if your R is version 3.6 or later). Then, within a replicate() loop, (1) partition the dataset into test and training sets with p=0.5 and using dat$y to generate your indices, (2) train a linear model predicting y from x, (3) generate predictions on the test set, and (4) calculate the RMSE of that model. Then, report the mean and standard deviation (SD) of the RMSEs from all 100 models.

Report all answers to at least 3 decimal places.

Mean:  correct

2.49 Loading

Standard deviation (SD):  correct

0.124 Loading

Submit

You have used 10 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

### **Q2**

0.4/2 points (graded)

Now we will repeat the exercise above but using larger datasets. Write a function that takes a size n, then (1) builds a dataset using the code provided at the top of Q1 but with n observations instead of 100 and without the set.seed(1), (2) runs the replicate() loop that you wrote to answer Q1, which builds **100 linear models** and returns a vector of RMSEs, and (3) calculates the mean and standard deviation of the 100 RMSEs.

Set the seed to 1 (if using R 3.6 or later, use the argument sample.kind="Rounding") and then use sapply() or map() to apply your new function to n <- c(100, 500, 1000, 5000, 10000).

You only need to set the seed once before running your function; do not set a seed within your function. Also be sure to use sapply() or map() as you will get different answers running the simulations individually due to setting the seed.

Mean, 100:  correct

2.498

2.49 Loading

SD, 100:  correct

0.118

0.118 Loading

Mean, 500:  incorrect

2.72

2.49 Loading

SD, 500:  incorrect

0.08

0.1289 Loading

Mean, 1000:  incorrect

2.5555

2.48 Loading

SD, 1000:  incorrect

0.0456

0.130 Loading

Mean, 5000:  incorrect

2.6248

2.478 Loading

SD, 5000:  incorrect

0.0231

0.125 Loading

Mean, 10000:  incorrect

2.6184

2.478 Loading

SD, 10000:  incorrect

0.0169

0.127 Loading

**Explanation**

The code below can be used to do this calculation:

set.seed(1) # if R 3.6 or later, set.seed(1, sample.kind="Rounding")

n <- c(100, 500, 1000, 5000, 10000)

res <- sapply(n, function(n){

Sigma <- 9\*matrix(c(1.0, 0.5, 0.5, 1.0), 2, 2)

dat <- MASS::mvrnorm(n, c(69, 69), Sigma) %>%

data.frame() %>% setNames(c("x", "y"))

rmse <- replicate(100, {

test\_index <- createDataPartition(dat$y, times = 1, p = 0.5, list = FALSE)

train\_set <- dat %>% slice(-test\_index)

test\_set <- dat %>% slice(test\_index)

fit <- lm(y ~ x, data = train\_set)

y\_hat <- predict(fit, newdata = test\_set)

sqrt(mean((y\_hat-test\_set$y)^2))

})

c(avg = mean(rmse), sd = sd(rmse))

})

res

Submit

You have used 10 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

What happens to the RMSE as the size of the dataset becomes larger?

On average, the RMSE does not change much as n gets larger, but the variability of the RMSE decreases.

Because of the law of large numbers the RMSE decreases; more data means more precise estimates.

n = 10000 is not sufficiently large. To see a decrease in the RMSE we would need to make it larger.

The RMSE is not a random variable.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q4**

2/2 points (graded)

Now repeat the exercise from Q1, this time making the correlation between x and y larger, as in the following code:

set.seed(1)

n <- 100

Sigma <- 9\*matrix(c(1.0, 0.95, 0.95, 1.0), 2, 2)

dat <- MASS::mvrnorm(n = 100, c(69, 69), Sigma) %>%

data.frame() %>% setNames(c("x", "y"))

Note what happens to RMSE - set the seed to 1 as before.

Mean:  correct

0.91

0.90 Loading

SD:  correct

0.0624

0.062 Loading

**Explanation**

The same code as in Q1 can be used:

set.seed(1)

rmse <- replicate(100, {

test\_index <- createDataPartition(dat$y, times = 1, p = 0.5, list = FALSE)

train\_set <- dat %>% slice(-test\_index)

test\_set <- dat %>% slice(test\_index)

fit <- lm(y ~ x, data = train\_set)

y\_hat <- predict(fit, newdata = test\_set)

sqrt(mean((y\_hat-test\_set$y)^2))

})

mean(rmse)

sd(rmse)

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q5**

1/1 point (graded)

Which of the following best explains why the RMSE in question 4 is so much lower than the RMSE in question 1?

It is just luck. If we do it again, it will be larger.

The central limit theorem tells us that the RMSE is normal.

When we increase the correlation between x and y, x has more predictive power and thus provides a better estimate of y.

These are both examples of regression so the RMSE has to be the same.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q6**

1/1 point (graded)

Create a data set using the following code.

set.seed(1)

Sigma <- matrix(c(1.0, 0.75, 0.75, 0.75, 1.0, 0.25, 0.75, 0.25, 1.0), 3, 3)

dat <- MASS::mvrnorm(n = 100, c(0, 0, 0), Sigma) %>%

data.frame() %>% setNames(c("y", "x\_1", "x\_2"))

Note that y is correlated with both x\_1 and x\_2 but the two predictors are independent of each other, as seen by cor(dat).

Set the seed to 1, then use the **caret** package to partition into a test and training set of equal size. Compare the RMSE when using just x\_1, just x\_2 and both x\_1 and x\_2. Train a single linear model for each (not 100 like in the previous questions).

Which of the three models performs the best (has the lowest RMSE)?

x\_1

x\_2

x\_1 and x\_2

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q7**

1/1 point (graded)

Report the lowest RMSE of the three models tested in Q6.  correct

0.307 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q8**

1/1 point (graded)

Repeat the exercise from Q6 but now create an example in which x\_1 and x\_2 are highly correlated.

set.seed(1)

Sigma <- matrix(c(1.0, 0.75, 0.75, 0.75, 1.0, 0.95, 0.75, 0.95, 1.0), 3, 3)

dat <- MASS::mvrnorm(n = 100, c(0, 0, 0), Sigma) %>%

data.frame() %>% setNames(c("y", "x\_1", "x\_2"))

Set the seed to 1, then use the **caret** package to partition into a test and training set of equal size. Compare the RMSE when using just x\_1, just x\_2, and both x\_1 and x\_2.

Compare the results from Q6 and Q8. What can you conclude?

Unless we include all predictors we have no predictive power.

Adding extra predictors improves RMSE regardless of whether the added predictors are correlated with other predictors or not.

Adding extra predictors results in over fitting.

Adding extra predictors can improve RMSE substantially, but not when the added predictors are highly correlated with other predictors.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## Comprehension Check: Logistic Regression

 Bookmark this page

### **Q1**

1/1 point (graded)

Define a dataset using the following code:

set.seed(2) #if you are using R 3.5 or earlier

set.seed(2, sample.kind="Rounding") #if you are using R 3.6 or later

make\_data <- function(n = 1000, p = 0.5,

mu\_0 = 0, mu\_1 = 2,

sigma\_0 = 1, sigma\_1 = 1){

y <- rbinom(n, 1, p)

f\_0 <- rnorm(n, mu\_0, sigma\_0)

f\_1 <- rnorm(n, mu\_1, sigma\_1)

x <- ifelse(y == 1, f\_1, f\_0)

test\_index <- createDataPartition(y, times = 1, p = 0.5, list = FALSE)

list(train = data.frame(x = x, y = as.factor(y)) %>% slice(-test\_index),

test = data.frame(x = x, y = as.factor(y)) %>% slice(test\_index))

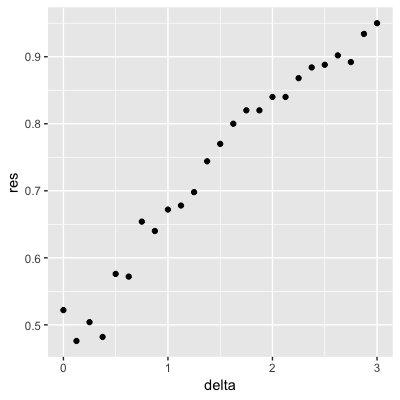
}

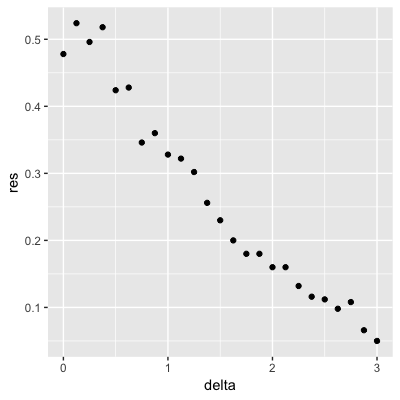
dat <- make\_data()

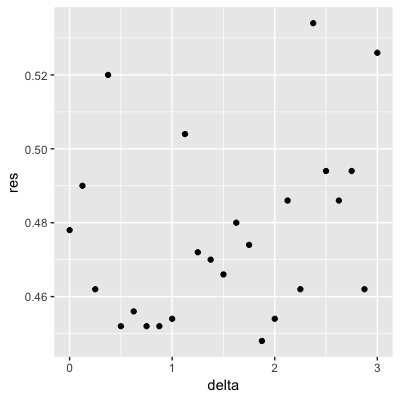
Note that we have defined a variable x that is predictive of a binary outcome y: dat$train %>% ggplot(aes(x, color = y)) + geom\_density().

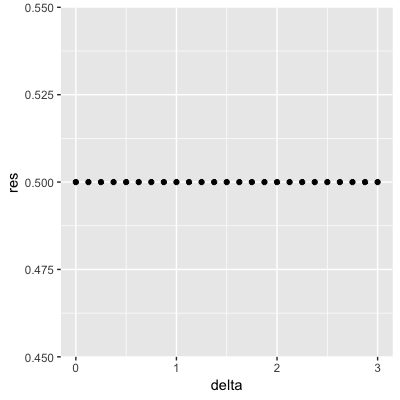
Set the seed to 1, then use the make\_data() function defined above to generate 25 different datasets with mu\_1 <- seq(0, 3, len=25). Perform logistic regression on each of the 25 different datasets (predict 1 if p>0.5) and plot accuracy (res in the figures) vs mu\_1 (delta in the figures).”

