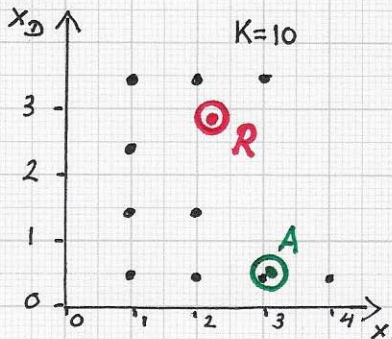
■ GENERATION OF SYNTHETIC DATA – Cont'd.

• Stochastic method for multi-dim. case:



D-dimensional feature space.

Given: K=10 original feature points; R = 0.25.

Tuples **A** and **R** are generated randomly.

A and **R** represent the centers of 2 balls (=disks).

A's ball includes 1 of the K=10 points; **R's ball** does not include an original point.

Tuple **A** is accepted with probability 0.1 = 1/10 as a synthetic point; tuple **R** is rejected as its probability for acceptance is **0**.

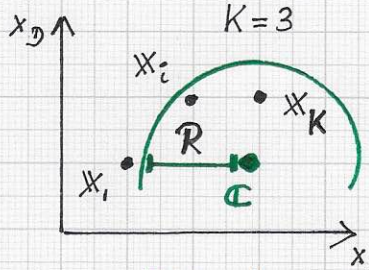
THE VALUE OF BALL RADIUS R IS MOST IMPORTANT FOR THE "SHAPE" OF THE RESULTING SYNTHETIC DATA.

It is possible to use a given discrete, finite sample set of feature points / vectors $(x_1, \ldots, x_D)$ distributed in D-dimensional feature space "directly" for the creation of "similar" feature point distributions. Such a direct method does not require one to define a data parametrization, a grid, a (local) polynomial approximation of the distribution function $p(x_1, \ldots, x_D)$ etc. A stochastic method for synthetic feature point distributions can operate directly on the given discrete data – generating synthetic distributions that are also pure point set data, i.e., scattered data without a grid.

A random number generator is used to produce a tuple $(x_1, \ldots, x_D)$, where the random values of the components x_i lie in the range intervals of the respective components. Two parameters are crucial: (i) the number K of original sample points and (ii) the radius R of a D-dimensional ball having a random tuple $(x_1, \ldots, x_D)$ as its center.

■ GENERATION OF SYNTHETIC DATA – Cont'd.

• Stochastic method:



conditional acceptance of random value $\mathbb{C}$ based on associated ball and number of original samples in ball;

"very small" values of R ensuring "very close" proximity between an accepted point $\mathbb{C}$ and the original samples

These are the necessary computations:

→ randomly compute a tuple/point $\mathbb{C}$;

→ determine the number k of original feature points $\mathbb{X}_i$ lying inside the disk / ball with radius R with center $\mathbb{C}$;

$\mathbb{X}_i$ inside ball $\Leftrightarrow \| \mathbb{X}_i - \mathbb{C} \| \leq R$;

→ randomly compute a value $p \in [0, 1]$ for the random point $\mathbb{C}$;

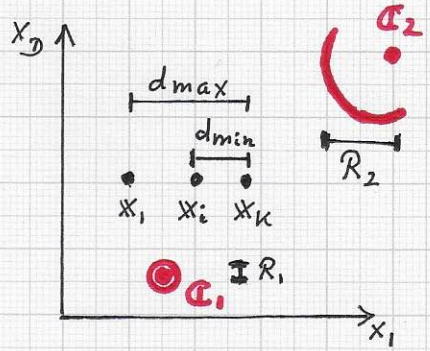
→ case 1: $k = 0 \Rightarrow$ reject ;

case 2: $p_t = k / K$;

$p \leq p_t \Rightarrow$ accept ;

/* threshold value p_t determines */

/* whether $\mathbb{C}$ is accepted or rejected. */



minimal and maximal distances – considering all pairwise distances between given samples $\mathbb{X}_i$ and $\mathbb{X}_j$, $i, j = 1 \ldots K$:

$$d_{min} = \min \{ \| \mathbb{X}_j - \mathbb{X}_i \| \}_{i,j=1}^{K}$$

$$d_{max} = \max \{ \| \mathbb{X}_j - \mathbb{X}_i \| \}_{i,j=1}^{K}$$

$\mathbb{C}_1, \mathbb{C}_2$: randomly generated synthetic feature points with associated ball radii R_1 and R_2; $R_1 < \frac{1}{2} d_{min}$; $R_2 = \frac{1}{2} d_{max}$

→ terminate the process when K synthetic values have been accepted, assuming that synthetic distributions should have the cardinality of the given sample distribution.

