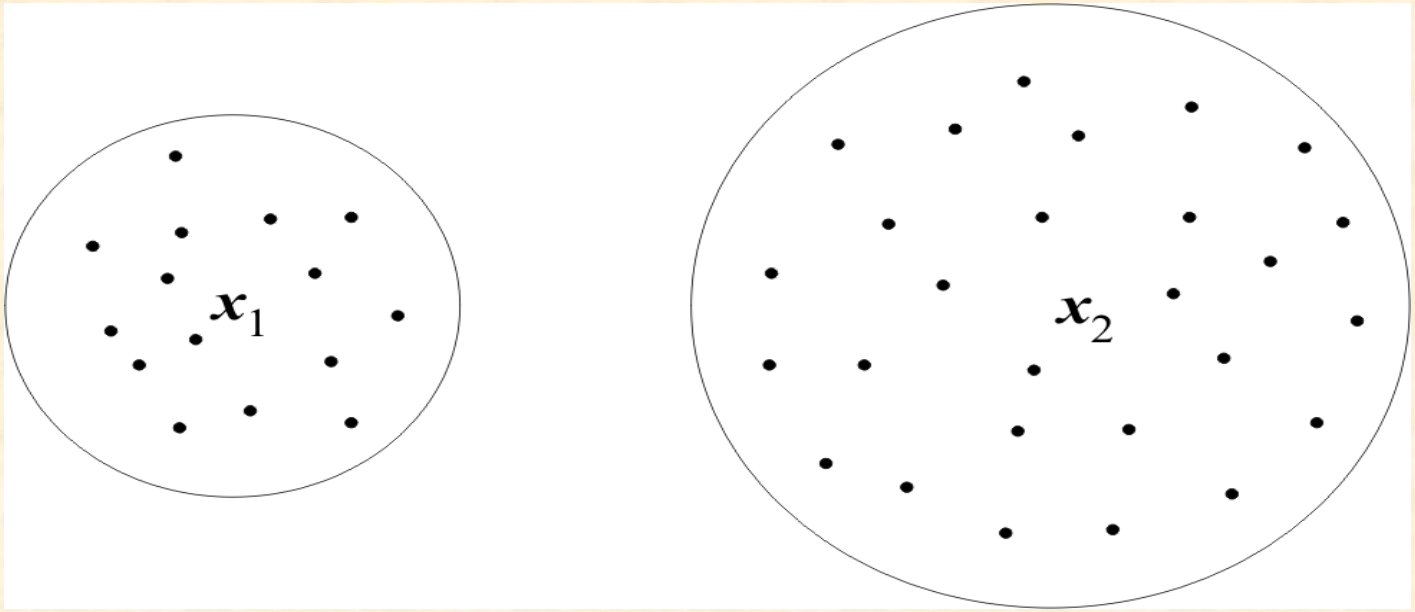
❖ Εκτίμηση πυκνότητας με βάση τους K-εγγύτερους γείτονες (K Nearest Neighbor Density Estimation)

- Στα παράθυρα Parzen:
  - Ο όγκος είναι σταθερός
  - Ο αριθμός των σημείων στον όγκο μεταβάλλεται
- Τώρα:
  - Κρατάμε τον αριθμό των σημείων σταθερό $k_N = k$
  - Επιτρέπουμε στον όγκο να μεταβάλλεται
  - $\hat{p}(\underline{x}) = \dfrac{k}{NV(\underline{x})}$



Κανόνας του Bayes:
($\theta = P(\omega_2)/P(\omega_1)$)

$$\frac{\frac{k}{N_1V_1}}{\frac{k}{N_2V_2}} = \frac{N_2V_2}{N_1V_1}(><)\theta$$

❖ ΚΑΤΑΡΑ ΤΗΣ ΔΙΑΣΤΑΤΙΚΌΤΗΤΑΣ (CURSE OF DIMENSIONALITY)

- Σε όλες τις μεθόδους, μέχρι τώρα, είδαμε ότι όσο μεγαλύτερος είναι ο αριθμός των σημείων, N, τόσο καλύτερη είναι η προκύπτουσα εκτίμηση.

- Αν στο μονοδιάστατο χώρο ένα διάστημα, που περιέχει N σημεία, is αρκετό (για καλή εκτίμηση), στον δισδιάστατο χώρο το αντίστοιχο τετράγωνο θα απαιτεί N^2 και στον ℓ-διάστατο χώρο ο ℓ-διάστατος κύβος θα απαιτεί N^ℓ σημεία.

- Η εκθετικά αύξηση του αριθμού των αναγκαίων σημείων είναι γνωστή ως κατάρα της διαστατικότητας (curse of dimensionality). Πρόκειται για σημαντικό πρόβλημα που αντιμετωπίζει κανείς σε χώρους υψηλής διάστασης.

❖ ΑΠΛΟΙΚΟΣ ΤΑΞΙΝΟΜΗΤΗΣ BAYES (NAIVE – BAYES CLASSIFIER)

- Έστω $\underline{x} \in \Re^{\ell}$ και ότι στόχος είναι η εκτίμηση της $p(\underline{x} \mid \omega_i)$ $i = 1, 2, \ldots, M$. Για μία "καλή" εκτίμηση της pdf χρειάζονται, ας πούμε, N^{ℓ} σημεία.

