K Mean Algorithm

Initialize $\underline{\mu_1, \mu_2, \ldots, \mu_K}$ in some arbitrary fashion

Step 1: Given $\underline{\mu_1}$ to $\underline{\mu_K}$, define associated $S_1, S_2, \ldots, S_K$ via problem #2

Step 2: Given $S_1, S_2, \ldots, S_K$, use problem # to find updated $\underline{\mu_1, \mu_2, \ldots, \mu_K}$

Repeat Step 1 and Step 2 until convergence

Density Estimation

$D = \{\underline{x_1, x_2, \ldots, x_N}\}$ iid (independent identically distributed)

$P(\underline{x}) \longrightarrow$ Unknown

Output: $\hat{P}(\underline{x})$ an approximation of $P(\underline{x})$

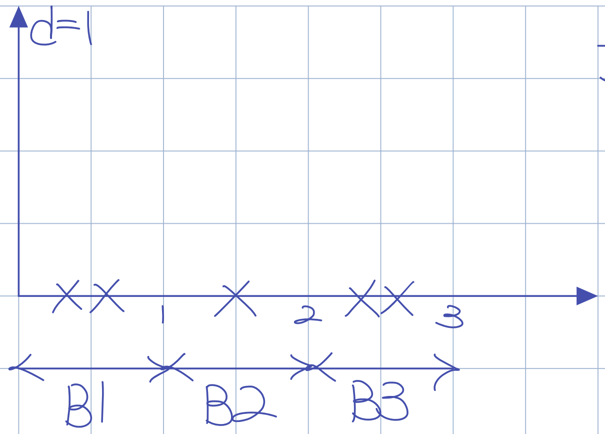
Histogram Method

Assume that $P(\underline{x})$ is none zero over a bounded set $\mathcal{X}$

$\underline{x} \in \mathbb{R}^d \qquad \mathcal{X} \in [0, 1]^d$

$$\begin{vmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{vmatrix} \qquad 0 \leq x_i \leq 1$$

Idea: Cover $\mathcal{X}$ using uniform hypercubes $B_1, B_2, \ldots, B_K$ each of volumn V



$$\mathcal{X} = [0, 3] \qquad k=3$$

$$B_1 = [0, 1)$$

$$B_2 = [1, 2)$$

$$B_3 = [2, 3]$$

N_i : # of points in B_i

N : Total # of points

$$P_X(x \in B_i) = \frac{N_i}{N}$$

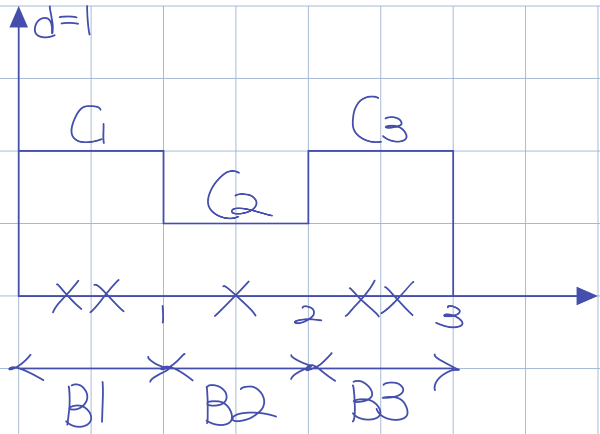
$$P_X(x \in B_2) = \frac{1}{5}$$

$$P_X(x \in B_3) = \frac{2}{5}$$

Our target is to compute $\hat{p}(\underline{x})$

Assume that for each $\underline{x} \in B_i$, $\hat{p}(\underline{x})$ is a constant function

$\hat{p}(\underline{x}) \longrightarrow$ pdf



$$P[\underline{x} \in B_1] = \frac{2}{5}$$
$$P[\underline{x} \in B_2] = \frac{1}{5}$$
$$P[\underline{x} \in B_3] = \frac{2}{5}$$

$$\int_{\underline{x} \in B_1} \hat{p}(\underline{x})\, d\underline{x} = P_x(\underline{x} \in B_1)$$
$$= \int_0^1 C_1\, dx$$
$$= \frac{2}{5}$$

$V = 1$ since $d = 1$

General Case

Let $\underline{x} \in \mathbb{R}^d$

Cover $\mathcal{X}$ using uniform hypercube $B1, B2, \ldots, B_K$, each of volume $= V$

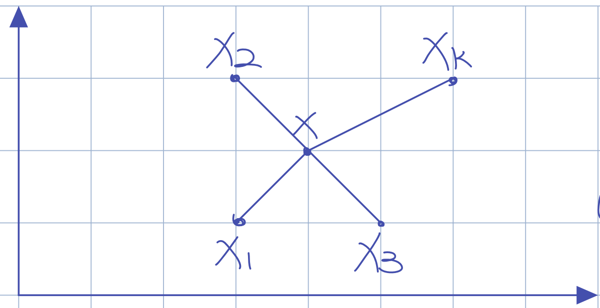
N_i = # of points in D that belong to B_i

$$P_x(x \in B_i) = \frac{N_i}{N} \qquad \hat{p}(\underline{x}) = \frac{N_i}{NV}, \quad \underline{x} \in B_i$$

Nearest Neighbour Estimation

$D = \{\underline{x_1, x_2, \ldots, x_N}\}, \ \underline{x_i} \in \mathbb{R}^d$

Estimate $\hat{p}(\underline{x}), \ \underline{x} \in \mathbb{R}^d$



Given $\underline{x}$, compute k nearest neighbors

Call them $\underline{x[1]}, \underline{x[2]}, \ldots, \underline{x[k]}$

$k=2$, $\underline{x[1]} = x_2$

$\underline{x[2]} = x_1$

Compute
$$d_1 = d(x, x[1])$$
$$d_2 = d(x, x[2])$$
$$\vdots$$
$$d_k = d(x, x[k])$$

$V_k(\underline{x})$ = Volume of a sphere in $\mathbb{R}^d$ with radius d_k

$$\hat{p}(\underline{x}) = \frac{c}{V_k(x)}$$

c is a constant from
$$\int \hat{p}(\underline{x})\,dx = 1$$
don't compute c