Which is the correct plot?









correct

**Explanation**

The correct plot can be generated using the following code:

set.seed(1) #if you are using R 3.5 or earlier

set.seed(1, sample.kind="Rounding") #if you are using R 3.6 or later

delta <- seq(0, 3, len = 25)

res <- sapply(delta, function(d){

dat <- make\_data(mu\_1 = d)

fit\_glm <- dat$train %>% glm(y ~ x, family = "binomial", data = .)

y\_hat\_glm <- ifelse(predict(fit\_glm, dat$test) > 0.5, 1, 0) %>% factor(levels = c(0, 1))

mean(y\_hat\_glm == dat$test$y)

})

qplot(delta, res)

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 3.2 Smoothing

## Comprehension Check: Smoothing

 Bookmark this page

### **Q1**

1/1 point (graded)

In the Wrangling course of this series, PH125.6x, we used the following code to obtain mortality counts for Puerto Rico for 2015-2018:

library(tidyverse)

library(lubridate)

library(purrr)

library(pdftools)

fn <- system.file("extdata", "RD-Mortality-Report\_2015-18-180531.pdf", package="dslabs")

dat <- map\_df(str\_split(pdf\_text(fn), "\n"), function(s){

s <- str\_trim(s)

header\_index <- str\_which(s, "2015")[1]

tmp <- str\_split(s[header\_index], "\\s+", simplify = TRUE)

month <- tmp[1]

header <- tmp[-1]

tail\_index <- str\_which(s, "Total")

n <- str\_count(s, "\\d+")

out <- c(1:header\_index, which(n==1), which(n>=28), tail\_index:length(s))

s[-out] %>%

str\_remove\_all("[^\\d\\s]") %>%

str\_trim() %>%

str\_split\_fixed("\\s+", n = 6) %>%

.[,1:5] %>%

as\_data\_frame() %>%

setNames(c("day", header)) %>%

mutate(month = month,

day = as.numeric(day)) %>%

gather(year, deaths, -c(day, month)) %>%

mutate(deaths = as.numeric(deaths))

}) %>%

mutate(month = recode(month, "JAN" = 1, "FEB" = 2, "MAR" = 3, "APR" = 4, "MAY" = 5, "JUN" = 6,

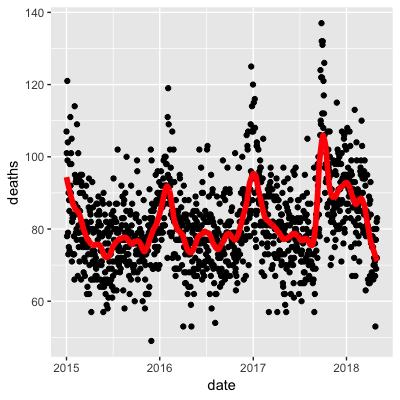
"JUL" = 7, "AGO" = 8, "SEP" = 9, "OCT" = 10, "NOV" = 11, "DEC" = 12)) %>%

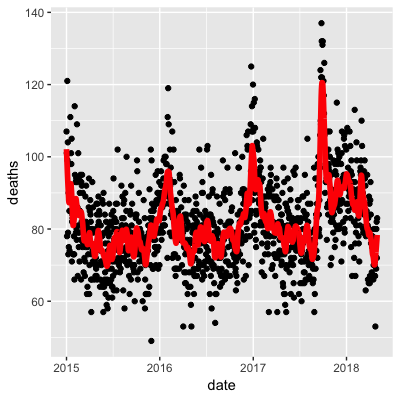
mutate(date = make\_date(year, month, day)) %>%

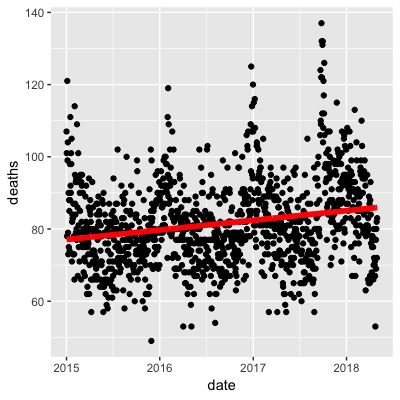
filter(date <= "2018-05-01")

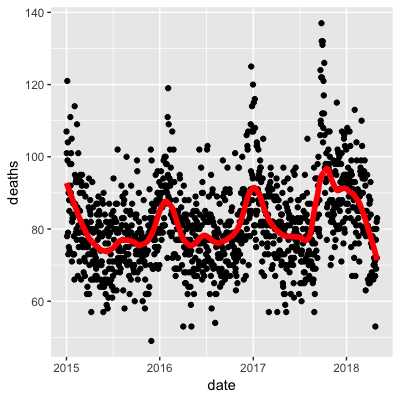
Use the loess() function to obtain a smooth estimate of the expected number of deaths as a function of date. Plot this resulting smooth function. Make the span about two months long.

Which of the following plots is correct?









correct

**Explanation**

The following code makes the correct plot:

span <- 60 / as.numeric(diff(range(dat$date)))

fit <- dat %>% mutate(x = as.numeric(date)) %>% loess(deaths ~ x, data = ., span = span, degree = 1)

dat %>% mutate(smooth = predict(fit, as.numeric(date))) %>%

ggplot() +

geom\_point(aes(date, deaths)) +

geom\_line(aes(date, smooth), lwd = 2, col = "red")

The second plot uses a shorter span, the third plot uses the entire timespan, and the fourth plot uses a longer span.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Q2**

1/1 point (graded)

Work with the same data as in Q1 to plot smooth estimates against day of the year, all on the same plot, but with different colors for each year.

Which code produces the desired plot?



dat %>%

mutate(smooth = predict(fit), day = yday(date), year = as.character(year(date))) %>%

ggplot(aes(day, smooth, col = year)) +

geom\_line(lwd = 2)



dat %>%

mutate(smooth = predict(fit, as.numeric(date)), day = mday(date), year = as.character(year(date))) %>%

ggplot(aes(day, smooth, col = year)) +

geom\_line(lwd = 2)



dat %>%

mutate(smooth = predict(fit, as.numeric(date)), day = yday(date), year = as.character(year(date))) %>%

ggplot(aes(day, smooth)) +

geom\_line(lwd = 2)



dat %>%

mutate(smooth = predict(fit, as.numeric(date)), day = yday(date), year = as.character(year(date))) %>%

ggplot(aes(day, smooth, col = year)) +

geom\_line(lwd = 2)

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q3**

1/1 point (graded)

Suppose we want to predict 2s and 7s in the mnist\_27 dataset with just the second covariate. Can we do this? On first inspection it appears the data does not have much predictive power.

In fact, if we fit a regular logistic regression the coefficient for x\_2 is not significant!

This can be seen using this code:

library(broom)

library(dslabs)

data(mnist\_27)

mnist\_27$train %>% glm(y ~ x\_2, family = "binomial", data = .) %>% tidy()

Plotting a scatterplot here is not useful since y is binary:

qplot(x\_2, y, data = mnist\_27$train)

Fit a loess line to the data above and plot the results. What do you observe?

There is no predictive power and the conditional probability is linear.

There is no predictive power and the conditional probability is non-linear.

There is predictive power and the conditional probability is linear.

There is predictive power and the conditional probability is non-linear.

correct

**Explanation**

Note that there is indeed predictive power, but that the conditional probability is non-linear.

The loess line can be plotted using the following code:

mnist\_27$train %>%

mutate(y = ifelse(y=="7", 1, 0)) %>%

ggplot(aes(x\_2, y)) +

geom\_smooth(method = "loess")

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

## 3.3 Working with Matrices

## Comprehension Check: Working with Matrices

 Bookmark this page

### **Q1**

1/1 point (graded)

Which line of code correctly creates a 100 by 10 matrix of randomly generated normal numbers and assigns it to x?



x <- matrix(rnorm(1000), 100, 100)



x <- matrix(rnorm(100\*10), 100, 10)



x <- matrix(rnorm(100\*10), 10, 10)



x <- matrix(rnorm(100\*10), 10, 100)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q2**

2/2 points (graded)

Write the line of code that would give you the specified information about the matrix x that you generated in q1. Do not include any spaces in your line of code.

Dimension of x  correct

Number of rows of x  correct

Number of columns of x  correct

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Q3**

1/1 point (graded)

Which of the following lines of code would add the scalar 1 to row 1, the scalar 2 to row 2, and so on, for the matrix x?

Select ALL that apply.



x <- x + seq(nrow(x))



x <- 1:nrow(x)



x <- sweep(x, 2, 1:nrow(x),"+")



x <- sweep(x, 1, 1:nrow(x),"+")

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q4**

1/1 point (graded)

Which of the following lines of code would add the scalar 1 to column 1, the scalar 2 to column 2, and so on, for the matrix x?

Select ALL that apply.



x <- 1:ncol(x)



x <- 1:col(x)



x <- sweep(x, 2, 1:ncol(x), FUN = "+")



x <- -x

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q5**

2/2 points (graded)

Which code correctly computes the average of each row of x?



mean(x)



rowMedians(x)



sapply(x,mean)



rowSums(x)



rowMeans(x)

correct

Which code correctly computes the average of each column of x?



mean(x)



sapply(x,mean)



colMeans(x)



colMedians(x)



colSums(x)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Q6**

1/1 point (graded)

For each observation in the mnist training data, compute the proportion of pixels that are in the **grey area**, defined as values between 50 and 205 (but not including 50 and 205). (To visualize this, you can make a boxplot by digit class.)

What proportion of the 60000\*784 pixels in the mnist training data are in the grey area overall, defined as values between 50 and 205?  correct

0.0618

0.0618 Loading

**Explanation**

The matrix and plot can be calculated using the following code:

mnist <- read\_mnist()

y <- rowMeans(mnist$train$images>50 & mnist$train$images<205)

qplot(as.factor(mnist$train$labels), y, geom = "boxplot")

The proportion of pixels can be calculated using mean(y).

