CHAPTER II

BACKGROUND

2.1 Conventions

Random variables are uppercase characters such as X, Y, and Z.

We usually denote a time series or a finite dimensional vector by a boldface letter such as $\mathbf{s}$, and its length by $\#\mathbf{s}$ or just l when the mentioned time series is obvious. The i-th value of the time series $\mathbf{s}$ is written s_i or $\mathbf{s}\langle i\rangle$, and by $\mathbf{s}$ and $[s_1, \ldots, s_l]$ we mean the same thing.

The time series $[s_2, \ldots, s_l]$ is called the *tail* of $\mathbf{s}$, denoted by $\mathbf{s}_\sim$. $\mathbf{0}$ and $[]$ are $[0]$ and the time series of length zero, respectively. There is one and only one place, in Section 2.4, where the square brackets enclosing a letter $[x]$ will be used to denote the equivalence class of x and we do not mention time series there.

Functions that change vectors or time series or one object to another, called *morphs*, are denoted by Greek letters using prefix notation. For example, $\mu(x)$ or μx is understood as the morphed object from x by the morph μ. Compositions of functions such as $\mu(\nu(x))$ may be written as $\mu\nu x$ or $\mu \circ \nu x$. $\mathbb{1}$ denotes the identity map.

Calligraphic scripts such as $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{M}$ are used to denote sets of functions. For a set $\mathcal{F}$ of functions on a space Ω and $x \in \Omega$, we let $\mathcal{F}(x)$ be the set $\{f(x) | f \in \mathcal{F}\}$.

We denote infinite sequences by the list of its elements enclosed in a parentheses e.g. $(1, 2, 3, 4, \ldots)$. Depending on the context, sometimes we regard a finite sequence as an infinite sequence entailed with zeros, or as an infinite sequence entailed with a constant sequence of its last element. For example, we may think of $[0.5, 1]$ as $(0.5, 1, 0, 0, 0, \ldots)$ or $(0.5, 1, 1, 1, \ldots)$, subject to the context.

Given a finite sequence $\mathbf{s}$ and a another sequence $\mathbf{t}$, the concatenation of $\mathbf{s}$ and $\mathbf{t}$ is written as $\mathbf{st}$, $\mathbf{s}^1$ is the same as $\mathbf{s}$ and $\mathbf{s}^n$ is defined recursively as $\mathbf{ss}^{n-1}$.

2.2 Classification Problem Settings

We follow the same setting as in the work of Devroye et al. (1996). The c-class classification problem in a probabilistic setting is formalized as follows. Let (X, Y) be a pair of random variables taking values in the Cartesian product $\Omega \times \Lambda$ of a metric space Ω of all possible examples and the class labels $\Lambda = \{1, \ldots, c\}$. A function $f : \Omega \to \Lambda$ deciding the class label based solely on the observation of examples from Ω is called a *classifier*. A *rule*, upon a given finite set of i.i.d. pairs of values observed from the random pair (X, Y), constructs a classifier. For example, in the case of 1-NN rule, given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, it constructs the decision function

$$g_n(x) = y_k \ ,$$

where x_k is closest to x.

An error occurs if $f(X) \neq Y$, and the probability of error for a classifier f is

$$L(f) = \mathbf{P}\{f(X) \neq Y\} \ .$$

For a fixed rule, the classifier f_n constructed according to n observations from the random pair (X, Y) depends randomly on the data sequence, so as the conditional probability of error

$$L_n = L(f_n) = \mathbf{P}\{f_n(X) \neq Y | X_1, Y_1, \ldots, X_n, y_n\} \ .$$

2.3 Bayes Classifier

The Bayes classifier is the following decision function

$$g^*(x) = \text{argmax}_{i \in \Lambda} \mathbf{P}\{Y = 1 | X = x\} \ .$$

It can be shown (Devroye et al., 1996, chap. 2) that for any classifier g,

$$L^* = \mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\} \ .$$

2.4 Distance, Metric and Norm

The asymptotic results of k-NN contains different assumptions on the distance measure and the probability distribution of the data. In order to get a good grasp of the different assumptions on the distance measures, we begin with the definition of metric space and its relatives.

Definition 1. *A* metric space (Ω, d) *is a set* Ω *together with a non-negative extended-real-valued function* $d : \Omega \times \Omega \to [0, \infty]$ *(called a* metric*) such that, for every* $x, y, z \in \Omega$,

i) $d(x, x) = 0$,

ii) $d(x, y) = 0$ *implies* $x = y$,

iii) $d(x, y) = d(y, x)$,

iv) $d(x, z) \leq d(x, y) + d(y, z)$.

