cross_valid_question

Group 4 - Cross Validation Problem

Our way of thinking: for each fold test all the ks -> plot the average error of all the folds for ks

```
# help https://genomicsclass.github.io/book/pages/crossvalidation.html - seems like for every fold chec
library(class)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
load("data/id100.Rda")

# shuffle dataset
set.seed(423)
shuffled_df <- id[sample(nrow(id)),]

# for every fold check all the ks
ks <- c(1,5,9,13,17,21,25,29)
mse_k_all_folds_avg <- rep(0, length(ks)) # vector for storing error of each k

# do the folds
folds <- createFolds(shuffled_df$X1, k = 10)

fold_no <- 0 # just for printing the progress

#current fold - to ignore in training data but to use as a validation dataset: fold -> test!
for (fold in folds){
  fold_no <- fold_no + 1
  print(paste0("fold_no =", fold_no))

  fold <- sample(fold) # shuffle id numbers of fold

  # use not-fold data as training and fold data as test
  not_f_df <- shuffled_df[ -fold, ]
  f_df <- shuffled_df[ fold, ]

  # get labels
  f_train_labels <- not_f_df[,1]
  f_test_labels <- f_df[,1]

  # predict with knn - test all k's for each fold

  k_no <- 0 # just to keep track
  k_error <- 0
  for (k in ks){
    k_no <- k_no + 1
```

```r
    f_test_pred <- knn(train = not_f_df, test = f_df, cl = f_train_labels, k=k)

    k_error <- mean(f_test_labels != f_test_pred) # error of current k in this fold
    print(paste0("      k =", k, " k_error =", k_error))
    print( mse_k_all_folds_avg) # Error vector that we want to plot

    # we just want to plot the results so we came up with idea of saving the errors in vector
    # for each k in the vector all the folds are summed up together, and then we get the average
    mse_k_all_folds_avg[k_no] <- mse_k_all_folds_avg[k_no] + k_error
  }
}
```