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

Answers are displayed within the problem

## [Section 4: Distance, Knn, Cross-validation, and Generative Models](https://courses.edx.org/courses/course-v1:HarvardX+PH125.8x+1T2020/course/#block-v1:HarvardX+PH125.8x+1T2020+type@chapter+block@7c50a2fc6d44433fa902be8ed097f301)

## 4.1 Nearest Neighbours

## Comprehension Check: Distance

 Bookmark this page

### **Q1**

1/1 point (graded)

Load the following dataset:

library(dslabs)

data(tissue\_gene\_expression)

This dataset includes a matrix x:

dim(tissue\_gene\_expression$x)

This matrix has the gene expression levels of 500 genes from 189 biological samples representing seven different tissues. The tissue type is stored in y:

table(tissue\_gene\_expression$y)

Which of the following lines of code computes the Euclidean distance between each observation and stores it in the object d?



d <- dist(tissue\_gene\_expression$x, distance='maximum')



d <- dist(tissue\_gene\_expression)



d <- dist(tissue\_gene\_expression$x)



d <- cor(tissue\_gene\_expression$x)

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q2**

1/1 point (graded)

Using the dataset from Q1, compare the distances between observations 1 and 2 (both cerebellum), observations 39 and 40 (both colon), and observations 73 and 74 (both endometrium).

Distance-wise, are samples from tissues of the same type closer to each other?

No, the samples from the same tissue type are not necessarily closer.

The two colon samples are closest to each other, but the samples from the other two tissues are not.

The two cerebellum samples are closest to each other, but the samples from the other two tissues are not.

Yes, the samples from the same tissue type are closest to each other.

correct

**Explanation**

You can calculate the distances using the following code:

ind <- c(1, 2, 39, 40, 73, 74)

as.matrix(d)[ind,ind]

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

Make a plot of all the distances using the image() function to see if the pattern you observed in Q2 is general.

Which code would correctly make the desired plot?



image(d)



image(as.matrix(d))



d



image()

correct

**Explanation**

When we examine the plot, we do see that the pattern holds and that samples from the same tissue are closest to each other, although there do appear to be some additional close distances between tissue types as well.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Comprehension Check: Nearest Neighbors

 Bookmark this page

### **Q1**

1/2 points (graded)

Previously, we used logistic regression to predict sex based on height. Now we are going to use knn to do the same. Set the seed to 1, then use the **caret** package to partition the **dslabs** heights data into a training and test set of equal size. Use the sapply() or map function to perform knn with k values of seq(1, 101, 3) and calculate F1 scores with the F\_meas() function using the default value of the relevant argument.

What is the max value of F\_1?  correct

0.6190476

At what value of k does the max occur?

If there are multiple values of k with the maximum value, report the smallest such k.

  incorrect

52

Submit

You have used 10 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

### **Q2**

2.0/2.0 points (graded)

Next we will use the same gene expression example used in the Comprehension Check: Distance exercises. You can load it like this:

library(dslabs)

data("tissue\_gene\_expression")

First, set the seed to 1 and split the data into training and test sets. Then, report the accuracy you obtain from predicting tissue type using KNN with k = 1, 3, 5, 7, 9, 11 using sapply() or map\_df(). Note: use the createDataPartition() function outside of sapply() or map\_df().

k=1  correct

0.980

k=3  correct

0.969

k=5  correct

0.947

k=7  correct

0.916

k=9  correct

0.916

k=11  correct

0.906

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 4.2 Cross Validation

## Comprehension Check: Cross-validation

 Bookmark this page

### **Q1**

1/1 point (graded)

Generate a set of random predictors and outcomes using the following code:

set.seed(1996) #if you are using R 3.5 or earlier

set.seed(1996, sample.kind="Rounding") #if you are using R 3.6 or later

n <- 1000

p <- 10000

x <- matrix(rnorm(n\*p), n, p)

colnames(x) <- paste("x", 1:ncol(x), sep = "\_")

y <- rbinom(n, 1, 0.5) %>% factor()

x\_subset <- x[ ,sample(p, 100)]

Because x and y are completely independent, you should not be able to predict y using x with accuracy greater than 0.5. Confirm this by running cross-validation using logistic regression to fit the model. Because we have so many predictors, we selected a random sample x\_subset. Use the subset when training the model.

Which code correctly performs this cross-validation?



fit <- train(x\_subset, y)

fit$results



fit <- train(x\_subset, y, method = "glm")

fit$results



fit <- train(y, x\_subset, method = "glm")

fit$results



fit <- test(x\_subset, y, method = "glm")

fit$results

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint Show Answer

Correct (1/1 point)

Review

### **Q2**

1/1 point (graded)

Now, instead of using a random selection of predictors, we are going to search for those that are most predictive of the outcome. We can do this by comparing the values for the y=1 group to those in the y=0 group, for each predictor, using a t-test. You can do perform this step like this:

install.packages("BiocManager")

BiocManager::install("genefilter")

library(genefilter)

tt <- colttests(x, y)

Which of the following lines of code correctly creates a vector of the p-values called pvals?



pvals <- tt$dm



pvals <- tt$statistic



pvals <- tt



pvals <- tt$p.value

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q3**

1/1 point (graded)

Create an index ind with the column numbers of the predictors that were "statistically significantly" associated with y. Use a p-value cutoff of 0.01 to define "statistically significantly."

How many predictors survive this cutoff?  correct

108

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q4**

1/1 point (graded)

Now re-run the cross-validation after redefinining x\_subset to be the subset of x defined by the columns showing "statistically significant" association with y.

What is the accuracy now?  correct

0.7597842

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q5**

1/1 point (graded)

Re-run the cross-validation again, but this time using kNN. Try out the following grid k = seq(101, 301, 25) of tuning parameters. Make a plot of the resulting accuracies.

Which code is correct?



fit <- train(x\_subset, y, method = "knn", tuneGrid = data.frame(k = seq(101, 301, 25)))

ggplot(fit)



fit <- train(x\_subset, y, method = "knn")

ggplot(fit)



fit <- train(x\_subset, y, method = "knn", tuneGrid = data.frame(k = seq(103, 301, 25)))

ggplot(fit)



fit <- train(x\_subset, y, method = "knn", tuneGrid = data.frame(k = seq(101, 301, 5)))

ggplot(fit)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q6**

1/1 point (graded)

In the previous exercises, we see that despite the fact that x and y are completely independent, we were able to predict y with accuracy higher than 70%. We must be doing something wrong then.

What is it?

The function train() estimates accuracy on the same data it uses to train the algorithm.

We are overfitting the model by including 100 predictors.

We used the entire dataset to select the columns used in the model.

The high accuracy is just due to random variability.

correct

**Explanation**

Because we used the entire dataset to select the columns in the model, the accuracy is too high. The selection step needs to be included as part of the cross-validation algorithm, and then the cross-validation itself is performed **after** the column selection step.

As a follow-up exercise, try to re-do the cross-validation, this time including the selection step in the cross-validation algorithm. The accuracy should now be close to 50%.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Q7**

1/1 point (graded)

Use the train() function with kNN to select the best k for predicting tissue from gene expression on the tissue\_gene\_expression dataset from **dslabs**. Try k = seq(1,7,2) for tuning parameters. For this question, do not split the data into test and train sets (understand this can lead to overfitting, but ignore this for now).

What value of k results in the highest accuracy?  correct

1

1

**Explanation**

The following code will allow you to pick the best value of k:

data("tissue\_gene\_expression")

fit <- with(tissue\_gene\_expression, train(x, y, method = "knn", tuneGrid = data.frame( k = seq(1, 7, 2))))

ggplot(fit)

fit$results

Submit

You have used 4 of 10 attempts

## 4.3 Generative Models

## Comprehension Check: Generative Models

 Bookmark this page

In the following exercises, we are going to apply LDA and QDA to the tissue\_gene\_expression dataset from **dslabs.** We will start with simple examples based on this dataset and then develop a realistic example.

### **Q1**

1/1 point (graded)

Create a dataset of samples from just cerebellum and hippocampus, two parts of the brain, and a predictor matrix with 10 randomly selected columns using the following code:

library(dslabs)

library(caret)

library(tidyverse)

data("tissue\_gene\_expression")

set.seed(1993) #if using R 3.6 or later set.seed(1993, sample.kind="Rounding")

ind <- which(tissue\_gene\_expression$y %in% c("cerebellum", "hippocampus"))

y <- droplevels(tissue\_gene\_expression$y[ind])

x <- tissue\_gene\_expression$x[ind, ]

x <- x[, sample(ncol(x), 10)]

Use the train() function to estimate the accuracy of LDA. For this question, use the version of x and y created with the code above: do not split them or tissue\_gene\_expression into training and test sets (understand this can lead to overfitting). Report the accuracy from the train() results (do not make predictions).

What is the accuracy?  correct

0.871

0.8707879

**Explanation**

The following code can be used to estimate the accuracy of the LDA:

fit\_lda <- train(x, y, method = "lda")

fit\_lda$results["Accuracy"]

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q2**

1/1 point (graded)

In this case, LDA fits two 10-dimensional normal distributions. Look at the fitted model by looking at the finalModel component of the result of train(). Notice there is a component called means that includes the estimated means of both distributions. Plot the mean vectors against each other and determine which predictors (genes) appear to be driving the algorithm.

Which TWO genes appear to be driving the algorithm (i.e. the two genes with the highest means)?

PLCB1

RAB1B

MSH4

OAZ2

SPI1

SAPCD1

HEMK1

correct

**Explanation**

The following code can be used to make the plot:

t(fit\_lda$finalModel$means) %>% data.frame() %>%

mutate(predictor\_name = rownames(.)) %>%

ggplot(aes(cerebellum, hippocampus, label = predictor\_name)) +

geom\_point() +

geom\_text() +

geom\_abline()

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

Repeat the exercise in Q1 with QDA.

Create a dataset of samples from just cerebellum and hippocampus, two parts of the brain, and a predictor matrix with 10 randomly selected columns using the following code:

library(dslabs)

library(caret)

data("tissue\_gene\_expression")

set.seed(1993) #set.seed(1993, sample.kind="Rounding") if using R 3.6 or later

ind <- which(tissue\_gene\_expression$y %in% c("cerebellum", "hippocampus"))

y <- droplevels(tissue\_gene\_expression$y[ind])

x <- tissue\_gene\_expression$x[ind, ]

x <- x[, sample(ncol(x), 10)]

Use the train() function to estimate the accuracy of QDA. For this question, use the entire tissue\_gene\_expression dataset: do not split it into training and test sets (understand this can lead to overfitting).