If the condition (ii) is omitted then we have a pseudometric space *and* d *will be called a* pseudometric. *If the condition (ii) and (iv) are dropped, we have a* distance space *and call* d *a* distance.

Note that we allow d to take the value ∞ so that the definition of a metric and its relatives are technically applicable in a wider situation. For example, if some pair of elements in Ω are not comparable, we let their distance be ∞.

Pseudometric space is a salient structure to perform nearest neighbor queries. For a pseudometric space, numerous techniques (Roussopoulos et al., 1995; Barros et al., 1996; Ciaccia et al., 1997; Dohnal et al., 2003; Guttman, 1984) could be readily applied to speed up nearest neighbor queries, and sometimes k-means

algorithms. Those works are based on the following bounding scheme or their variants, each of them is derivable from the triangle law,

$$d(x,z) \geq |d(x,y) - d(y,z)| \ ,$$
$$d(x,y) \geq \frac{1}{2}d(x,z) - |d(y,z) - \frac{1}{2}d(x,z)| \ .$$

Examples of pseudometric spaces are, the real numbers with absolute difference, vector spaces with the Euclidean distance, a set of strings with edit distance, etc.

Having a pseudometric space (Ω, ρ) with a pseudometric ρ, we can always have a metric space by gluing together elements in Ω that are zero distance apart together. Precisely, we construct a *quotient space* of it using the equivalence relation

$$x \sim y \quad \text{iff} \quad \rho(x,y) = 0 \ .$$

Then we can think of an equivalence class as one point in the new space and define a new metric in the space $\Omega/\sim$ of equivalence classes,

$$\tilde{\rho}([x],[y]) = \rho(x,y) \ .$$

The new space $(\Omega/\sim, \tilde{\rho})$ is a metric space.

It is not uncommon that one encounters norms when working with vector spaces. Since we may think of a set of time series as a vector space of number sequences, norms are involved naturally. Some of the asymptotic results for k-NN hold for norms metrics. They will also be mention later in Chapter 3.

Definition 2. *Let V be a vector space. A function $\|\cdot\| : V \to [0,\infty)$ is a norm on V iff for every $x, y \in V$*

i) $\|x\| = 0$ *iff* $x = 0$,

ii) $\|\alpha x\| = |\alpha|\|x\|$ *for any* $\alpha \in \mathbb{R}$,

iii) $\|x + y\| \leq \|x\| + \|y\|$.

Having a normed space, the function $(x, y) \mapsto \|x - y\|$ is always a metric. Well known norms for the space of number sequences are the ℓ^p norms defined by $\|\mathbf{x}\|_p = \{\sum_{i=1}^{l} |x_i|^p\}^{\frac{1}{p}}$, for $p \in [1, \infty)$ and $\|\mathbf{x}\|_\infty = \max_i |x_i|$. The latter is called the *supremum norm*. The metrics induced by ℓ^p norms are called ℓ^p metrics.

2.5 Asymptotic Behavior of Metric Based k-NN

In terms of generality, are two major asymptotic results, the first holds for separable metric spaces but with a usual assumptions on the distribution.

Theorem 1 (Cover and Hart (1967)). *Let Ω be a separable metric space and (X, Y) admits class conditional densities. Let $f_1, \ldots, f_c$ be probability densities such that $f_i(x) = \mathbf{P}\{X = x | Y = y\}$ and f_i is continuous almost everywhere for each $i \in \{1, \ldots, c\}$. Then the k-NN probability of error L has the bounds*

$$L^* \leq \lim_{n \to \infty} \mathbf{E}\, L_n \leq L^* \left(2 - \frac{cL^*}{c-1}\right) \leq 2L^* \ .$$

These bounds are as tight as possible.

The second major result holds for every possible distribution but the distance is assumed to be a norm metric (Devroye et al., 1996, prob. 5.1, chap. 5).

Theorem 2 (Devroye et al. (1996), chap. 5). *Let the random pair (X, Y) take values in $\mathbb{R}^d \times \{1, 2\}$. Then the norm-metric based k-NN probability of error L has the bounds*

$$L^* \leq \lim_{n \to \infty} \mathbf{E}\, L_n \leq L^* (2 - 2L^*) \leq 2L^* \ .$$

Since our pseudometrics in Chapter 3 are not norm metrics we have to hinge on the former theorem and assume the regularity of the distributions. Note that a quick argument that the set of all time series with all rational values serves as a countable dense subset will be sufficient to establish that all of the pseudometric spaces of time series in Chapter 3 are separable. Such argument ensures that the space of time series equipped with DTW as the distance is a separable distance space, although not a metric space.

