

Assignment 3: Parameter Selection, Cross Validation and Feature Selection

Due: Oct. 6, 12pm (Parts 1 & 2)
Oct. 13, 12pm Part 3

In this assignment you will be extending your k-Nearest Neighbor classifier from **Assignment 1** to incorporate z-score normalization, parameter selection and feature selection.

1. Z-Score Normalization

You must use z-score normalized feature values rather than raw feature values for all sections of this assignment (unless otherwise specified). Write a function/method that implements z-score normalization on a set of examples S . Assume each example $\mathbf{x}_i \in S$ contains d numeric features, plus one class label.

The z-score is defined as: $z_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$, where $x_{i,j}$ is the i 'th example's value for feature j , μ_j is the mean of feature j , and σ_j is the sample standard deviation of feature j . Remember to divide by $N - 1$ when calculating the σ values (rather than N), so that the estimation is unbiased. Note that $N = |S|$. Use code libraries to calculate μ_j and σ_j where appropriate.

2. Parameter Selection: How to choose k ?

For this question, use `iris-prime.arff` (available on the course website) as the train/test dataset. Run leave-one-out cross validation (LOOCV) to calculate the classification accuracy achieved by kNN for $k = 1, 3, 5, \dots, 19, 21$ on this dataset.

- (a) What is the LOOCV classification accuracy for each value of k ?
- (b) What would be the best value(s) to choose for k if you wanted this classifier to achieve the highest possible classification accuracy on an unseen *test* dataset over the same problem domain?

- (c) Observe the accuracy rates on the raw (unnormalized) data. Did z-score normalization of the feature values result in improved kNN performance on this problem domain?
- (d) What does this imply about the utility of z-score normalization for a different classifier such as a Decision Tree? And for kNN on a different problem domain?

3. Feature Selection

Feature selection is used to remove irrelevant or correlated features in order to improve classification performance. You will be performing feature selection on a variant of the UCI *vehicle* dataset in the file `veh-prime.arff` (available on the course website). You will be comparing 3 different feature selection methods: the *Filter* method which doesn't make use of cross-validation performance, the *Wrapper* method which does, and finally one method that you will create.

Fix the kNN parameter to be $k = 7$ for all runs of LOOCV used in this question.

(a) Filter Method

Make the class labels numeric (set 'noncar'=0 and 'car'=1) and calculate the Pearson Correlation Coefficient (PCC) of each feature with the numeric class label. The PCC value is commonly referred to as r . For a simple method to calculate the PCC that is both computationally efficient and numerically stable, see the *Assignment 3* section of the course website.

- i. List the features from highest $|r|$ (the absolute value of r) to lowest, along with their $|r|$ values. Why would one be interested in the absolute value of r rather than the raw value?
- ii. Select the features that have the highest m values of $|r|$, and run LOOCV on the dataset restricted to only those m features. Which value of m gives the highest LOOCV classification accuracy, and what is the value of this optimal accuracy?

(b) **Wrapper Method**

Starting with the empty set of features, use a greedy approach to add the single feature that improves performance by the largest amount when added to the feature set. This is Sequential Forward Selection. Define performance as the LOOCV classification accuracy of the kNN classifier using only the features in the selection set (including the ‘candidate’ feature). Stop adding features only when there is no candidate that when added to the selection set increases the LOOCV accuracy.

- i. Show the set of selected features at each step, as it grows from size zero to its final size (increasing in size by exactly one feature at each step).
- ii. What is the LOOCV accuracy over the final set of selected features?

(c) **Your own method**

Create your own feature selection method using a combination of PCC and Wrapper. You may do anything that is computationally feasible within the constraint of turning the assignment in on time. Report your best method. Note that you do NOT have to beat the performance of the Wrapper or Filter methods presented. Implementing a single method with lower performance than either the Wrapper or Filter Method will still get you points.

- i. Describe your method and why you thought it would perform well.
- ii. Which features does it select?
- iii. What is the LOOCV accuracy over the selected features?

(d) What are the advantages and disadvantages of the *Filter* approach.

(e) What are the advantages and disadvantages of the *Wrapper* approach.

Submission and Due Date

Please ensure that you answer all parts of each problem.

Answers to the problems should be printed in hardcopy and submitted at the beginning of class. It is highly recommended that you do not submit handwritten work, as many of the problems ask for data which will necessarily be output from your program.

Submit your code via Provide:

Parts 1 & 2: `provide comp135 hw3-1 file1 file2 ... fileN.`

Part 3: `provide comp135 hw3-2 file1 file2 ... fileN.`

Parts 1 & 2 due: **October 6** at the start of class.

Part 3 is due: **October 13** at the start of class.