What is the accuracy?  correct

0.8109429

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q4**

1/1 point (graded)

Which TWO genes drive the algorithm when using QDA instead of LDA (i.e. the two genes with the highest means)?

PLCB1

RAB1B

MSH4

OAZ2

SPI1

SAPCD1

HEMK1

correct

**Explanation**

t(fit\_qda$finalModel$means) %>% data.frame() %>%

mutate(predictor\_name = rownames(.)) %>%

ggplot(aes(cerebellum, hippocampus, label = predictor\_name)) +

geom\_point() +

geom\_text() +

geom\_abline()

The following code can be used to make the plot to evaluate which genes are driving the algorithm:

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q5**

1/1 point (graded)

One thing we saw in the previous plots is that the values of the predictors correlate in both groups: some predictors are low in both groups and others high in both groups. The mean value of each predictor found in colMeans(x) is not informative or useful for prediction and often for purposes of interpretation, it is useful to center or scale each column. This can be achieved with the preProcess argument in train(). Re-run LDA with preProcess = "center". Note that accuracy does not change, but it is now easier to identify the predictors that differ more between groups than based on the plot made in Q2.

Which TWO genes drive the algorithm after performing the scaling?

C21orf62

PLCB1

RAB1B

MSH4

OAZ2

SPI1

SAPCD1

IL18R1

correct

**Explanation**

The following code can be used to make the plot to evaluate which genes are driving the algorithm after scaling:

fit\_lda <- train(x, y, method = "lda", preProcess = "center")

fit\_lda$results["Accuracy"]

t(fit\_lda$finalModel$means) %>% data.frame() %>%

mutate(predictor\_name = rownames(.)) %>%

ggplot(aes(predictor\_name, hippocampus)) +

geom\_point() +

coord\_flip()

You can see that it is different genes driving the algorithm now. This is because the predictor means change.

In the previous exercises we saw that both LDA and QDA approaches worked well. For further exploration of the data, you can plot the predictor values for the two genes with the largest differences between the two groups in a scatter plot to see how they appear to follow a bivariate distribution as assumed by the LDA and QDA approaches, coloring the points by the outcome, using the following code:

d <- apply(fit\_lda$finalModel$means, 2, diff)

ind <- order(abs(d), decreasing = TRUE)[1:2]

plot(x[, ind], col = y)

Submit

You have used 2 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q6**

1/1 point (graded)

Now we are going to increase the complexity of the challenge slightly. Repeat the LDA analysis from Q5 but using all tissue types. Use the following code to create your dataset:

library(dslabs)

library(caret)

data("tissue\_gene\_expression")

set.seed(1993) #set.seed(1993, sample.kind="Rounding") if using R 3.6 or later

y <- tissue\_gene\_expression$y

x <- tissue\_gene\_expression$x

x <- x[, sample(ncol(x), 10)]

What is the accuracy using LDA?  correct

0.8194837

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## [Section 5: Classification with More than Two Classes and the Caret Package](https://courses.edx.org/courses/course-v1:HarvardX+PH125.8x+1T2020/course/#block-v1:HarvardX+PH125.8x+1T2020+type@chapter+block@3f70797a568946e195863415b94d25fd)

## [5.1: Classification with More than Two Classes](https://courses.edx.org/courses/course-v1:HarvardX+PH125.8x+1T2020/course/#block-v1:HarvardX+PH125.8x+1T2020+type@sequential+block@78cfd6af400d452c9aa9f8070403e55c)

## Comprehension Check: Trees and Random Forests

 Bookmark this page

### **Q1**

1/1 point (graded)

Create a simple dataset where the outcome grows 0.75 units on average for every increase in a predictor, using this code:

library(rpart)

n <- 1000

sigma <- 0.25

set.seed(1) #set.seed(1, sample.kind = "Rounding") if using R 3.6 or later

x <- rnorm(n, 0, 1)

y <- 0.75 \* x + rnorm(n, 0, sigma)

dat <- data.frame(x = x, y = y)

Which code correctly uses rpart() to fit a regression tree and saves the result to fit?



fit <- rpart(y ~ .)



fit <- rpart(y, ., data = dat)



fit <- rpart(x ~ ., data = dat)



fit <- rpart(y ~ ., data = dat)

correct

Submit

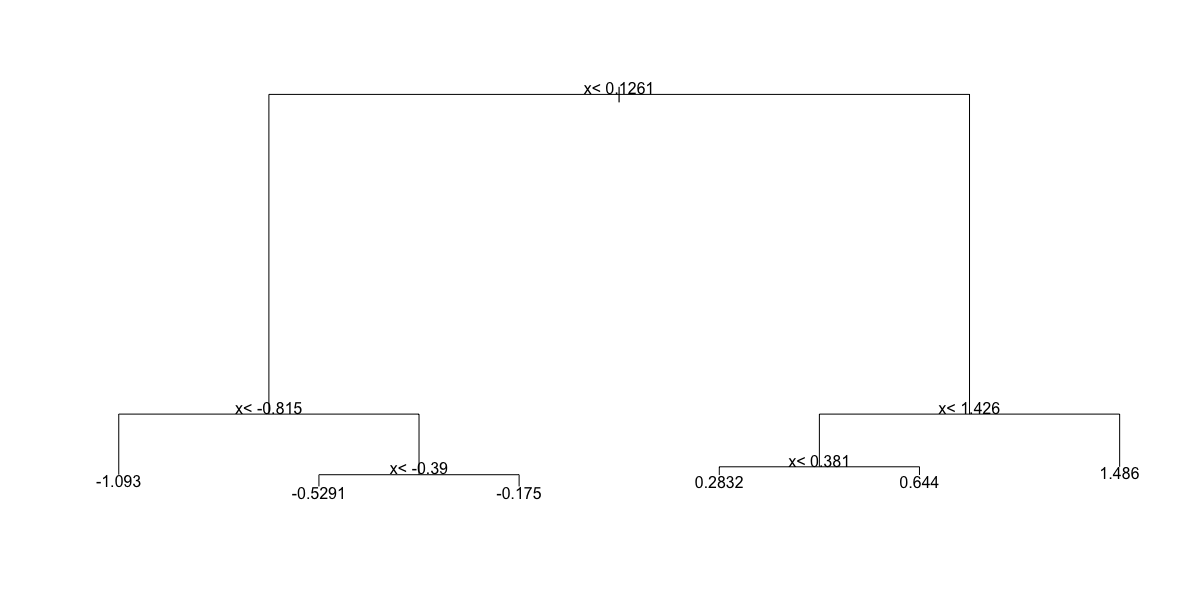
You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