2.6 Admissibility of 1-NN

Cover and Hart (1967) showed in their paper that 1-NN is admissible in the sense that there is a distribution of data such that k-NN will be strictly worse than 1-NN in terms of probability of misclassification for every $k > 1$. They also give an example of such distribution in the paper and noted that if the between-class distances are always greater than the within-class distances then 1-NN is strictly better than any other k-NN.

2.7 DTW Distance

In computing the DTW distance, one searches for a *warping path* with the lowest possible associated cost. Given two finite sequences of real numbers $\mathbf{s}$ and $\mathbf{t}$, a warping path between $\mathbf{s}$ and $\mathbf{t}$ is a sequence of pairs

$$(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N) \ . \tag{2.1}$$

Where $N \leq \#\mathbf{s} + \#\mathbf{t} - 1$, $(i_1, j_1) = (1, 1)$, $(i_N, j_N) = (\#\mathbf{s}, \#\mathbf{t})$, and (i_k, j_k) must be one of $(i_{k-1}, j_{k-1}), (i_{k-1}, j_k)$ or (i_k, j_{k-1}) for all $2 \leq k \leq N$.

One may perceive a warping path as a continuous monotonic sequence of coordinates whose start and end are fixed in a two dimensional grid. The cost associated with a warping path in Equation (2.1) is $\sum_{k=1}^{N} d(s_{i_k} - t_{j_k})$, where d is any distance measure — common choices are absolute difference and squared difference. Figure 2.1 shows an example of the optimal warping path of two time series $[1, 0, 0]$ and $[1, 2, 0]$, the path is $(1, 1), (2, 1), (3, 2), (3, 3)$ from the bottom left of the grid to top right. Suppose for concreteness that d is absolute difference. The DTW distance can be expressed in terms of its partial solutions as,

$$\mathrm{DTW}(\mathbf{s}, \mathbf{t}) = |s_1 - t_1|^p + \min \begin{cases} \mathrm{DTW}(\mathbf{s}_\sim, \mathbf{t}_\sim), \\ \mathrm{DTW}(\mathbf{s}, \mathbf{t}_\sim), \\ \mathrm{DTW}(\mathbf{s}_\sim, \mathbf{t}) \ . \end{cases}$$

Where $\mathrm{DTW}([s_1], \mathbf{t}) = \mathrm{DTW}(\mathbf{t}, [s_1]) = \sum_{i=1}^{\#\mathbf{t}} |s_1 - t_i|^p$ for the base cases and p is usually 1 or 2 as mentioned above. $\mathrm{DTW}(\mathbf{s}, \mathbf{t})$ can be computed in $O(\#\mathbf{s}\#\mathbf{t})$

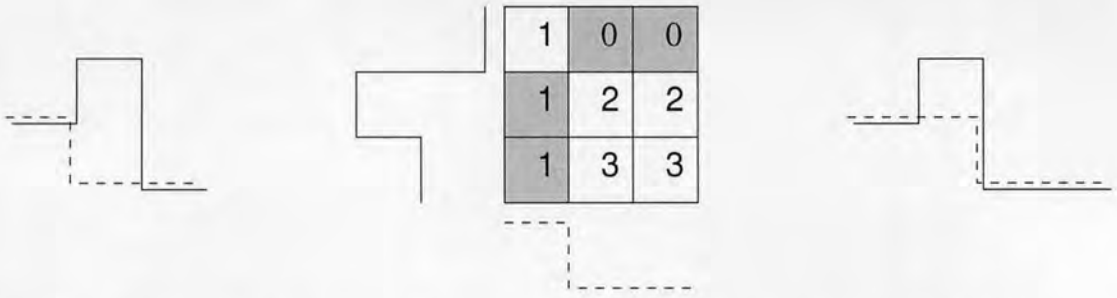

Figure 2.1: A visualization of a warping path, which is also an optimal warping path whose associated cost is 1. The path is $(1,1),(2,1),(3,2),(3,3)$ from bottom left to top right. On the left, the dotted line is the sequence $[1,2,0]$ and other in solid line is the sequence $[1,0,0]$. On the right, the dotted line is the sequence $[1,2,0,0]$, which is a *stretch* (see Definition 6) of $[1,2,0]$; i.e. $[1,2,0,0] \in \mathcal{S}([1,2,0])$. The solid line on the right is $[1,1,0,0]$, which is a stretch of $[1,1,0]$. Note that the distance between $[1,2,0,0]$ and $[1,1,0,0]$ is equal to the cost of the optimal warping path.

