Machine Learning Lab1

Lepeng Zhang, Xuan Wang, Priyarani Patil

2023-11-08

Assignment 1. Handwritten digit recognition with Knearest neighbors.

1.Import the data into R and divide it into training, validation and test sets (50%/25%/25%) by using the partitioning principle specified in the lecture slides.

Answer:

```
##   X0 X1 X6 X15 X12 X1.1 X0.1 X0.2 X0.3 X7 X16 X6.1 X6.2 X10 X0.4 X0.5 X0.6 X8
## 1  0  0 10  16   6    0    0    0    0  7  16    8   16   5    0    0    0 11
## 2  0  0  8  15  16   13    0    0    0  1  11    9   11  16    1    0    0  0
## 3  0  0  0   3  11   16    0    0    0  0   5   16   11  13    7    0    0  3
## 4  0  0  5  14   4    0    0    0    0  0  13    8    0   0    0    0    0  3
## 5  0  0 11  16  10    1    0    0    0  4  16   10   15   8    0    0    0  4
##   X16.1 X2 X0.7 X11 X2.1 X0.8 X0.9 X5 X16.2 X3 X0.10 X5.1 X7.1 X0.11 X0.12 X7.2
## 1    16  0    6  14    3    0    0 12    12  0     0   11   11     0     0   12
## 2     0  0    7  14    0    0    0  0     3  4    14   12    2     0     0    1
## 3    15  8    1  15    6    0    0 11    16 16    16   16   10     0     0    1
## 4    14  4    0   0    0    0    0  6    16 14     9    2    0     0     0    4
## 5    16  3   11  13    0    0    0  1    14  6     9   14    0     0     0    0
##   X13 X3.1 X0.13 X8.1 X7.3 X0.14 X0.15 X4 X12.1 X0.16 X1.2 X13.1 X5.2 X0.17
## 1  12    0     0    8   12     0     0  7    15     1    0    13   11     0
## 2  16   16    16   16   10     0     0  2    12    16   10     0    0     0
## 3   4    4    13   10    2     0     0  0     0     0   15     4    0     0
## 4  16    3     4   11    2     0     0  0    14     3    0     4   11     0
## 5   0    0    12   10    0     0     0  0     0     6   16     6    0     0
##   X0.18 X0.19 X14 X9 X15.1 X9.1 X0.20 X0.21 X0.22 X0.23 X6.3 X14.1 X7.4 X1.3
## 1     0     0  16  8    10   15     3     0     0     0   10    16   15    3
## 2     0     0   2 16     4    0     0     0     0     0    9    14    0    0
## 3     0     0   0  3    16    0     0     0     0     0    0     1   15    2
## 4     0     0  10  8     4   11    12     0     0     0    4    12   14    7
## 5     0     0   5 15    15    8     8     3     0     0   10    16   16   16
##   X0.24 X0.25 X0.26
## 1     0     0     0
## 2     0     0     7
## 3     0     0     4
## 4     0     0     6
## 5    16     6     2
```

2.Use training data to fit 30-nearest neighbor classifier with function kknn() and kernel="rectangular" from package kknn and estimate • Confusion matrices for the training and test data (use table()) • Misclassification errors for the training and test data Comment on the quality of predictions for different digits and on the overall prediction quality.

Answer:

3.Find any 2 cases of digit "8" in the training data which were easiest to classify and 3 cases that were hardest to classify (i.e. having highest and lowest probabilities of the correct class). Reshape features for each of these cases as matrix 8x8 and visualize the corresponding digits (by using e.g. heatmap() function with parameters Colv=NA and Rowv=NA) and comment on whether these cases seem to be hard or easy to recognize visually.

4.Fit a K-nearest neighbor classifiers to the training data for different values of = 1,2, … , 30 and plot the dependence of the training and validation misclassification errors on the value of K (in the same plot). How does the model complexity change when K increases and how does it affect the training and validation errors? Report the optimal according to this plot. Finally, estimate the test error for the model having the optimal K, compare it with the training and validation errors and make necessary conclusions about the model quality.

5.Fit K-nearest neighbor classifiers to the training data for different values of = 1,2, … , 30, compute the error for the validation data as cross-entropy (when computing log of probabilities add a small constant within log, e.g. 1e-15, to avoid numerical problems) and plot the dependence of the validation error on the value of . What is the optimal value here? Assuming that response has multinomial distribution, why might the cross-entropy be a more suitable choice of the error function than the misclassification error for this problem?

Assignment 2. Linear regression and ridge regression

Assignment 3. Logistic regression and basis function expansion

Appendix:

knearest.R

```r
# import data set
optdigits_data <- read.table('optdigits.csv', sep=",", header = 1)
head(optdigits_data, 5)

n=dim(optdigits_data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.5))
train_data=optdigits_data[id,]
id1=setdiff(1:n, id)
id2=sample(id1, floor(n*0.25))
valid_data=optdigits_data[id2,]
id3=setdiff(id1,id2)
test_data=optdigits_data[id3,]


# training data to fit 30-nearest neighbor classifier
```