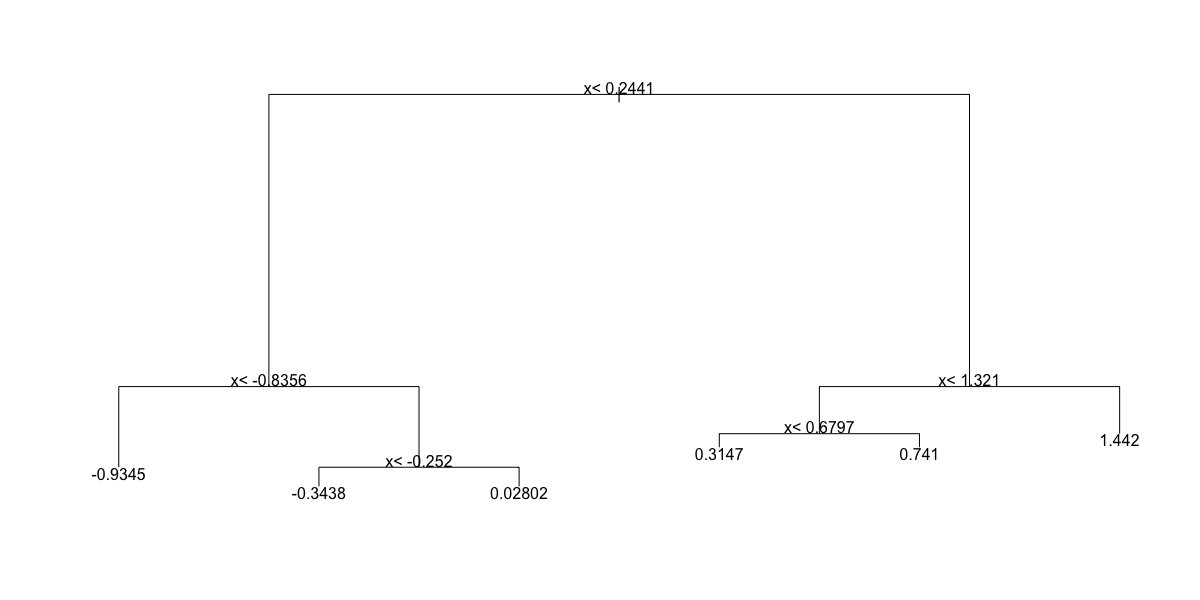
SaveSave Your Answer Show Answer

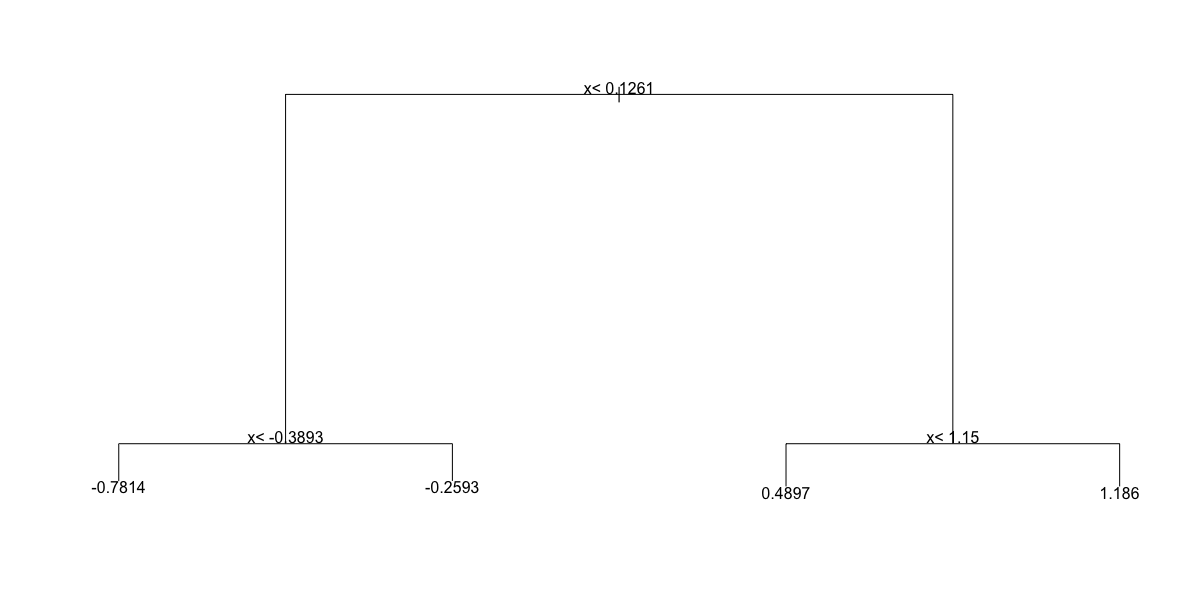
### **Q2**

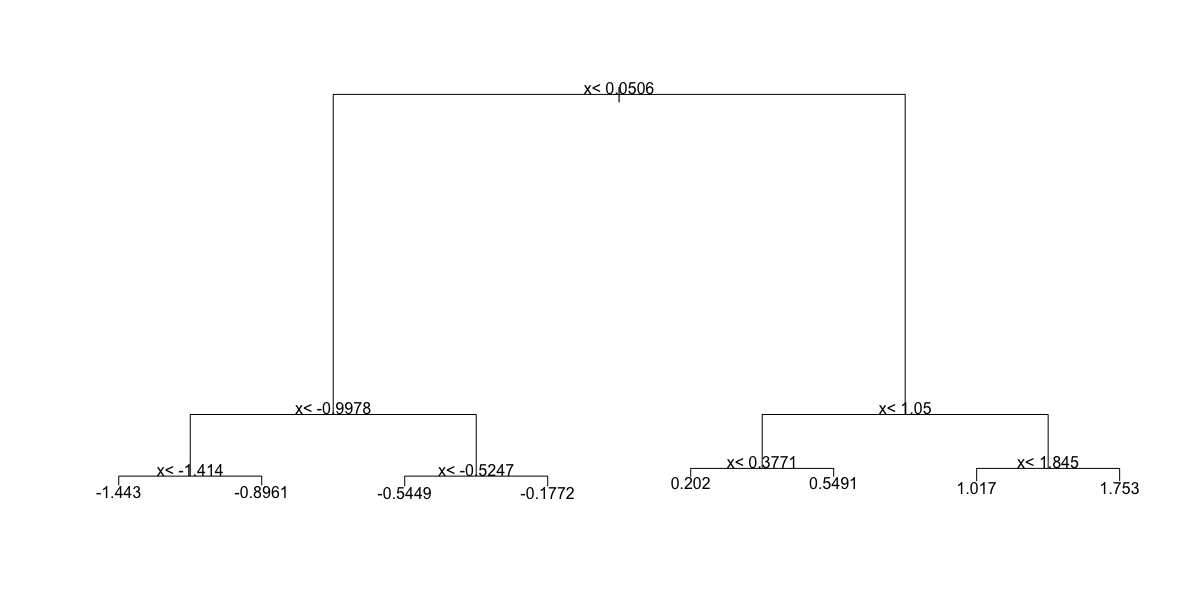
1/1 point (graded)

Which of the following plots has the same tree **shape** obtained in Q1?









correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q3**

1/1 point (graded)

Below is most of the code to make a scatter plot of y versus x along with the predicted values based on the fit.

dat %>%

mutate(y\_hat = predict(fit)) %>%

ggplot() +

geom\_point(aes(x, y)) +

#BLANK

Which line of code should be used to replace #BLANK in the code above?



geom\_step(aes(x, y\_hat), col=2)



geom\_smooth(aes(y\_hat, x), col=2)



geom\_quantile(aes(x, y\_hat), col=2)



geom\_step(aes(y\_hat, x), col=2)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q4**

1/1 point (graded)

Now run Random Forests instead of a regression tree using randomForest() from the **randomForest** package, and remake the scatterplot with the prediction line. Part of the code is provided for you below.

library(randomForest)

fit <- #BLANK

dat %>%

mutate(y\_hat = predict(fit)) %>%

ggplot() +

geom\_point(aes(x, y)) +

geom\_step(aes(x, y\_hat), col = "red")

What code should replace #BLANK in the provided code?



randomForest(y ~ x, data = dat)



randomForest(x ~ y, data = dat)



randomForest(y ~ x, data = data)



randomForest(x ~ y)

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

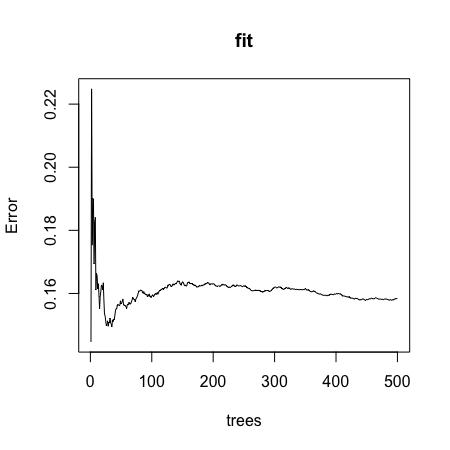
Show Answer

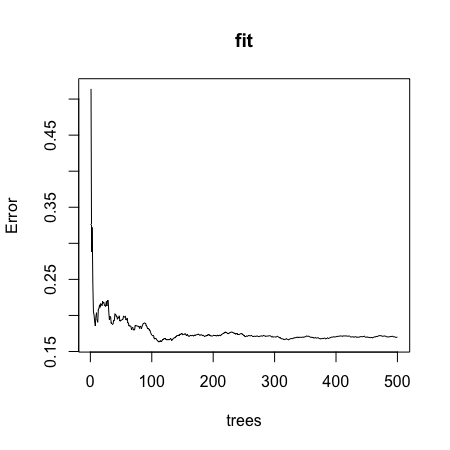
### **Q5**

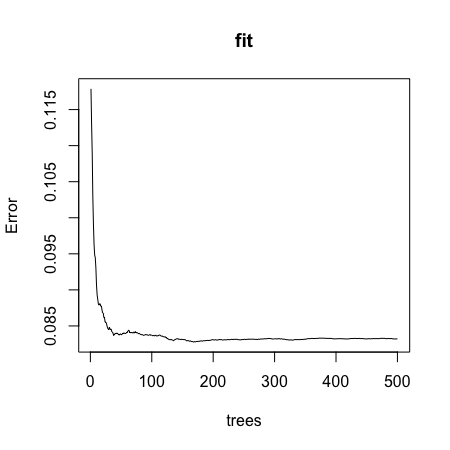
1/1 point (graded)

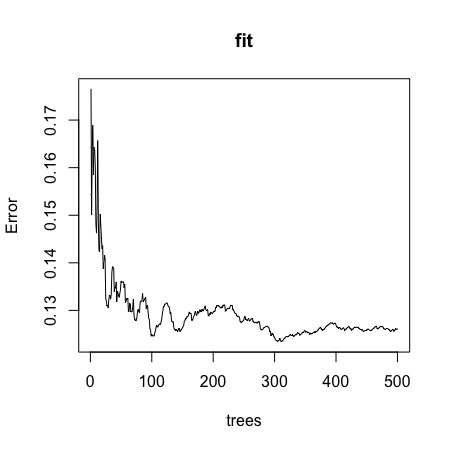
Use the plot() function to see if the Random Forest from Q4 has converged or if we need more trees.

Which of these graphs is produced by plotting the random forest?









correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q6**

1/1 point (graded)

It seems that the default values for the Random Forest result in an estimate that is too flexible (unsmooth). Re-run the Random Forest but this time with a node size of 50 and a maximum of 25 nodes. Remake the plot.

Part of the code is provided for you below.

library(randomForest)

fit <- #BLANK

dat %>%

mutate(y\_hat = predict(fit)) %>%

ggplot() +

geom\_point(aes(x, y)) +

geom\_step(aes(x, y\_hat), col = "red")

What code should replace #BLANK in the provided code?



randomForest(y ~ x, data = dat, nodesize = 25, maxnodes = 25)



randomForest(y ~ x, data = dat, nodes = 50, max = 25)



randomForest(x ~ y, data = dat, nodes = 50, max = 25)



randomForest(y ~ x, data = dat, nodesize = 50, maxnodes = 25)



randomForest(x ~ y, data = dat, nodesize = 50, maxnodes = 25)

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 5.2 Caret Package

## 

## Comprehension Check: Caret Package

 Bookmark this page

These exercises take you through an analysis using the tissue\_gene\_expression dataset.

### **Q1**

1/1 point (graded)

Load the **rpart** package and then use the caret::train() function with method = "rpart" to fit a classification tree to the tissue\_gene\_expression dataset. Try out cp values of seq(0, 0.1, 0.01). Plot the accuracies to report the results of the best model. Set the seed to 1991.

Which value of cp gives the highest accuracy?  correct

0

0.0

**Explanation**

The following code can be used to do generate the plot and get the value of cp:

library(caret)

library(rpart)

library(dslabs)

set.seed(1991)

data("tissue\_gene\_expression")

fit <- with(tissue\_gene\_expression,

train(x, y, method = "rpart",

tuneGrid = data.frame(cp = seq(0, 0.1, 0.01))))

ggplot(fit)

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q2**

1/1 point (graded)

Note that there are only 6 placentas in the dataset. By default, rpart() requires 20 observations before splitting a node. That means that it is difficult to have a node in which placentas are the majority. Rerun the analysis you did in the exercise in Q1, but this time, allow rpart() to split any node by using the argument control = rpart.control(minsplit = 0). Look at the confusion matrix again to determine whether the accuracy increases. Again, set the seed to 1991.

What is the accuracy now?  correct

0.9141

0.9147869

**Explanation**

The following code can be used to re-run the analysis and view the confusion matrix:

library(rpart)

set.seed(1991)

data("tissue\_gene\_expression")

fit\_rpart <- with(tissue\_gene\_expression,

train(x, y, method = "rpart",

tuneGrid = data.frame(cp = seq(0, 0.10, 0.01)),

control = rpart.control(minsplit = 0)))

ggplot(fit\_rpart)

confusionMatrix(fit\_rpart)

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

Plot the tree from the best fitting model of the analysis you ran in Q2.

