Q1(3) From the Warping Path, draw a figure to show how these two data points are aligned.

Ans.) The Warping Path between p1 and p3 is: [(0,0), (0,1), (0,2), (1,2), (2,3), (3,4)]

Q1(4) Give an Example to show that the Cosine Similarity can better capture the similarity than Simple Matching Coefficient (SMC).

Ans.) Given the points p3 = {1, 1, 1, 0, 0} & p4 = {1, 0, 0, 0, 1}

We know that

* SMC = No. of Matches / No. of Attributes.
* M00 indicates the number of 00 attribute values.
* M01 indicates the number of 01 attribute values.
* M10 indicates the number of 10 attribute values.
* M11 indicates the number of 11 attribute values.

M00 = 1, M01 = 1, M10  = 2, M11 = 1

Therefore, SMC = 1+1 / 1+1+2+1

= 2 /5

= 0.4.

We know that

* **‘.’** indicates vector dot product.
* **||p||** is the length of the vector p.

P3 = 1, 1, 1, 0, 0

P4 = 1, 0, 0, 0, 1

P3.p4 = 1(1) + 1(0) + 1(0) + 0(0) + 0(1)

= 1

||p3|| = 2 + 12 + 12 + 02 + 02 = = 1.732.

||p4|| = 2 + 02 + 02 + 02 + 12 = = 1.414.

Therefore, cos(p3,p4) = 1 / 1.732 \* 1.414

= 1 / 2.449048

= 0.4083.

Hence, this proves that the cosine similarity can better capture the similarity than Simple matching coefficient (SMC).

Q2(3) In the random sample and the stratified sample that you generated in Q2(1 & 2), are there any major differences regarding the number of instances for each class label? Why?

Ans.)

Text

Description automatically generated

* Sampling in general we know that it is used for both the preliminary investigation of the data and the final data analysis.
* Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.
* **Simple random sampling**: there is an equal probability of selecting any particular item.
* **Sampling without replacement**: as each item is selected, it is removed from the population.

As the above image shows, the number of random samples selected from that of the entire dataset is 10.

A black screen with white text

Description automatically generated with low confidence

The above image shows, the stratified sampling selected from the entire dataset of random samples is 10.

* **Stratified random sampling**: Split the data into several partitions, then draw random samples from each partition.
* No, there is no major difference between the two random samplings for its instances of the class labels.
* The simple random sampling selects the random samples from the dataset irrespective of the class labels. Whereas the stratified sampling splits the data into partitions and then drew the random partitions from each partition.