- Έστω ότι $x_1, x_2, \ldots, x_{\ell}$ είναι αμοιβαίως ανεξάρτητα. Τότε:

$$p(\underline{x} \mid \omega_i) = \prod_{j=1}^{\ell} p(x_j \mid \omega_i)$$

- Σ' αυτή την περίπτωση, κάποιος θα χρειαζόταν, περίπου, N σημεία για κάθε pdf. Έτσι, $O(N \cdot \ell)$ σημεία θα αρκούσαν.

- Ο απλοϊκός ταξινομητής Bayes δουλεύει καλά ακόμα και σε περιπτώσεις όπου παραβιάζεται η υπόθεση της ανεξαρτησίας

❖ Ο κανόνας του εγγύτερου γείτονα (The Nearest Neighbor Rule)

- **Για δεδομένο** $\underline{x}$
- Από δεδομένο σύνολο N διανυσμάτων εκπαίδευσης, προσδιόρισε τα k πλησιέστερα στο $\underline{x}$

- Από αυτά τα k προσδιόρισε τα k_i που ανήκουν στην κλάση ω_i

$$\text{Καταχώρησε } \underline{x} \rightarrow \omega_i : \quad k_i > k_j \quad \forall i \neq j$$

- Η απλούστερη περίπτωση

$k=1$!!!

- Για μεγάλα N παρουσιάζει καλή συμπεριφορά. Αποδεικνύεται ότι: Αν P_B είναι η βέλτιστη πιθανότητα λάθους κατά Bayes, τότε:

$$P_B \leq P_{NN} \leq P_B\left(2 - \frac{M}{M-1} P_B\right) \leq 2P_B$$

- $P_B \leq P_{kNN} \leq P_B + \sqrt{\frac{2P_{NN}}{k}}$ Για δύο κλάσεις

- $k \to \infty,\ P_{kNN} \to P_B$

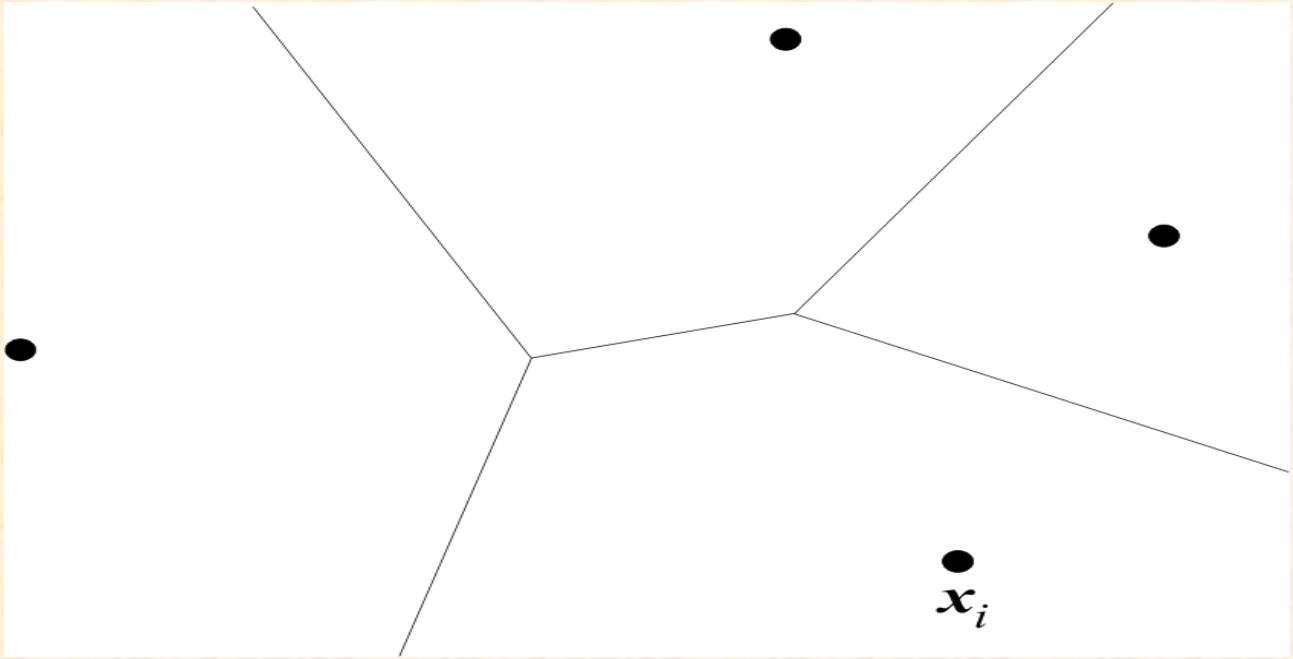
- Για μικρό P_B:

$$P_{NN} \cong 2P_B$$

$$P_{3NN} \cong P_B + 3(P_B)^2$$

Πρόβλημα: Μεγάλο υπολογιστικό κόστος ($O(kN)$ ανά δείγμα)

❖ Ψηφοθέτηση Voronoi (Voronoi tesselation) (1-NN)



$$R_i = \left\{ \underline{x} : d(\underline{x}, \underline{x}_i) < d(\underline{x}, \underline{x}_j)\, i \neq j \right\}$$