Which gene is at the first split?

B3GNT4

CAPN3

CES2

CFHR4

CLIP3

GPA33

HRH1

correct

**Explanation**

The first split is at GPA33 >= 8.794. The following code will give the tree:

plot(fit\_rpart$finalModel)

text(fit\_rpart$finalModel)

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q4**

1/1 point (graded)

We can see that with just seven genes, we are able to predict the tissue type. Now let's see if we can predict the tissue type with even fewer genes using a Random Forest. Use the train() function and the rf method to train a Random Forest model and save it to an object called fit. Try out values of mtry ranging from seq(50, 200, 25) (you can also explore other values on your own). What mtry value maximizes accuracy? To permit small nodesize to grow as we did with the classification trees, use the following argument: nodesize = 1.

Note: This exercise will take some time to run. If you want to test out your code first, try using smaller values with ntree. Set the seed to 1991 again.

What value of mtry maximizes accuracy?  correct

100

100

**Explanation**

The following code can be used to do the analysis:

set.seed(1991)

library(randomForest)

fit <- with(tissue\_gene\_expression,

train(x, y, method = "rf",

nodesize = 1,

tuneGrid = data.frame(mtry = seq(50, 200, 25))))

ggplot(fit)

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q5**

1/1 point (graded)

Use the function varImp() on the output of train() and save it to an object called imp:

imp <- #BLANK

imp

What should replace #BLANK in the code above?

Do not include spaces in your answer.

  correct

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q6**

1/1 point (graded)

The rpart() model we ran above produced a tree that used just seven predictors. Extracting the predictor names is not straightforward, but can be done. If the output of the call to train was fit\_rpart, we can extract the names like this:

tree\_terms <- as.character(unique(fit\_rpart$finalModel$frame$var[!(fit\_rpart$finalModel$frame$var == "<leaf>")]))

tree\_terms

Calculate the variable importance in the Random Forest call from Q4 for these seven predictors and examine where they rank.

What is the importance of the CFHR4 gene in the Random Forest call?

Enter a number.

  correct

35.0

35

What is the rank of the CFHR4 gene in the Random Forest call?

Enter a number.

  correct

7

7

**Explanation**

The following code can be used to calculate the rank and importance in the Random Forest call for the predictors from the rpart() model:

data\_frame(term = rownames(imp$importance),

importance = imp$importance$Overall) %>%

mutate(rank = rank(-importance)) %>% arrange(desc(importance)) %>%

filter(term %in% tree\_terms)

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 5.3 Titanic Exercise

## Titanic Exercises Part 1

 Bookmark this page

### Titanic Exercises

These exercises cover everything you have learned in this course so far. You will use the background information to provided to train a number of different types of models on this dataset.

### Background

The Titanic was a British ocean liner that struck an iceberg and sunk on its maiden voyage in 1912 from the United Kingdom to New York. More than 1,500 of the estimated 2,224 passengers and crew died in the accident, making this one of the largest maritime disasters ever outside of war. The ship carried a wide range of passengers of all ages and both genders, from luxury travelers in first-class to immigrants in the lower classes. However, not all passengers were equally likely to survive the accident. You will use real data about a selection of 891 passengers to predict which passengers survived.

### Libraries and data

Use the titanic\_train data frame from the **titanic** library as the starting point for this project.

library(titanic) # loads titanic\_train data frame

library(caret)

library(tidyverse)

library(rpart)

# 3 significant digits

options(digits = 3)

# clean the data - `titanic\_train` is loaded with the titanic package

titanic\_clean <- titanic\_train %>%

mutate(Survived = factor(Survived),

Embarked = factor(Embarked),

Age = ifelse(is.na(Age), median(Age, na.rm = TRUE), Age), # NA age to median age

FamilySize = SibSp + Parch + 1) %>% # count family members

select(Survived, Sex, Pclass, Age, Fare, SibSp, Parch, FamilySize, Embarked)

### **Question 1: Training and test sets**

3/3 points (graded)

Split titanic\_clean into test and training sets - after running the setup code, it should have 891 rows and 9 variables.

Set the seed to 42, then use the **caret** package to create a 20% data partition based on the Survived column. Assign the 20% partition to test\_set and the remaining 80% partition to train\_set.

How many observations are in the training set?  correct

712 Loading

How many observations are in the test set?  correct

179 Loading

What proportion of individuals in the training set survived?  correct

0.383 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2: Baseline prediction by guessing the outcome**

1.0/1.0 point (graded)

The simplest prediction method is randomly guessing the outcome without using additional predictors. These methods will help us determine whether our machine learning algorithm performs better than chance. How accurate are two methods of guessing Titanic passenger survival?

Set the seed to 3. For each individual in the test set, randomly guess whether that person survived or not by sampling from the vector c(0,1) (Note: use the default argument setting of prob from the sample function). Assume that each person has an equal chance of surviving or not surviving.

What is the accuracy of this guessing method?  correct

0.475 Loading

Submit

You have used 6 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3a: Predicting survival by sex**

2.0/2.0 points (graded)

Use the training set to determine whether members of a given sex were more likely to survive or die. Apply this insight to generate survival predictions on the test set.

What proportion of training set females survived?  correct

0.731 Loading

What proportion of training set males survived?  correct

0.197 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3b: Predicting survival by sex**

1.0/1.0 point (graded)

Predict survival using sex on the test set: if the survival rate for a sex is over 0.5, predict survival for all individuals of that sex, and predict death if the survival rate for a sex is under 0.5.

What is the accuracy of this sex-based prediction method on the test set?  correct

0.821 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4a: Predicting survival by passenger class**

1.0/1.0 point (graded)

In the training set, which class(es) (Pclass) of passengers were more likely to survive than die?

Select ALL that apply.

1

2

3

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 4b: Predicting survival by passenger class**

1.0/1.0 point (graded)

Predict survival using passenger class on the test set: predict survival if the survival rate for a class is over 0.5, otherwise predict death.

What is the accuracy of this class-based prediction method on the test set?  correct

0.704 Loading

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4c: Predicting survival by passenger class**

1.0/1.0 point (graded)

Use the training set to group passengers by both sex and passenger class.

Which sex and class combinations were more likely to survive than die?

Select ALL that apply.

female 1st class

female 2nd class

female 3rd class

male 1st class

male 2nd class

male 3rd class

correct

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4d: Predicting survival by passenger class**

1.0/1.0 point (graded)

Predict survival using both sex and passenger class on the test set. Predict survival if the survival rate for a sex/class combination is over 0.5, otherwise predict death.

What is the accuracy of this sex- and class-based prediction method on the test set?  correct

0.821 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 5a: Confusion matrix**

2/2 points (graded)

Use the confusionMatrix() function to create confusion matrices for the sex model, class model, and combined sex and class model. You will need to convert predictions and survival status to factors to use this function.

What is the "positive" class used to calculate confusion matrix metrics?

0

1

correct

Which model has the highest sensitivity?

sex only

class only

sex and class combined

correct

Which model has the highest specificity?

sex only

class only

sex and class combined

correct

Which model has the highest balanced accuracy?

sex only

class only

sex and class combined

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Question 5b: Confusion matrix**

1/1 point (graded)

What is the maximum value of balanced accuracy?  correct

0.806 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6: F1 scores**

2/2 points (graded)

Use the F\_meas() function to calculate F1 scores for the sex model, class model, and combined sex and class model. You will need to convert predictions to factors to use this function.

Which model has the highest F1 score?

sex only

class only

sex and class combined

correct

What is the maximum value of the F1 score?  correct

0.872 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

## Titanic Exercises, part 2

 Bookmark this page

### **Question 7: Survival by fare - LDA and QDA**

2/2 points (graded)

Set the seed to 1. Train a model using linear discriminant analysis (LDA) with the **caret** lda method using fare as the only predictor.

What is the accuracy on the test set for the LDA model?  correct

0.659 **or** 0.693

0.693 Loading

**Explanation**

The accuracy can be determined using the following code:

#set.seed(1) # R 3.5

set.seed(1, sample.kind = "Rounding") # R 3.6

train\_lda <- train(Survived ~ Fare, method = "lda", data = train\_set)

lda\_preds <- predict(train\_lda, test\_set)

mean(lda\_preds == test\_set$Survived)

Set the seed to 1. Train a model using quadratic discriminant analysis (QDA) with the **caret** qda method using fare as the only predictor.

What is the accuracy on the test set for the QDA model?  correct

0.659 **or** 0.693

0.693 Loading

**Explanation**

The accuracy can be determined using the following code:

#set.seed(1) # R 3.5

set.seed(1, sample.kind = "Rounding") # R 3.6

train\_qda <- train(Survived ~ Fare, method = "qda", data = train\_set)

qda\_preds <- predict(train\_qda, test\_set)

mean(qda\_preds == test\_set$Survived)

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 8: Logistic regression models**

3/3 points (graded)

Set the seed to 1. Train a logistic regression model using **caret** train() with the glm method using age as the only predictor.

What is the accuracy on the test set using age as the only predictor?  correct

0.615

0.615 Loading

**Explanation**

The accuracy can be determined using the following code:

train\_glm\_age <- train(Survived ~ Age, method = "glm", data = train\_set)

glm\_preds\_age <- predict(train\_glm\_age, test\_set)

mean(glm\_preds\_age == test\_set$Survived)

Set the seed to 1. Train a logistic regression model using **caret** train() with the glm method using four predictors: sex, class, fare, and age.

What is the accuracy on the test set using these four predictors?  correct

0.821 **or** 0.849

0.849 Loading

**Explanation**

The accuracy can be determined using the following code:

train\_glm <- train(Survived ~ Sex + Pclass + Fare + Age, method = "glm", data = train\_set)

glm\_preds <- predict(train\_glm, test\_set)

mean(glm\_preds == test\_set$Survived)

Set the seed to 1. Train a logistic regression model using **caret** train() with the glm method using all predictors. Ignore warnings about rank-deficient fit.

