Comprehension Check: Nearest Neighbors

Q1

2/2 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` function to perform knn with `k` values of `seq(1, 101, 3)` and calculate `F_1` scores.

What is the max value of `F_1`?

0.6019417 ✔ **Answer:** 0.60194

0.6019417

At what value of `k` does the max occur?

46 ✔ **Answer:** 46

46

Explanation

This exercise can be accomplished using the following code:

```
library(dslabs)
library(tidyverse)
library(caret)
data("heights")

set.seed(1)
test_index <- createDataPartition(heights$sex, times = 1, p = 0.5, list = FALSE)
test_set <- heights[test_index, ]
train_set <- heights[-test_index, ]

ks <- seq(1, 101, 3)
F_1 <- sapply(ks, function(k){
        fit <- knn3(sex ~ height, data = train_set, k = k)
        y_hat <- predict(fit, test_set, type = "class") %>%
                factor(levels = levels(train_set$sex))
        F_meas(data = y_hat, reference = test_set$sex)
})
plot(ks, F_1)
max(F_1)
ks[which.max(F_1)]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Q2

2/2 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")
```

Split the data into training and test sets, and report the accuracy you obtain. Try it for `k = 1, 3, 5, 7, 9, 11`. Set the seed to 1 before splitting the data.

k=1

0.9895833 ✔ **Answer:** 0.989

0.9895833

k=3

0.9687500 ✔ **Answer:** 0.968

0.9687500

k=5

0.9479167 ✔ **Answer:** 0.947

0.9479167

k=7

0.9166667 ✔ **Answer:** 0.916

0.9166667

k=9

0.9166667 ✔ **Answer:** 0.916

0.9166667

k=11

0.9062500 ✔ **Answer:** 0.906

0.9062500

Explanation
This exercise can be accomplished using the following code:

```
set.seed(1)
library(caret)
y <- tissue_gene_expression$y
x <- tissue_gene_expression$x
test_index <- createDataPartition(y, list = FALSE)
sapply(seq(1, 11, 2), function(k){
        fit <- knn3(x[-test_index,], y[-test_index], k = k)
        y_hat <- predict(fit, newdata = data.frame(x=x[test_index,]),
                                 type = "class")
mean(y_hat == y[test_index])
})
```

You have used 1 of 10 attempts

Submit

Answers are displayed within the problem

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

Some reminders:

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

Discussion: Nearest Neighbors

Show Discussion

Topic: Section 4: Distance, Knn, Cross-Validation, and Generative Models / 4.1.2: Nearest Neighbors