CSE250B Homework 1

Qiao Zhang

January 15, 2016

1 Prototype selection

1.1 Description

Condensation 1 approach:
For each point in the training set, we can find 2 nearest neighbors where the 1-nn is usually itself and the 2-nn is its true nearest neighbor. If they share the same label, we can say that this data point can be easily recognized by nearest neighbor classification. We want to put as many points which are difficult to recognize and near the decision boundary in the prototype as possible.
Condensation 2 approach:
The prototype should contain enough interior points which helps recognize regular test images in addition to points described in Condensation 1. Due to the former 30000 points are easier to recognize, we need to randomly shuffle the data to ensure better performance. This approach can be seen as a modification to the first method.
Kmeans approach:
We first divide the train data into 10 categories by their labels. Within each category, we group the points into M/10 clusters using k-means algorithm. Finally we can obtain (M/10) * 10 centers, which forms the prototype naturally.

1.2 Pseudo code

```
Input: training set - X, training label - Y, subset size - M
Output: prototype
Algorithm:
    for x,y in zip(X,Y)
        neighbor1 = firstNearestNeighbor(x)
        neighbor2 = secondNearestNeighbor(x)
        if neighbor1.label == neighbor2.label
            add x,y to prototype_same
        else
            add x,y to prototype_diff
    Condensation 1
    prototype = prototype_diff + prototype_same
    prototype = prototype[:M]
    Condensation 2
    prototype = randomly_shuffled_prototype_diff[:0.1M] + randomly_shuffled_prototype_same[:0.9M]
```

```
kmeans
prototype = []
for l = 0:9
    X,Y = train data with label of l
    prototype += M / 10 centroids of X using kmeans
```

1.3 Performance

Test error = $\frac{wrongly\ classified}{correctly\ classified + wrongly\ classified}$
The result is reported in the following table:

M	uniform-random	condensation1	condensation2	kmeans
10000	0.0508	0.0558	0.0501	0.0288
5000	0.0651	0.0824	0.0648	0.034
1000	0.1197	0.6145	0.1206	0.0431

Table 1: Testing error using 1NN

We can see from the table of result that the condensation1 approach works even worse than uniform-random selection when M = 1000, 5000, 10000. This provided approach intends to retain the data points which are closer to the boundary decision, in other words, difficult to classify while removes the interior data points, thus speed up the nearest neighbor search. But the problem is: (1) the prototype must contain enough interior points to ensure basic performance, (2)the former 30000 data points are easier to recognize thus shuffling is needed. Thus I obtained a better condensation2 solution, which is barely better than uniform-random.
But the k-means approach promote the performance profoundly. The error rate is reduced by around 0.075 when M = 1000.

1.4 Reference

http://www.ijarcce.com/upload/2013/december/IJARCCE8D-s-shikha-Prototype_Selection.pdf

2 Bayes Optimality

2.1

The Bayes classifier is

$$h(x) = \begin{cases} 0 & if\ x < -0.5 \\ 1 & if\ -0.5 \leq x \leq 0.5 \\ 0 & if\ x > 0.5 \end{cases}$$

The optimal risk is

$R^* = 0.2 \times \frac{3}{8} + 0.2 \times \frac{2}{8} + 0.4 \times \frac{3}{8} = 0.275$

2.2

The decision boundary of 1-NN is -0.6 and 0.5.

$$h(x) = \begin{cases} 0 & if\ x < 0.6 \\ 1 & if\ -0.6 \leq x \leq 0.5 \\ 0 & if\ x > 0.5 \end{cases}$$

Error rate = $P_r(h(X) = 1, Y = 0) + P_r(h(X) = 0, Y = 1) = \int_{-1}^{-0.6} 0.2\|x\|dx + \int_{0.5}^{1} 0.4\|x\|dx + \int_{-0.6}^{-0.5} 0.8\|x\|dx + \int_{-0.5}^{0.5} 0.2\|x\|dx = 0.308$

2.3

Let the classifier be as follows:

$$h(x) = \begin{cases} 0 & if\ x < a \\ 1 & if\ a \leq x \leq b \\ 0 & if\ x > b \end{cases}$$

Because $c_{01} > c_{10}$, $a \leq -0.5$ and $b \geq 0.5$.

