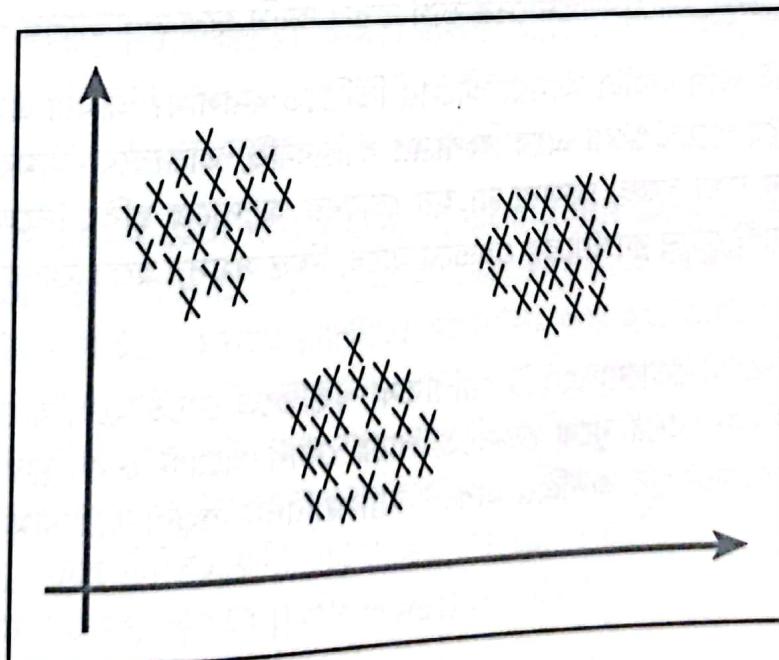


অধ্যায় ৭ : কে-মিনস ক্লাস্টারিং (K-Means Clustering)

কে-মিনস ক্লাস্টারিং অ্যালগরিদমটি প্রথম ব্যবহার করেন ম্যাককুইন (James MacQueen) নামের একজন বিজ্ঞানী 1967 সালে, যদিও মূল ধারণাটি ছিল হ্যুগো স্টেইনহাউস (Hugo Steinhaus, 1887 - 1972) নামে একজন বিজ্ঞানীর (1957 সাল)। ম্যাককুইন ছিলেন ক্যালিফোর্নিয়া বিশ্ববিদ্যালয়ের পরিসংখ্যান বিভাগের একজন অধ্যাপক।

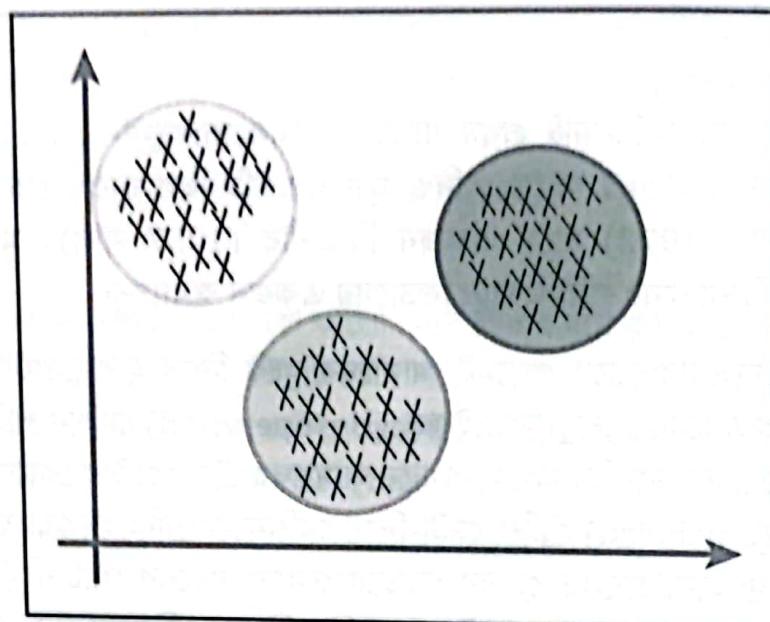
মূল অ্যালগরিদমে যাওয়ার আগে প্রথমেই আমাদের বুঝে নিতে হবে ক্লাস্টারিং আসলে কী? **ক্লাস্টারিং** হচ্ছে এক ধরনের আনসুপারভাইজড (Unsupervised) মেশিন লার্নিং পদ্ধতি, যেটি আমরা প্রথমেই আলোচনা করেছি। অর্থাৎ, এখানে আমাদের **ট্রেনিং ডেটার কোনো লেবেল থাকবে না। আনলেবেলড (unlabeled) ট্রেনিং ডেটা দিয়ে মেশিনকে ট্রেনিং দেওয়া হবে এবং মেশিন চেষ্টা করবে সেইসব আনলেবেলড ডেটার ভেতরে কোনো ধরনের প্যাটার্ন খুঁজে বের করতে** (unlabeled ডেটা সম্পর্কে আমরা প্রথমেই ধারণা দিয়েছিলাম পরিচ্ছেদ ২.২-এ)। **তারপরে যে সব ডেটা পয়েন্ট একই প্যাটার্নের, তাদেরকে একটি গুচ্ছ/ক্লাস্টারে নিয়ে নেবে এই অ্যালগরিদম।** এই হচ্ছে সাধারণভাবে বলতে গেলে ক্লাস্টারিং। নিচের ছবিটি দেখি :