What is the accuracy on the test set using all predictors?  correct

0.827 **or** 0.849

0.849 Loading

**Explanation**

The accuracy can be determined using the following code:

train\_glm\_all <- train(Survived ~ ., method = "glm", data = train\_set)

glm\_all\_preds <- predict(train\_glm\_all, test\_set)

mean(glm\_all\_preds == test\_set$Survived)

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 9a: kNN model**

1/1 point (graded)

Set the seed to 6. Train a kNN model on the training set using the **caret** train function. Try tuning with k = seq(3, 51, 2).

What is the optimal value of the number of neighbors k?  correct

15 **or** 11

11 Loading

**Explanation**

The optimal value can be calculated using the following code:

#set.seed(6)

set.seed(6, sample.kind = "Rounding") # simulate R 3.5

train\_knn <- train(Survived ~ .,

method = "knn",

data = train\_set,

tuneGrid = data.frame(k = seq(3, 51, 2)))

train\_knn$bestTune

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Hint SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 9b: kNN model**

1/1 point (graded)

Plot the kNN model to investigate the relationship between the number of neighbors and accuracy on the training set.

Of these values of k, which yields the highest accuracy?

7

11

17

21

correct

**Explanation**

The plot can be generated using the following code:

ggplot(train\_knn)

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 9c: kNN model**

1/1 point (graded)

What is the accuracy of the kNN model on the test set?  correct

0.709 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 10: Cross-validation**

2/2 points (graded)

Set the seed to 8 and train a new kNN model. Instead of the default training control, use 10-fold cross-validation where each partition consists of 10% of the total.

Try tuning with k = seq(3, 51, 2). What is the optimal value of k using cross-validation?  correct

23 **or** 5

5 Loading

**Explanation**

The optimal value of k can be found using the following code:

#set.seed(8)

set.seed(8, sample.kind = "Rounding") # simulate R 3.5

train\_knn\_cv <- train(Survived ~ .,

method = "knn",

data = train\_set,

tuneGrid = data.frame(k = seq(3, 51, 2)),

trControl = trainControl(method = "cv", number = 10, p = 0.9))

train\_knn\_cv$bestTune

What is the accuracy on the test set using the cross-validated kNN model?  correct

0.737 **or** 0.648

0.648 Loading

**Explanation**

The accuracy can be calculated using the following code:

knn\_cv\_preds <- predict(train\_knn\_cv, test\_set)

mean(knn\_cv\_preds == test\_set$Survived)

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Question 11a: Classification tree model**

2/2 points (graded)

Set the seed to 10. Use **caret** to train a decision tree with the rpart method. Tune the complexity parameter with cp = seq(0, 0.05, 0.002).

What is the optimal value of the complexity parameter (cp)?  correct

0.016 Loading

What is the accuracy of the decision tree model on the test set?  correct

0.838 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 11b: Classification tree model**

1/1 point (graded)

Inspect the final model and plot the decision tree.

Which variables are used in the decision tree?

Select ALL that apply.

Survived

Sex

Pclass

Age

Fare

Parch

Embarked

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 11c: Classification tree model**

1.7142857142857142/2 points (graded)

Using the decision rules generated by the final model, predict whether the following individuals would survive.

**A 28-year-old male**                           

correct

**A female in the second passenger class**                           

correct

**A third-class female who paid a fare of $8**                           

incorrect

**A 5-year-old male with 4 siblings**                           

correct

**A third-class female who paid a fare of $25**                           

correct

**A first-class 17-year-old female with 2 siblings**                           

correct

**A first-class 17-year-old male with 2 siblings**                           

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Partially correct (1.71/2 points)

Review

### **Question 12: Random forest model**

3/3 points (graded)

Set the seed to 14. Use the **caret** train() function with the rf method to train a random forest. Test values of mtry ranging from 1 to 7. Set ntree to 100.

What mtry value maximizes accuracy?  correct

2 Loading

What is the accuracy of the random forest model on the test set?  correct

0.844 Loading

Use varImp() on the random forest model object to determine the importance of various predictors to the random forest model.

What is the most important variable?

Be sure to report the variable name exactly as it appears in the code.

  correct

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

## Section 6: Model Fitting and Recommendation System

## 6.1 Case Study: MNIST

## Comprehension Check: Dimension Reduction

 Bookmark this page

The [dimension reduction section of the textbook](https://rafalab.github.io/dsbook/large-datasets.html#dimension-reduction) may help with some of these exercises.

### **Q1**

1/1 point (graded)

We want to explore the tissue\_gene\_expression predictors by plotting them.

library(dslabs)

data("tissue\_gene\_expression")

dim(tissue\_gene\_expression$x)

We want to get an idea of which observations are close to each other, but, as you can see from the dimensions, the predictors are 500-dimensional, making plotting difficult. Plot the first two principal components with color representing tissue type.

Which tissue is in a cluster by itself?

cerebellum

colon

endometrium

hippocampus

kidney

liver

placenta

correct

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q2**

1/1 point (graded)

The predictors for each observation are measured using the same device and experimental procedure. This introduces biases that can affect all the predictors from one observation. For each observation, compute the average across all predictors, and then plot this against the first PC with color representing tissue. Report the correlation.

What is the correlation?  correct

0.596 Loading

Submit

You have used 5 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q3**

1/1 point (graded)

We see an association with the first PC and the observation averages. Redo the PCA but only after removing the center. Part of the code is provided for you.

#BLANK

pc <- prcomp(x)

data.frame(pc\_1 = pc$x[,1], pc\_2 = pc$x[,2],

tissue = tissue\_gene\_expression$y) %>%

ggplot(aes(pc\_1, pc\_2, color = tissue)) +

geom\_point()

Which line of code should be used to replace #BLANK in the code block above?



x <- with(tissue\_gene\_expression, sweep(x, 1, mean(x)))



x <- sweep(x, 1, rowMeans(tissue\_gene\_expression$x))



x <- tissue\_gene\_expression$x - mean(tissue\_gene\_expression$x)



x <- with(tissue\_gene\_expression, sweep(x, 1, rowMeans(x)))

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q4**

1/1 point (graded)

For the first 10 PCs, make a boxplot showing the values for each tissue.

For the 7th PC, which two tissues have the greatest median difference?

Select the TWO tissues that have the greatest median difference.

cerebellum

colon

endometrium

hippocampus

kidney

liver

placenta

correct

Submit

You have used 2 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q5**

1/1 point (graded)

Plot the percent variance explained by PC number. Hint: use the summary function.

How many PCs are required to reach a cumulative percent variance explained greater than 50%?  correct

3 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## 6.2 Recommendation Systems

## Comprehension Check: Recommendation Systems

 Bookmark this page

The following exercises all work with the movielens data, which can be loaded using the following code:

library(tidyverse)

library(lubridate)

library(dslabs)

data("movielens")

### **Q1**

1/1 point (graded)

Compute the number of ratings for each movie and then plot it against the year the movie came out. Use the square root transformation on the y-axis when plotting.

What year has the highest median number of ratings?  correct

1995 Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q2**

1/1 point (graded)

We see that, on average, movies that came out after 1993 get more ratings. We also see that with newer movies, starting in 1993, the number of ratings decreases with year: the more recent a movie is, the less time users have had to rate it.

Among movies that came out in 1993 or later, select the top 25 movies with the highest average number of ratings per year (n/year), and caculate the average rating of each of them. To calculate number of ratings per year, use 2018 as the end year.

What is the average rating for the movie The Shawshank Redemption ("Shawshank Redemption, The")?  correct

4.49

4.49 Loading

What is the average number of ratings per year for the movie Forrest Gump?  correct

14.2

14.2 Loading

**Explanation**

The top 25 movies with the most ratings per year, along with their average ratings, can be found using the following code:

movielens %>%

filter(year >= 1993) %>%

group\_by(movieId) %>%

summarize(n = n(), years = 2018 - first(year),

title = title[1],

rating = mean(rating)) %>%

mutate(rate = n/years) %>%

top\_n(25, rate) %>%

arrange(desc(rate))

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q3**

1/1 point (graded)

From the table constructed in Q2, we can see that the most frequently rated movies tend to have above average ratings. This is not surprising: more people watch popular movies. To confirm this, stratify the post-1993 movies by ratings per year and compute their average ratings. To calculate number of ratings per year, use 2018 as the end year. Make a plot of average rating versus ratings per year and show an estimate of the trend.

What type of trend do you observe?

There is no relationship between how often a movie is rated and its average rating.

Movies with very few and very many ratings have the highest average ratings.

The more often a movie is rated, the higher its average rating.

The more often a movie is rated, the lower its average rating.

correct

**Explanation**

The plot can be generated using the following code:

movielens %>%

filter(year >= 1993) %>%

group\_by(movieId) %>%

summarize(n = n(), years = 2018 - first(year),

title = title[1],

rating = mean(rating)) %>%

mutate(rate = n/years) %>%

ggplot(aes(rate, rating)) +

geom\_point() +

geom\_smooth()

We see that the trend is that the more often a movie is rated, the higher its average rating.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Answers are displayed within the problem

Review

### **Q4**

1/1 point (graded)

Suppose you are doing a predictive analysis in which you need to fill in the missing ratings with some value.

Given your observations in the exercise in Q3, which of the following strategies would be most appropriate?

Fill in the missing values with the average rating across all movies.

Fill in the missing values with 0.

Fill in the missing values with a lower value than the average rating across all movies.

Fill in the value with a higher value than the average rating across all movies.

None of the above.

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q5**

1/1 point (graded)

The movielens dataset also includes a time stamp. This variable represents the time and data in which the rating was provided. The units are seconds since January 1, 1970. Create a new column date with the date.

Which code correctly creates this new column?



movielens <- mutate(movielens, date = as.date(timestamp))



movielens <- mutate(movielens, date = as\_datetime(timestamp))



movielens <- mutate(movielens, date = as.data(timestamp))



movielens <- mutate(movielens, date = timestamp)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q6**

1/1 point (graded)

Compute the average rating for each week and plot this average against date. Hint: use the round\_date() function before you group\_by().

What type of trend do you observe?

There is strong evidence of a time effect on average rating.

There is some evidence of a time effect on average rating.

There is no evidence of a time effect on average rating.

correct

Submit

You have used 1 of 1 attemptSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Q7**

1/1 point (graded)

Consider again the plot you generated in Q6.

