THU-70250043-0, Pattern Recognition (Spring 2021) **Homework: 6**

Metric and k-NN

Lecturer: Changshui Zhang `zcs@mail.tsinghua.edu.cn`

Hong Zhao `vzhao@tsinghua.edu.cn`

Student:

1 Property of Euclidean Distance

When the metric space is a finite-dimensional Euclidean space, please **prove** that the Voronoi cells induced by the single-nearest neighbor algorithm must always be convex. Does this property hold when the metric becomes Manhattan distance?

In mathematics, a Voronoi diagram is a partition of a plane into regions close to each of a given set of objects. In the simplest case, these objects are just finitely many points in the plane (called seeds, sites, or generators). For each seed there is a corresponding region consisting of all points of the plane closer to that seed than to any other. These regions are called Voronoi cells, as shown in Figure 1. Similarly, Voronoi cells of a discrete set in higher-order Euclidean space are known as generalized polyhedra.

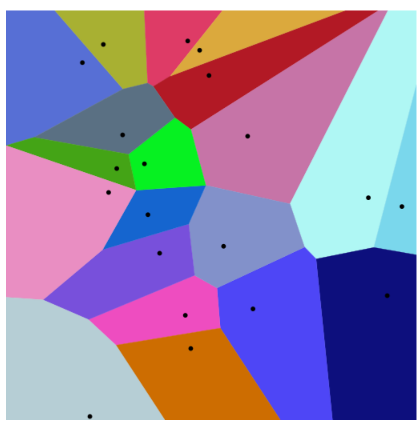

Figure 1: Voronoi cells with Euclidean distance, cited from [1].

Hint: Convex means for any two points x_1 and x_2 in a cell, all points on the segment linking x_1 and x_2 must also lie in the cell.

2 Properties of Metric

Please **prove** that the Minkowski metric indeed possesses the three properties required of all metrics.

Hint: A metric $D(,)$ must have three properties: for all vectors a, b and c,

1. *identity of indiscernibles:* $D(a, b) = 0$ *if and only if* $a = b$.
2. *symmetry:* $D(a, b) = D(b, a)$.
3. *triangle inequality:* $D(a, b) + D(b, c) \geq D(a, c)$.

3 k-NN Classifier

Let $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ be a set of n independent labelled samples and let $\mathcal{D}_k(\boldsymbol{x}) = \{\boldsymbol{x}'_1, \cdots, \boldsymbol{x}'_k\}$ be the k nearest neighbors of $\boldsymbol{x}$. Recall that the k-nearest-neighbor rule for classifying $\boldsymbol{x}$ is to give $\boldsymbol{x}$ the label most frequently represented in $\mathcal{D}_k(\boldsymbol{x})$. Consider a two-category problem with $P(\omega_1) = P(\omega_2) = 1/2$. Assume further that the conditional densities $p(x|\omega_i)$ are uniform within unit hyperspheres, and the two categories center on two points ten units apart. Figure 2 shows a diagram of this situation.

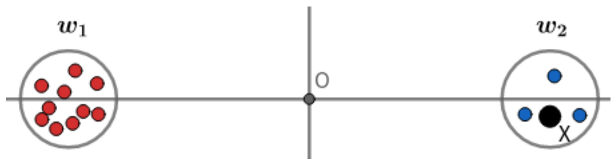

Figure 2: A diagram of assumed situation. When $k \geq 7$, X is misclassified as there are only 3 samples in w_2.

1. Show that if k is odd, the average probability of error is given by

$$P_n(e) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}.$$

2. Show that for this case the single-nearest neighbor rule has a lower error rate than the k-nearest-neighbor error rate for $k > 1$.
3. If k is odd and is allowed to increase with n but is restricted by $k < a\sqrt{n}$, where a is a positive constant, show that $P_n(e) \to 0$ as $n \to \infty$.

4 Programming: k-NN Classifier on MNIST

Please implement k-NN classifier and run on MNIST[2]. You need to follow the official train/test split of MNIST. Compare the performance with the following settings:

- Using 100, 300, 1000, 3000, 10000 training samples.
- Using different values of k.
- Using at least three different distance metrics.

In this assignment, you are *NOT* allowed to use any existing libraries or code snippets that provides k-NN algorithm.

5 Literature Reading

Please read the paper about metric learning.

Distance metric learning with application to clustering with side-information.[3]

Hint: You do not have to submit anything for this reading section.

References

[1] File:Euclidean Voronoi diagram.svg. (2020, October 10). Wikimedia Commons, the free media repository. Retrieved 09:39, March 29, 2021 from

`https://commons.wikimedia.org/wiki/File:Euclidean_Voronoi_diagram.svg`

[2] LeCun Y. The MNIST database of handwritten digits[J]. `http://yann.lecun.com/exdb/mnist/`, 1998.

[3] Xing E P, Ng A Y, Jordan M I, et al. Distance metric learning with application to clustering with side-information[C]//NIPS. 2002, 15(505–512): 12.