Error rate = $P_r(h(X) = 1, Y = 0) + 0.1P_r(h(X) = 0, Y = 1) = \int_{-1}^{a} 0.2\|x\|dx + \int_{b}^{1} 0.4\|x\|dx + \int_{a}^{-0.5} 0.8\|x\|dx + \int_{-0.5}^{0.5} 0.2\|x\|dx + \int_{0.5}^{b} 0.6\|x\|dx = 0.1b^2 + 0.3a^2 + 0.175$

In order to minimize error rate, the absolute value of a and b should be as small as possible. Thus a = -0.5 and b = 0.5.

2.4

The expression should be

$$h(x) = \begin{cases} 0 & if\ c_{01}(1 - \eta(x)) > c_{10}\eta(x) \\ 1 & if\ c_{01}(1 - \eta(x)) \leq c_{10}\eta(x) \end{cases}$$

This expression applies to 2.3.

3 Metrics

3.1 l_1 distance is metric

$\because$ d(x,y) = $\sum_{i=1}^{m} |x_i - y_i|$
$\therefore$ d(x,y) $\geq$ 0
If d(x,y) = 0,
$\therefore x_i = y_i \quad \forall i$
$\therefore$ x = y
$\because |x_i - y_i| = |y_i - x_i|$
$\therefore$ d(x,y) = d(y,x)
$\because |x_i - z_i| \leq |x_i - y_i| + |y_i - z_i| \quad \forall i$
$\therefore$ d(x,z) $\leq$ d(x,y) + d(y,z)

3.2 $d_1 + d_2$ distance is metric

$\because d_1, d_2$ are metrics
$\therefore d_1(x, y) \geq 0$, $d_2(x, y) \geq 0$
$\therefore d_1(x, y) = 0 \Rightarrow x = y$, $d_2(x, y) = 0 \Rightarrow x = y$
$\therefore d_1(x, y) = d_1(y, x)$, $d_2(x, y) = d_2(y, x)$
$\therefore d_1(x, z) \leq d_1(x, y) + d_1(y, z)$, $d_2(x, z) \leq d_2(x, y) + d_2(y, z)$
$\because d_1(x, y) \geq 0$, $d_2(x, y) \geq 0$
$\therefore (d_1 + d_2)(x, y) \geq 0$
$\therefore d_1(x, y) + d_2(x, y) = 0 \Rightarrow d_1(x, y) = 0$ and $d_2(x, y) = 0$
$\because d_1(x, y) = 0 \Rightarrow x = y$, $d_2(x, y) = 0 \Rightarrow x = y$
$\therefore (d_1 + d_2)(x, y) = 0 \Rightarrow x = y$
$\because d_1(x, y) = d_1(y, x)$, $d_2(x, y) = d_2(y, x)$
$\therefore (d_1 + d_2)(x, y) = (d_1 + d_2)(y, x)$
$\because d_1(x, z) \leq d_1(x, y) + d_1(y, z)$, $d_2(x, z) \leq d_2(x, y) + d_2(y, z)$
$\therefore (d_1 + d_2)(x, z) \leq (d_1 + d_2)(x, y) + (d_1 + d_2)(y, z) \forall x, y, z$

3.3 hamming distance is metric

$\because$ d(x,y) = # positions on which x and y differ
$\therefore d(x, y) \geq 0$
If d(x,y) = 0
$\therefore$ # positions on which x and y differ = 0
$\therefore x_i = y_i \quad \forall i$ x = y
$\because$ # positions on which x and y differ = # positions on which y and x differ
$\therefore$ d(x,y) = d(y,x)
$\because$ # positions on which x and z are identical $\geq$ # positions on which x and y are identical + # positions on which y and z are identical - m
$\therefore$ m - # positions on which x and z are identical $\leq$ 2m - (# positions on which x and y are identical + # positions on which y and z are identical)
$\therefore$ d(x,z) $\leq$ d(x,y) + d(y,z)

3.4 squared euclidean distance is not metric

Counterexample:
Let m = 1, x = 1, y = 2, z = 3
d(x,y) = 1, d(x,z) = 4, d(y,z) = 1
$\because$ d(x,z) $\geq$ d(x,y) + d(y,z)
$\therefore$ the inequality property of metric space does not suffice

3.5 Kullback-Leibler distance is not metric

Counterexample:
Let m = 2, p = [0.5, 0.5], q = [0.1, 0.9]
K(p,q) = 0.5(log0.5 - log0.1) + 0.5(log0.5 - log0.9) = 0.2218
K(q,p) = 0.1(log0.1 - log0.5) + 0.9(log0.9 - log0.5) = 0.1598
$\because$ K(p,q) $\neq$ K(q,p)
$\therefore$ the symmetric property of metric space does not suffice