

Exercises 02582
Module 2
Spring 2018

February 13, 2019

Topics: Model Selection

Exercises (coding tips at end of exercises):

- 1 Perform model selection for ridge regression (for the prostate data set):
 - (a) Consider using ridge-regression solutions for the prostate data set. What is a suitable range for the shrinkage parameter λ in which to search for an optimal solution in?
 - (b) Select a suitable value for the regularization parameter using K-fold cross-validation. Plot the resulting optimal value of lambda on a plot of the parameter trace (i.e. a plot of the $\hat{\beta}_j$ s as a function of λ).
 - (i) Try one of the common choices $K = 5$ and $K = 10$, and run the cross-validation a couple of times. Which would you prefer and why?
 - (ii) What is the value of K corresponding to leave-one-out cross-validation?
 - (c) Find a suitable value of λ using the one-standard-error rule. What is the difference between the two strategies (cross-validation and cross-validation with one-standard-error-rule)?
 - (d) Select suitable values for the regularization parameter using the AIC and BIC criteria (cf. 7.5-7.7 in ESL). What are the advantages and disadvantages of using cross-validation vs. information criteria?
 - (e) Use the bootstrap to estimate the variance of the parameters of the solution (β) for each value of lambda in exercise 1a. Plot the variance estimates as a function of lambda. What do you notice?

2 Perform model selection for KNN classification (for the silhouette data):

- (a) Load the dataset and plot the silhouettes.
 - (i) Matlab hint: Plot the silhouettes using `plot(Xa(:,1:65)', Xa(:,66:end)', '-')`. The variables `Male` and `Fem` contain row indices to male and female silhouettes.
 - (ii) R hint: Plot the silhouettes using `plot(X[1,1:65], X[1,66:p], type='l')`. The variable `class` contains information on whether it is a male or female silhouette.
 - (iii) Python hint: Plot the silhouettes using `plot(Xa[Fem,:65].T, Xa[Fem, 65:].T)`. The variables `Male` and `Fem` contain row indices to male and female silhouettes (with some adjustment, `mat['Fem'].ravel() - 1`, from Matlab indexing).
- (b) What size of fold would you use for cross-validation for the silhouette data set? Why?
- (c) Select a suitable number K for KNN classification on the silhouette profiles data using leave-one-out cross-validation.

3 Use the receiver operator curve (ROC) and determine specificity and sensitivity:

- (a) Make a function `[sens, spec] = roc_data(y, y_true, cut)` that takes as input the estimated response y , the true response y_{true} and the cut off value cut , and outputs the sensitivity and specificity.
 - (i) Matlab hint: `function [sens spec] = roc_data(y, y_true, cut)`.
 - (ii) R hint: `roccdata <- function(y, y_true, cut){ return(c(sens,spec))}`.
 - (iii) Python hint: Calculate sensitivity and specificity using the data in `X_ex2_3.txt` and `y_ex_2_3.txt`.
 - (iv) General hint: compute TP, TN, FP and FN. From there, sensitivity and specificity is easily computed.
- (b) Run the function `roc_gui.m` which takes your `roc_data` function as input. The output plot illustrates the data for the next exercise, and gives you a tool to adjust the cut off value for the classification rule, which is also plotted in the figure. Or simply plot the data in `X_ex2_3.txt` and `y_ex_2_3.txt`
- (c) Assume you are developing a mammography system for General Electrics, and that the GUI is showing the two features you have extracted to find suspicious image regions. GE has ordered the system to have a sensitivity of 95% to make sure very few lesions go undetected. Discuss this solution with your mates. What are you sacrificing to get such a sensitive system? Which sensitivity would you recommend based on the given data?

Exercises 02582
Module 2
Spring 2018

February 13, 2019

Topics: Model Selection

Resources for this exercise:

Listing 1: Resources in Matlab

```
Indices = crossvalind('Kfold', N, K); % create the random folds for CV
x_tr = x(Indices~=i,:); % the training data
trace(H); % trace of the matrix H
I = randi(N,M,1); % sample with replacement.
load silhouettes % load the data in silhouettes.mat (from CampusNet)
knn() % perform knn – file provided on CN
roc_gui.m % interface to play with Receiver Operator Curves (ROC)
```

Listing 2: Resources in R

```
library('cvTools') # load cvTools library (includes cvFolds)
folds <- cvFolds(n, K = 5, R = 1, type = "random") # folds for the CV
x_tr = x[folds$subsets[folds$which!=j],] # training data
library('psych') # load library psych (includes tr)
tr(H) # trace of the square matrix H
I = sample(1:n, size=n, replace = TRUE) # sample with replacement
library('class') # load library class (includes knn)
knn() # perform knn
```

Listing 3: Resources in Python

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.linalg as lng
import scipy.io
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve, auc

scipy.io.loadmat(path + '\\diabetes.mat') # load diabetes data
(i + 1) % K + 1 # use modulus K to get folds
np.random.permutation(N) # random permutation of N values
XTrain = X[i != I, :] # the training data
np.trace(outer) # trace of the matrix outer
np.random.randint(0, N, N) # sample with replacement.
scipy.io.loadmat(path + '\\Silhouettes.mat') # load silhouettes data
KNeighborsClassifier() # perform knn
```

End of exercise