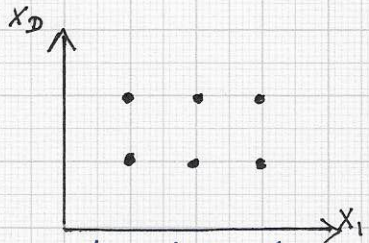
→ We assume that $0 < R < \infty$. Several "special cases" can be considered for R-values.

(i) $0 < R < \frac{1}{2} d_{min}$: [see figure] Only two cases are possible; a ball with radius R contains either 0 or 1 original point.

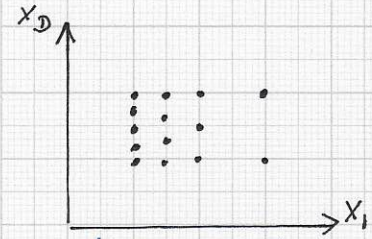
(ii) $\frac{1}{2} d_{min} < R \leq \frac{1}{2} d_{max}$: A ball with radius R can contain any number of original points, 0, 1, 2, ..., K. (A randomly generated point $\mathbb{C}$ with associated radius $R = \frac{1}{2} d_{max}$ can have an associated ball containing all points $\mathbb{X}_i$, $i = 1 \ldots K$)

GENERATION OF SYNTHETIC DATA - Cont'd.

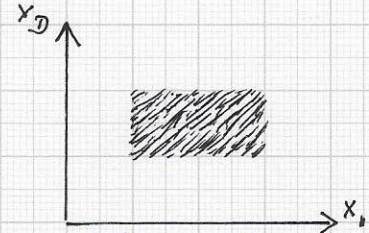
• Stochastic method:



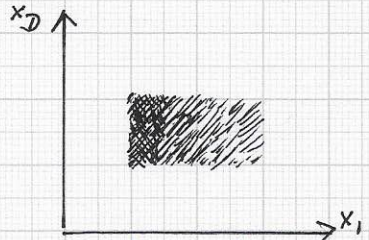
discrete samples, uniform distribution



discrete samples, non-uniform distribution



continuum domain, uniform distribution



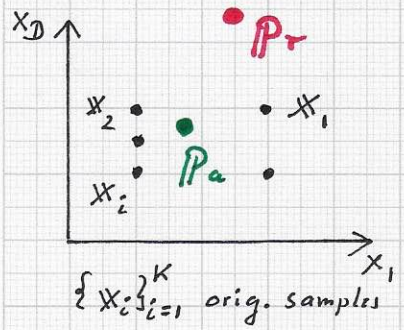
continuum domain, non-uniform distribution

Feature point/vector distributions defined in D-dim. feature space: definitions via finite sets of discrete samples or analytical functions over continuous domain regions.

In principle, one can define a distribution of feature points/vectors, obtained from original sample values, via a (high-resolution) finite set of points in D-dimensional feature space or an analytical definition, representing the distribution over a (bounded) continuum region of D-dimensional space. When employing a stochastic approach for generating additional synthetic feature point sets — finite and consisting of discrete synthetic samples — the generating process must be able to operate on a given distribution defined in either discrete or analytical form. Further, the generating process must produce synthetic samples with an associated distribution that must be "close" to the given distribution and, when desired, "converges" to the given distribution. Therefore, a meaningful measure for distribution similarity, computable with acceptable time complexity, is necessary for the generation of synthetic feature point distributions — in support of data augmentation.

■ GENERATION OF SYNTHETIC DATA - Cont'd.

• **Stochastic method:**

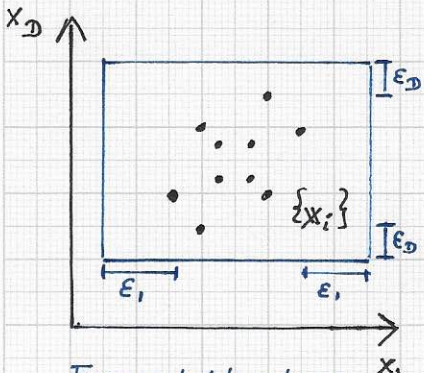


$\{x_i\}_{i=1}^{K}$ orig. samples

$\mathbb{P}_r$ rejected pnt.

$\mathbb{P}_a$ acceptable pnt. (subject to probabil.)

Efficient data structures and algorithms are needed to compute $\mathbb{P}_r$, $\mathbb{P}_a$.