DTW-DISTANCE$(A[1..n], B[1..m])$
1 $W[0][0] \leftarrow 0$
2 $W[0][1..m], W[1..n][0] \leftarrow \infty$
3 **for** $i \leftarrow 1$ **to** n
4 **do for** $j \leftarrow 1$ **to** m
5 **do** $d \leftarrow \min\{W[i][j-1], W[i-1][j], W[i-1][j-1]\}$
6 $W[i][j] \leftarrow |A[i] - B[j]|^p + d$
7 **return** $W[n][m]$

Figure 2.2: Pseudocode of the DTW algorithm.

time.

By the expression above, the DTW distance can be computed by dynamic programming paradigm. More detailed treatment of the subject can be found in other sources (Keogh and Ratanamahatana, 2004; Sakoe and Chiba, 1978; Itakura, 1975; Rabiner et al., 1978).

2.7.1 Non-Subadditivity of DTW

To see that the DTW distance is not a pseudometric, consider the following trivial example. Let $\mathbf{s} = [1,1]$ and $\mathbf{t} = [1,1,1]$, then $\mathrm{DTW}(\mathbf{0},\mathbf{s}) + \mathrm{DTW}(\mathbf{s},\mathbf{t}) = 2 + 0 = 2$, while $\mathrm{DTW}(\mathbf{0},\mathbf{t}) = 3$. Another interesting example is when $\mathbf{u} =$

$(1,0,0), \mathbf{v} = (1,2,0)$. We have $\mathrm{DTW}(\mathbf{0}, \mathbf{v}) = 3 > 2 = \mathrm{DTW}(\mathbf{0}, \mathbf{u}) + \mathrm{DTW}(\mathbf{u}, \mathbf{v})$. These demonstrate that the DTW distance is not subadditive, and hence not a pseudometric.

2.8 Levenshtein Distance

The Levenshtein distance is a metric used to measure difference between two strings. The following relation may be taken as its definition

$$\mathrm{Lev}(\mathbf{s}, \mathbf{t}) = \rho(s_1, t_1) + \min \begin{cases} \mathrm{Lev}(\mathbf{s}_\sim, \mathbf{t}_\sim), \\ \mathrm{Lev}(\mathbf{s}, \mathbf{t}_\sim), \\ \mathrm{Lev}(\mathbf{s}_\sim, \mathbf{t}) \ . \end{cases}$$

Where the function $\rho(x, y)$ is the *discrete metric* taking value 0 if x equals y and 1 otherwise. $\mathrm{Lev}([], []) = 0$ and $\mathrm{Lev}([], [s_1]) = 1$ for the base cases.

The Levenshtein distance between two strings is the minimum number of operations needed to transform one string into the other, where an operation is an insertion, deletion, or substitution of a single character.

The distance is subadditive, indeed it is a metric, one way to see this is by the fact that the distance is the minimum number of operations needed to transform one string to the other. For strings $\mathbf{s}$, $\mathbf{t}$ and $\mathbf{u}$, the sum $\mathrm{Lev}(\mathbf{s}, \mathbf{u}) + \mathrm{Lev}(\mathbf{u}, \mathbf{t})$ is the number of an operation sequence that transforms $\mathbf{s}$ to $\mathbf{t}$ (by changing $\mathbf{s}$ to $\mathbf{u}$ and then to $\mathbf{t}$), but that number is never greater than $\mathrm{Lev}\,\mathbf{s}, \mathbf{t}$ which is the minimum the length of such operations.

2.9 Edit Distance with Real Penalty

Edit Distance with Real Penalty (ERP) (Chen and Ng, 2004) is adapted from the Levenshtein distance. It is subadditive via a result for edit distance by Waterman et al. (1976).

For a real valued γ called "gap" ERP is defined by

$$\mathrm{ERP}(\mathbf{s},\mathbf{t}) = \min \begin{cases} |s_1 - t_1| + \mathrm{ERP}(\mathbf{s}_\sim, \mathbf{t}_\sim), \\ |\gamma - t_1| + \mathrm{ERP}(\mathbf{s}, \mathbf{t}_\sim), \\ |s_1 - \gamma| + \mathrm{ERP}(\mathbf{s}_\sim, \mathbf{t}) \ . \end{cases} \tag{2.2}$$

Where $\mathrm{ERP}([], \mathbf{s}) = \mathrm{ERP}(\mathbf{s}, []) = \sum_{i=1}^{l} |\gamma - s_i|$ for the base cases.

The value γ of the gap can be thought of as the default value in the sense that the constant sequence of γ, $(\gamma, \gamma, \ldots)$ is the null signal. It usually makes sense that the gap value is set to zero in practice because we usually perceive the null signal as a sequence of zeros. A side benefit is that we do not need to compute the difference of the gap value and the element of another sequence if the gap is zero.