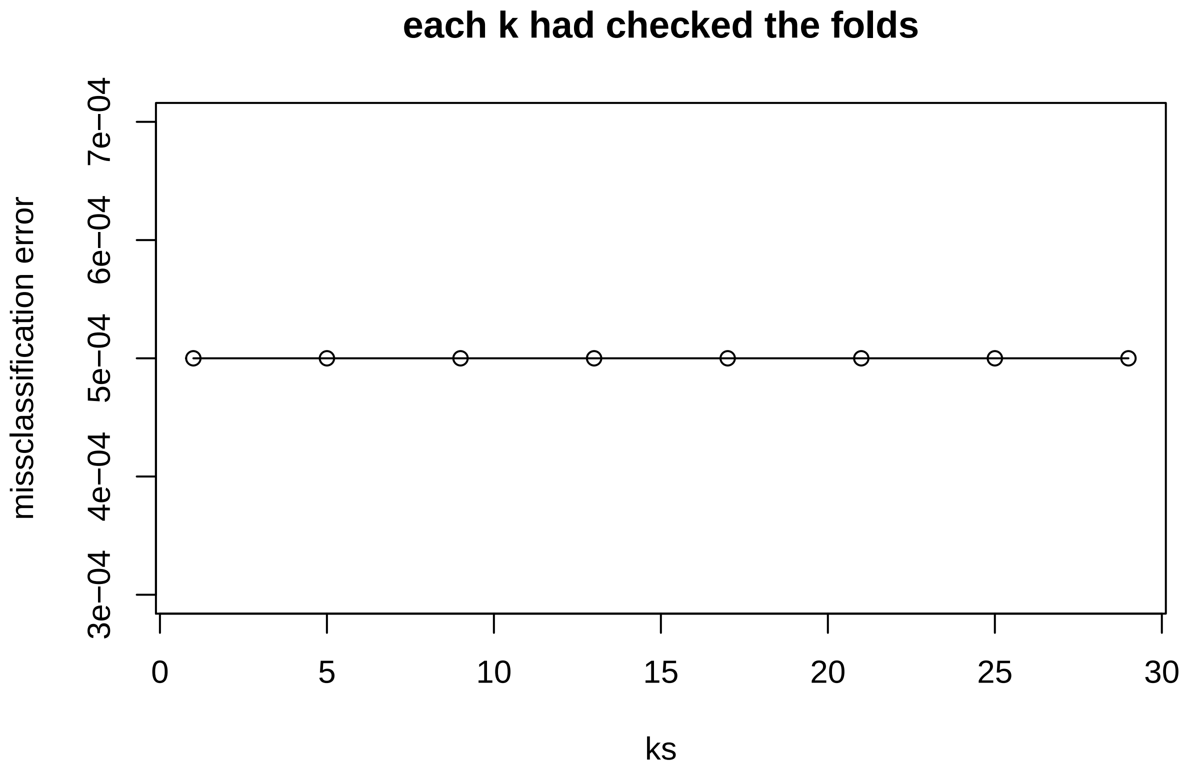
```
## [1] "fold_no =1"
## [1] "      k =1 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =5 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =9 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =13 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =17 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =21 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =25 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =29 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "fold_no =2"
## [1] "      k =1 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =5 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =9 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =13 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =17 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =21 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =25 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =29 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "fold_no =3"
## [1] "      k =1 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =5 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =9 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =13 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =17 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =21 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =25 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =29 k_error =0"
## [1] 0 0 0 0 0 0 0 0
## [1] "fold_no =4"
## [1] "      k =1 k_error =0.0025"
## [1] 0 0 0 0 0 0 0 0
## [1] "      k =5 k_error =0.0025"
## [1] 0.0025 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## [1] "      k =9 k_error =0.0025"
## [1] 0.0025 0.0025 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
## [1] "      k =13 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0000 0.0000 0.0000 0.0000 0.0000
## [1] "      k =17 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0000 0.0000 0.0000 0.0000
## [1] "      k =21 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0000 0.0000 0.0000
## [1] "      k =25 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0000 0.0000
## [1] "      k =29 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0000
## [1] "fold_no =5"
## [1] "      k =1 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =5 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =9 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =13 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =17 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =21 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =25 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =29 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "fold_no =6"
## [1] "      k =1 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =5 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =9 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =13 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =17 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =21 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =25 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =29 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "fold_no =7"
## [1] "      k =1 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =5 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =9 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =13 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =17 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =21 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =25 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =29 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "fold_no =8"
## [1] "      k =1 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =5 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =9 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =13 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =17 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =21 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =25 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =29 k_error =0"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "fold_no =9"
## [1] "      k =1 k_error =0.0025"
## [1] 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =5 k_error =0.0025"
## [1] 0.0050 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =9 k_error =0.0025"
## [1] 0.0050 0.0050 0.0025 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =13 k_error =0.0025"
## [1] 0.0050 0.0050 0.0050 0.0025 0.0025 0.0025 0.0025 0.0025
## [1] "      k =17 k_error =0.0025"
## [1] 0.0050 0.0050 0.0050 0.0050 0.0025 0.0025 0.0025 0.0025
## [1] "      k =21 k_error =0.0025"
## [1] 0.0050 0.0050 0.0050 0.0050 0.0050 0.0025 0.0025 0.0025
## [1] "      k =25 k_error =0.0025"
## [1] 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0025 0.0025
## [1] "      k =29 k_error =0.0025"
## [1] 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0050 0.0025
## [1] "fold_no =10"
## [1] "      k =1 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =5 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =9 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =13 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =17 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =21 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =25 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
## [1] "      k =29 k_error =0"
## [1] 0.005 0.005 0.005 0.005 0.005 0.005 0.005 0.005
```

```r
mse_k_all_folds_avg <- mse_k_all_folds_avg / length(folds)

print(paste0("mse_k_all_folds_avg = ", mse_k_all_folds_avg))
```

```
## [1] "mse_k_all_folds_avg = 5e-04" "mse_k_all_folds_avg = 5e-04"
## [3] "mse_k_all_folds_avg = 5e-04" "mse_k_all_folds_avg = 5e-04"
## [5] "mse_k_all_folds_avg = 5e-04" "mse_k_all_folds_avg = 5e-04"
## [7] "mse_k_all_folds_avg = 5e-04" "mse_k_all_folds_avg = 5e-04"
```

```r
plot(ks, mse_k_all_folds_avg, type="o", ylab="missclassification error", main="each k had checked the fo
```



# Our problem Why we don't get any error in some folds?

Although getting the training and test data for each fold works and the calculation of the error itself also works, probably the cross-validation is not implemented correctly. We are worried that we don't get any error in some folds.

Could you show us where is the mistake? Or maybe you can suggest other way to do cross-validation?