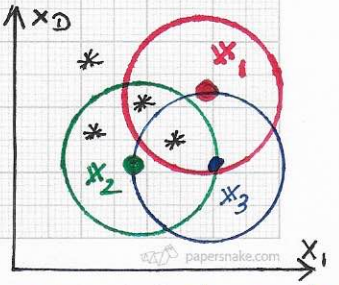
Expanded bounding box. Random point computations should be restricted to a simple finite region in D-dim. feature space, e.g., an ε-expanded bounding box of the original feature pnts.

All computations to be done as part of the random generation process of synthetic feature point distribution should be as efficient as possible. Computational complexity is mainly influenced by the number of dimensions of feature space (D) and "resolution" of synthetic feature points (e.g., considering the number of 'bins' used for each dimension).

For example, consider $K = 2^{30}$ as the approximate limit for any feature point set's cardinality; one already reaches this limit for $D = 10$ and resolution $R = 8$: $R^D = (2^3)^{10} = 2^{30} = K$. Thus, it is imperative to keep D and R values low.

(Further - as discussed before - one might be able to save significant storage and computation costs by representing distributions via analytical, locally best polynomial approximations.)

All random number computations should be done only inside an enlarged bounding box (left figure), and "point-inside-ball tests" must be efficient (right fig.).

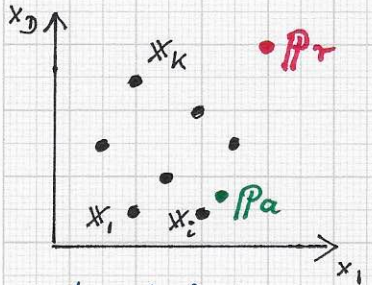


Original feature points x_i with associated ε-balls; reject/accept synthetic pnts '*'.

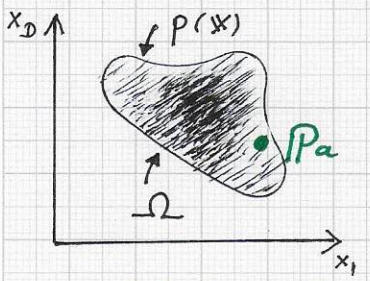
Stratovan

■ GENERATION OF SYNTHETIC DATA – Cont'd.

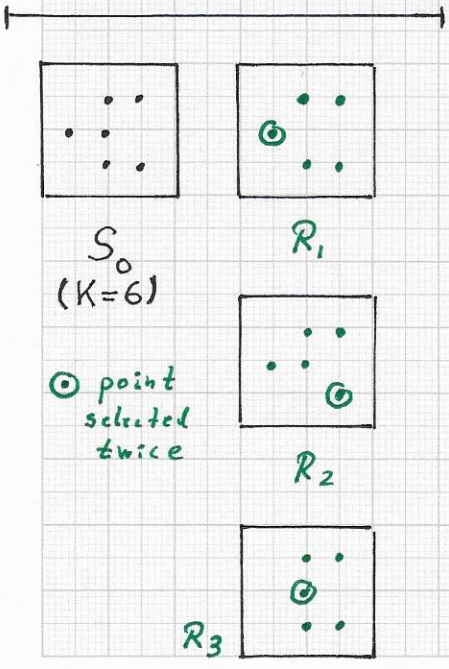
• Stochastic method – computations:



discrete feature points only



analytical distribution function $p(\mathbb{P})$



Bootstrapping and resampling: original sample S_0 and resamples R_i

One must consider the cost for generating synthetic feature point/vector data sets.

(i) Only a discrete sample of the distribution of feature values (of a specific class) is given in D-dimensional feature space. For the direct generation of a synthetic data set of cardinality K, one would randomly produce feature points $\mathbb{P}$ inside some feasible region. When a point $\mathbb{P}$ is not inside any ε-ball of the given K points $\{\mathbb{P}_i\}$, then it will be rejected ($\mathbb{P}_r$); when $\mathbb{P}$ lies inside k ε-balls, then it will be accepted ($\mathbb{P}_a$) WITH A PROBABILITY k/K. EXPENSIVE!

(ii) An analytical function $p(\mathbb{P})$ of the distribution of feature values over a domain Ω is given (e.g., a piecewise polynomial approximation). To generate a discrete synthetic sample randomly via $p(\mathbb{P})$, one would produce a point $\mathbb{P}$ randomly inside Ω and accept it ($\mathbb{P}_a$) when a randomly produced probability value $\hat{p}$ satisfies $0 < \hat{p} \leq p(\mathbb{P}) < p_{max}$. EXPENSIVE!

(iii) When adapting random resampling for our purpose, one can produce resample by randomly selecting K points from the given sample S_0. Points in the resamples can contain multiplicities > 1. (S_0 is de facto viewed as the 'true distribution.') ≈ BH