CS7616 Fall 2014 — Pattern Recognition
Problem Set 2a: KNN

DUE: Tuesday, March 4 at 11:55pm

Caution: PS 2b should be released by next tuesday and would be due on March 4 or March 6.

1 The Problem

In the lectures we discussed k-Nearest Neighbor based classification. In this assignment, we are going to use them. We discuss the methods briefly in the following, but refer to slides and textbook for more details on these techniques.

1.1 K-NN classifier

Using the notations from the class slide, for K-NN classifier we have

$$\hat{f}_{kNN}(x) = \arg\max_y \hat{p}_{kNN}(x|y)\hat{P}(y) = \arg\max_y \frac{k_y}{n_y \Delta_{k,x}} \frac{n_y}{N} = \arg\max_y k_y$$

Distance metric $d(X^i, X^j) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ between data points $\{X^i\} \in \mathbb{R}^n$ defines a measure of *similarity*. So distance between data-points belonging to same cluster should be lower than that between points belonging to separate clusters. Distance metric plays an important role for successful classification or density estimation. And its true for K-NN also. You are encouraged to try different distance metric for your K-NN classifier apart from the usual Euclidean distance metric $d(X^i, X^j) = \|X^i - X^j\| = \sqrt{(X^i - X^j)^T(X^i - X^j)}$. You are required to scale the different dimensions of the data differently. We discussed whitening the data in the previous assignment and people reported that it made little to no difference. That can be a starting point for doing the scaling, but you can imagine that for determining neighbors certain dimensions are more important and thus should be weighed more. Another method [1, 2] is to learn the distance metric from the data itself. Usually we choose fixed type of distance metric like Mahalanobis distance $d(X^i, X^j) = \|X^i - X^j\|_M = \sqrt{(X^i - X^j)^T M (X^i - X^j)}$ and we try to learn the parameters of the positive semi-definite matrix M. Learning the Mahalanobis distance metric is equivalent to ïňĄnding a rescaling of a data that replaces each point X with $M^{\frac{1}{2}}X$. You can try either scaling the data manually as suggested in previous assignment or use metric learning algorithms like [1, 2]. You can use code available online for [1].

We discussed in class the effect different k can have on the classification. Cross-validation can be used to decide between different values for k.

In this problem set you will also try to estimate the test-error from training samples. This you will do through cross-validation. Then you will compare these estimates to the actual test error. You

can create a test set by randomly selecting part of the dataset and holding them out of training and cross-validation.

2 Datasets

For this Problem Set you would be required to do the following for each of the three datasets listed later in this PS.

1. Randomly select 20% of the dataset for test data (TtD). The rest will be considered the training data (TD).
2. Randomly select a subset of d% from the TD where $d = \{20, 50, 80, 100\}$. For the first three cases (except the 100% case) you should generate five training sets (TS) for each of the d% data.
3. For each of the TS, do **F-fold** cross-validation (CV) on each of the generated TS. You are expected to do 2-fold, 5-fold and Leave-one-out cross validation to decide the best K. Report the CV error for the best K.
4. Now, test the TtD with the best K and its associated TD and compare the error rate to CV error. Report which F worked best.
5. Report on the stability of the results with respect to the TS of the same size.
6. Report on the results with respect to the size of the TS.

Following are the three datasets.

Dataset 1

You are required to use the **Alcoholism Dataset** and classify a person as being an alcoholic or not. [DESCR] [TRAINING-DATA] [TEST-DATA]
For this case, use you don't need to generate test data.

Dataset 2

You have to choose one dataset from the following list.

- Bank-Note Authentication [DESCR] [DATA]
- Diabetes in Pima Indians [DESCR] [DATA]
- Phoneme [DESCR] [DATA]
- Skin Color [DESCR] [DATA]
- Music Genre [DESCR] [DATA]
- Spam [DESCR & DATA]
- Eye-State [DESC] [DATA]

Dataset 3

You can pick another dataset from the previous list or use any other reasonable (according to suggestions of the previous assignment) dataset that you can find (or make).

3 Submit

A zip file with a PDF report and all of your code. You can use any programming language, but we will prefer if it is `Matlab` or `Python` or `C++`. The code should be easily runnable. Please include a README file that describes how the TAs are to run your code.

The Report should contain the following for each of the three datasets.

- A concise description of the methods you used that seemed to work. In this case you might want to talk about scaling or any distance metrics you tried.
- The accuracy you achieved for each of the cases described organized in tables and figures (if that helps).
- The results asked for in the previous section.
- A short description of the results, you should discuss what results remained consistent across datasets and what didn't.
- A succinct explanation of the why you think you achieved those results. Hypothesize why certain techniques or cases worked out better than the other. If you have data/plots to support your claim, even better.
- If part of your code was taken from someone else, please mention it here as well. You are encouraged to discuss and help others with anything short of giving them your code. We hope you use Piazza for this so more people benefit.

References

[1] Kilian Q. Weinberger and Lawrence K. Saul *Distance Metric Learning for Large Margin Nearest Neighbor Classification.* JMLR, 2009.

[2] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan and Stuart Russell *Distance metric learning, with application to clustering with side-information.* NIPS, 2002.