Introduction to Machine Learning: Autumn Semester 2013
Instructor: Prof. Andreas Krause

Course Project
Part 2

Classification of medical data

1 Introduction

You have seen several kinds of classification algorithms in class. In this project your task is to design a classification system for real medical data.

Magnetic resonance (MR) images of patients afflicted with Crohns Disease are annotated by clinical experts indicating the diseased regions, and also areas without any disease activity. When done manually, it is a tedious procedure which involves a considerable amount of effort from experts. Hence (semi-)automated systems are preferable in order to identify regions with similar disease activity in an unseen patient dataset. In order to design a robust system, different features are extracted from the images corresponding to the annotated regions, and used to train a classifier. When an unseen test dataset is provided, the corresponding features are extracted from different regions/pixels and put through the classifier for label prediction (diseased/non-diseased). Your task is to design a classifier which can reliably differentiate between diseased and normal samples.

2 Dataset Description

2.1 Input

Each data sample has 27 features derived from multiple neighborhoods around the annotated pixels. From each neighborhood 3 features are extracted - mean intensity, mean and variance of gradient values. We extract features across 9 neighborhoods ranging in size from 4×4 to 20×20.

2.2 Output

You are required to build a classifier to predict the class label of a given sample set. The output is either -1 (diseased) or $+1$ (normal).

2.3 Training Set

The data is formated as a comma-separated (CSV) file in which each line corresponds to a feature vector of 27 dimensions along with the class of the feature vector. In particular, for any line, the first 27 values correspond to the features of the image patch as described in Section 2.1. The 28^{th} value corresponds to the class of the image patch, and is either -1 (for the diseased class) or $+1$ (for the normal class). The training set is in the file "training.csv".

2.4 Validation and Test Sets

Both validation and test set contain feature vectors that have not been classified yet. Your task is to classify each feature vector as either belonging to the diseased class or normal class. You will be given several feature vectors each of length 27 and you are asked to predict the class for each feature vector. The data sets are given in the files "validation.csv" and "testing.csv". The formatting in both files is as follows:

- Same line format as the training set except that the class is not given (each line has only the 27 comma-separated configuration features).
- **Required output:** a file that contains the class predictions in order (the "i^{th}" line of the output file should contain the class prediction for the instance in the "i^{th}" row of the input file).

There are two submission pages: One for the validation set and one for the test set. You will receive feedback for each submission on the validation set and you can use it as a way to compare the prediction performance of your algorithm with other submissions. **You will not receive immediate feedback for the submissions on the test set: it will be used to calculate your grade.**

3 Evaluation and Grading

Each submission (upload of a prediction file for a given data set) will be ranked according to the classification error (CE) of the predictions. Here we shall employ a asymmetrical penalty function for mis-classifications i.e., false positives are penalized more than false negatives. This is illustrated according to the following table

True Label	Predicted Label	
	+1	-1
+1	0	1
-1	C=5	0

A false negative (FN) occurs when a positive sample is predicted to be negative, and a false positive (FP) occurs when a negative sample is predicted to be positive. In this scenario a false negative is acceptable but a false positive is highly undesirable since failure to detect a patient with disease can be very damaging. Therefore we choose $C > 1$ to impose higher penalty on false positives. For this project you may choose $C = 5$. Since we have the class prediction for each feature vector in the validation set and test set we can calculate

$$\text{CE} = \frac{5 \cdot |FP| + |FN|}{m},$$

where m denotes number of instances and $|FP|, |FN|$ denote the total number of false positives and false negatives respectively.

Now we compare the cost of the submission to two baseline predictions: a weak one (called "baseline easy") and a strong one (called "baseline hard"). These will have costs of CE_{BE} and CE_{BH} respectively, calculated as described above. Both baselines will appear in the rankings together with the error measure of your submitted predictions.

Performing better than the weak baseline on the **validation set** will give you 50% of the grade, and matching or exceeding the strong baseline on the **test set** will give you 100% of the grade. This allows you to check if you are getting at least 50% of the grade by looking at the ranking. If your prediction performance on the **test set** (CE_{test}) is in between the baselines, the grade is computed as:

$$\text{Grade} = \left(1 - \left(\frac{\text{CE}_{\text{test}} - \text{CE}_{\text{BH,test}}}{\text{CE}_{\text{BE,test}} - \text{CE}_{\text{BH,test}}}\right)\right) \times 50\% + 50\%$$

Your last submission on the test set will be used for grading.

3.1 Submission

In addition to your predictions on the test set you need to provide a brief report that explains how you obtained your results. We include a template for LaTeX in the file “report.tex”. If you do not want to use LaTeX, please use the same sections as shown in “report.pdf”.

Upload a zip file with the report (as a PDF file) along with your code or parameters/screenshots of the tools you used. For further instructions refer to the report template.

We might ask you to show us what you did, so please keep the necessary files until the end of the semester.

3.2 Deadline

You will be able to submit predictions starting from **Tuesday, 05.11.2013, 17:00** until **Sunday, 24.11.2013, 23:59:59**.