Assignment 1: KNN and Cross Validation

CS 725 (Foundations of Machine Learning), IITB

January 26, 2015

Introduction

- This assignment is aimed at making you familiar with the standard machine learning pipeline. It focuses on Knn algorithm and cross validation as discussed in the class.
- The assignment has three parts (10 + 10 + 5 points), described in section 1,2 and 3. Each section consists of an explanation section describing the dataset and a "To-do" section. The "To-do" section describes what needs to be done. Section 4 describes a complete submission.
- Please create a Moodle post for any clarification regarding this assignment.
- You are free to use **any** language, library and tool of your choice.
- This is an Individual assignment. Please do not copy. We have zero tolerance of cheating & plagiarism.

1 Cross validation and Leave one out validation, 10 points

1.1 Dataset

This is an image recognition dataset [1]. It consists of selective features for an image of English letters (A-Z). There are total **16** features corresponding to an image of a single letter. The format of the dataset is as follows:

- **class label**
  - lettr capital letter (26 values from A to Z)
- **features**
  - x-box horizontal position of box (integer)
  - y-box vertical position of box (integer)
  - width width of box (integer)
  - high height of box (integer)

[1] derived from: `http://archive.ics.uci.edu/ml/datasets/Letter+Recognition`

- onpix total # on pixels (integer)
- x-bar mean x of on pixels in box (integer)
- y-bar mean y of on pixels in box (integer)
- x2bar mean x variance (integer)
- y2bar mean y variance (integer)
- xybar mean x y correlation (integer)
- x2ybr mean of x * x * y (integer)
- xy2br mean of x * y * y (integer)
- x-ege mean edge count left to right (integer)
- xegvy correlation of x-ege with y (integer)
- y-ege mean edge count bottom to top (integer)
- yegvx correlation of y-ege with x (integer)

We have modified the original dataset in several ways, do not use the original dataset.

Dataset file: image_recognition.csv

1.2 To-do

Use different strategies of model selection to choose the best value of k for the given dataset. You are required to generate two plots, as explained below.

- The first plot: Leave one out average accuracy for different values of k (i.e. accuracy vs k for different k) (**5 points**)
- The second plot: 5-fold Cross validation accuracy for different values of k (the number of neighbors). (**5 points**) In each of the cases, also add a brief explanation for the behavior of the plots.

2 The R3 Dataset, 5 + 5 points

2.1 Format

The format is (x,y,z,class_label) where x,y,z ϵ R and class_label $\epsilon\{0, 1\}$. To be clear, x, y and z are the three features (real numbers) and the fourth column is the class label (0 or 1). For example, a row looks like:

```
 x,         y,         z,         class_label
-4.5338,    0.23073,  -0.39844,  1
```

2.2 Files

- **trainr3.csv**
  The training file.
- **testr30.csv**
  The first test file.
- **testr31.csv**
  The second test file.

2.3 To-do

Run KNN with training dataset trainr3.csv and generate two plots.

- Plot 1: Plot of accuracy vs k for different values of k on testr30.csv. (**5 points**)
- Plot 2: Plot of accuracy vs k different values of k on testr31.csv. (**5 points**)

Note that, $accuracy = \frac{\#samples\ in\ the\ test\ dataset\ correctly\ classified}{\#samples\ in\ the\ test\ set}$

Your final submission should consist of the plots along with a brief explanation.

3 Demo, 5 points

Check out the demo at `http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html`.

Step 2 and step 3 can be used to adjust the number of samples (m) and the number of neighbors (k) respectively.

3.1 To-do

Try picking various values of m and k. Can you relate your observations to the discussion about bayes optimality in class?

Add your comments in a file titled demo_inferences.txt.

4 Final Submission

Your final submission should be in the form: rollnumber_assignment1.zip. The zip file should contain the following:

- Four plots with explanations as above.
- Well commented code.
- demo_inferences.txt.