If we define du,i as the day for user's u rating of movie i, which of the following models is most appropriate?

Yu,i=μ+bi+bu+du,i+εu,i

Yu,i=μ+bi+bu+du,iβ+εu,i

Yu,i=μ+bi+bu+du,iβi+εu,i

Yu,i=μ+bi+bu+f(du,i)+εu,i, with f a smooth function of du,i

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q8**

1/1 point (graded)

The movielens data also has a genres column. This column includes every genre that applies to the movie. Some movies fall under several genres. Define a category as whatever combination appears in this column. Keep only categories with more than 1,000 ratings. Then compute the average and standard error for each category. Plot these as error bar plots.

Which genre has the lowest average rating?

Enter the name of the genre exactly as reported in the plot, including capitalization and punctuation.

  correct

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q9**

1/1 point (graded)

The plot you generated in Q8 shows strong evidence of a genre effect. Consider this plot as you answer the following question.

If we define gu,i as the genre for user u's rating of movie i, which of the following models is most appropriate?

Yu,i=μ+bi+bu+gu,i+εu,i

Yu,i=μ+bi+bu+gu,iβ+εu,i

Yu,i=μ+bi+bu+∑Kk=1xku,iβk+εu,i, with xku,i=1 if gu,i is genre k

Yu,i=μ+bi+bu+f(gu,i)+εu,i, with f a smooth function of gu,i

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## 6.3 Regularization

## Comprehension Check: Regularization

 Bookmark this page

The exercises in Q1-Q8 work with a simulated dataset for 1000 schools. This pre-exercise setup walks you through the code needed to simulate the dataset.

If you have not done so already since the Titanic Exercises, please restart R or reset the number of digits that are printed with options(digits=7).

An education expert is advocating for smaller schools. The expert bases this recommendation on the fact that among the best performing schools, many are small schools. Let's simulate a dataset for 1000 schools. First, let's simulate the number of students in each school, using the following code:

set.seed(1986) #for R 3.5 or earlier

#if using R 3.6 or later, use `set.seed(1986, sample.kind="Rounding")` instead

n <- round(2^rnorm(1000, 8, 1))

Now let's assign a **true** quality for each school that is completely independent from size. This is the parameter we want to estimate in our analysis. The true quality can be assigned using the following code:

set.seed(1) #for R 3.5 or earlier

#if using R 3.6 or later, use `set.seed(1, sample.kind="Rounding")` instead

mu <- round(80 + 2\*rt(1000, 5))

range(mu)

schools <- data.frame(id = paste("PS",1:1000),

size = n,

quality = mu,

rank = rank(-mu))

We can see the top 10 schools using this code:

schools %>% top\_n(10, quality) %>% arrange(desc(quality))

Now let's have the students in the school take a test. There is random variability in test taking, so we will simulate the test scores as normally distributed with the average determined by the school quality with a standard deviation of 30 percentage points. This code will simulate the test scores:

set.seed(1) #for R 3.5 or earlier

#if using R 3.6 or later, use `set.seed(1, sample.kind="Rounding")` instead

mu <- round(80 + 2\*rt(1000, 5))

scores <- sapply(1:nrow(schools), function(i){

scores <- rnorm(schools$size[i], schools$quality[i], 30)

scores

})

schools <- schools %>% mutate(score = sapply(scores, mean))

### **Q1**

1/1 point (graded)

What are the top schools based on the average score? Show just the ID, size, and the average score.

Report the ID of the top school and average score of the 10th school.

What is the ID of the top school?

Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

  correct

567 Loading

What is the average score of the 10th school?  correct

87.95 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q2**

1/1 point (graded)

Compare the median school size to the median school size of the top 10 schools based on the score.

What is the median school size overall?  correct

261 Loading

What is the median school size of the of the top 10 schools based on the score?  correct

185.5 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q3**

1/1 point (graded)

According to this analysis, it appears that small schools produce better test scores than large schools. Four out of the top 10 schools have 100 or fewer students. But how can this be? We constructed the simulation so that quality and size were independent. Repeat the exercise for the worst 10 schools.

What is the median school size of the bottom 10 schools based on the score?  correct

219 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q4**

1/1 point (graded)

From this analysis, we see that the worst schools are also small. Plot the average score versus school size to see what's going on. Highlight the top 10 schools based on the **true** quality.

What do you observe?

There is no difference in the standard error of the score based on school size; there must be an error in how we generated our data.

The standard error of the score has larger variability when the school is smaller, which is why both the best and the worst schools are more likely to be small.

The standard error of the score has smaller variability when the school is smaller, which is why both the best and the worst schools are more likely to be small.

The standard error of the score has larger variability when the school is very small or very large, which is why both the best and the worst schools are more likely to be small.

The standard error of the score has smaller variability when the school is very small or very large, which is why both the best and the worst schools are more likely to be small.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Q5**

1/1 point (graded)

Let's use regularization to pick the best schools. Remember regularization **shrinks** deviations from the average towards 0. To apply regularization here, we first need to define the overall average for all schools, using the following code:

overall <- mean(sapply(scores, mean))

Then, we need to define, for each school, how it deviates from that average.

Write code that estimates the score above the average for each school but dividing by n+α instead of n, with n the school size and α a regularization parameter. Try α=25.

What is the ID of the top school with regularization?

Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

  correct

191

191 Loading

What is the regularized score of the 10th school?  correct

87.15

87.2 Loading

**Explanation**

The regularization and reporting of scores can be done using the following code:

alpha <- 25

score\_reg <- sapply(scores, function(x) overall + sum(x-overall)/(length(x)+alpha))

schools %>% mutate(score\_reg = score\_reg) %>%

top\_n(10, score\_reg) %>% arrange(desc(score\_reg))

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q6**

1/1 point (graded)

Notice that this improves things a bit. The number of small schools that are not highly ranked is now lower. Is there a better α? Using values of α from 10 to 250, find the α that minimizes the RMSE.

RMSE=11000∑i=11000(quality−estimate)2−−−−−−−−−−−−−−−−−−−−−−−−⎷

What value of α gives the minimum RMSE?  correct

135

135 Loading

**Explanation**

The value of α that minimizes the MSE can be calculated using the following code:

alphas <- seq(10,250)

rmse <- sapply(alphas, function(alpha){

score\_reg <- sapply(scores, function(x) overall+sum(x-overall)/(length(x)+alpha))

sqrt(mean((score\_reg - schools$quality)^2))

})

plot(alphas, rmse)

alphas[which.min(rmse)]

Submit

You have used 3 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q7**

1/1 point (graded)

Rank the schools based on the average obtained with the best α. Note that no small school is incorrectly included.

What is the ID of the top school now?

Note that the school IDs are given in the form "PS x" - where x is a number. Report the **number** only.

  correct

191

191 Loading

What is the regularized average score of the 10th school now?  correct

85.4

85.5 Loading

**Explanation**

The new ranking can be done using the following code:

alpha <- alphas[which.min(rmse)]

score\_reg <- sapply(scores, function(x)

overall+sum(x-overall)/(length(x)+alpha))

schools %>% mutate(score\_reg = score\_reg) %>%

top\_n(10, score\_reg) %>% arrange(desc(score\_reg))

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

Review

### **Q8**

1/1 point (graded)

A common mistake made when using regularization is shrinking values towards 0 that are not centered around 0. For example, if we don't subtract the overall average before shrinking, we actually obtain a very similar result. Confirm this by re-running the code from the exercise in Q6 but without removing the overall mean.

What value of α gives the minimum RMSE here?  correct

10

10 Loading

**Explanation**

The code here is nearly the same as in Q6, but we don't subtract the overall mean. The value of α that minimizes the RMSE can be calculated using the following code:

alphas <- seq(10,250)

rmse <- sapply(alphas, function(alpha){

score\_reg <- sapply(scores, function(x) sum(x)/(length(x)+alpha))

sqrt(mean((score\_reg - schools$quality)^2))

})

plot(alphas, rmse)

alphas[which.min(rmse)]

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Answers are displayed within the problem

## Comprehension Check: Clustering

 Bookmark this page

These exercises will work with the tissue\_gene\_expression dataset, which is part of the **dslabs** package.

### **Q1**

1/1 point (graded)

Load the tissue\_gene\_expression dataset. Remove the row means and compute the distance between each observation. Store the result in d.

Which of the following lines of code correctly does this computation?



d <- dist(tissue\_gene\_expression$x)



d <- dist(rowMeans(tissue\_gene\_expression$x))



d <- dist(rowMeans(tissue\_gene\_expression$y))



d <- dist(tissue\_gene\_expression$x - rowMeans(tissue\_gene\_expression$x))

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q2**

1/1 point (graded)

Make a hierarchical clustering plot and add the tissue types as labels.

You will observe multiple branches.

Which tissue type is in the branch farthest to the left?

cerebellum

colon

endometrium

hippocampus

kidney

liver

placenta

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Q3**

0/1 point (graded)

Run a k-means clustering on the data with K=7. Make a table comparing the identified clusters to the actual tissue types. Run the algorithm several times to see how the answer changes.

What do you observe for the clustering of the **liver** tissue?

Liver is always classified in a single cluster.

Liver is never classified in a single cluster.

Liver is classified in a single cluster roughly 20% of the time and in more than one cluster roughly 80% of the time.

Liver is classified in a single cluster roughly 80% of the time and in more than one cluster roughly 20% of the time.

incorrect

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Q4**

1/1 point (graded)

Select the 50 most variable genes. Make sure the observations show up in the columns, that the predictor are centered, and add a color bar to show the different tissue types. Hint: use the ColSideColors argument to assign colors. Also, use col = RColorBrewer::brewer.pal(11, "RdBu") for a better use of colors.

Part of the code is provided for you here:

library(RColorBrewer)

sds <- matrixStats::colSds(tissue\_gene\_expression$x)

ind <- order(sds, decreasing = TRUE)[1:50]

colors <- brewer.pal(7, "Dark2")[as.numeric(tissue\_gene\_expression$y)]

#BLANK

Which line of code should replace #BLANK in the code above?



heatmap(t(tissue\_gene\_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = colors)



heatmap(t(tissue\_gene\_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = rev(colors))



heatmap(t(tissue\_gene\_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = sample(colors))



heatmap(t(tissue\_gene\_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = sample(colors))

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.