ছবি 7.1

ঘবিতে, প্রতিটি ক্রস চিহ্ন হচ্ছে একটি ডেটা পয়েন্ট। সব পয়েন্টই যেহেতু এখানে সাদাকালো, তাহলে কিন্তু সবার চেহারা একই হয়ে গেল। এদের এখন আলাদা করা কিন্তু মুশকিল। এখন যদি বলা হয়, এই সাদাকালো পয়েন্টগুলোকে ডেটার বিন্যাস থেকে কোনো প্যাটার্ন বের করে তার

ওপরে ভিত্তি করে কতগুলো গুচ্ছ/ক্লাস্টারে ভাগ করতে, যাতে একই ধরনের প্যাটার্নের সব ডেটা একসঙ্গে থাকে, তখন আপনারা কী করতেন?



ছবি 7.2

আমার আন্দাজ বলছে, আপনারা অনেকটা এভাবে সবগুলো ডেটা তিনটি গুচ্ছ/ক্লাস্টারে ভাগ করতেন, তাই না? ঠিকই আছে আপনাদের ভাগ করা। আমি হলেও এভাবেই করতাম।

এখন কথা হচ্ছে, এই ভাগ করাটা আমরা কীসের ভিত্তিতে করলাম? আমরা এই ভাগটা করে চোখের আন্দাজে। যে পয়েন্টগুলো একে অপরের কাছাকাছি, তাদেরকে আমরা একই ক্লাস্টারে নিয়ে নিয়েছি। এখন কথা হচ্ছে, আমরা না হয় চোখের আন্দাজে ছবির দিকে তাকিয়ে বুঝ পারছি যে কোন পয়েন্ট কোন ক্লাস্টারের ভেতরে যাবে, কিন্তু সমস্যা হচ্ছে কম্পিউটারকে সেটি করে বোঝাই?

সেজন্যই আমাদের এখন কোনো একটি গাণিতিক পদ্ধতিতে যেতে হবে, যাতে করে আমা
র কম্পিউটার হিসাবনিকাশ করে বুঝে ফেলতে পারে কোন পয়েন্ট কোন ক্লাস্টারে যাবে।
ক্লাস্টারিংয়ের জন্য আমরা খুব জনপ্রিয় একটি অ্যালগরিদম পড়ব, যার নাম হচ্ছে K-Means Clustering।

পরিচেদ ৭.১ : কে-মিনস ক্লাস্টারিংয়ের সংক্ষিপ্ত গাণিতিক বর্ণনা

বিস্তারিত বর্ণনায় যাওয়ার আগে, একটু সংক্ষেপে ব্যাখ্যা করে নিই যে কাজটি আসলে কীভাবে। আমাদের যেটি করতে হবে, সেটি হচ্ছে প্রথমে আমাদেরকে যে ট্রেনিং ডেটা দেওয়া থাকে

অধ্যায় ৭ : কে-মিনস ক্লাস্টারিং (K-Means Clustering)

সেখান থেকে ঠিক করে নিতে হবে যে আমরা কয়টি ক্লাস্টার নিয়ে কাজ করতে চাই। ধরা যাক, আমরা K-সংখ্যক ক্লাস্টার নিয়ে কাজ করতে চাই। প্রতিটি ক্লাস্টারের একটি কেন্দ্রবিন্দু বা সেন্টার পয়েন্ট থাকবে, যেটাকে আমরা বলব সেন্ট্রয়েড (centroid)। এই সেন্ট্রয়েডের মান হবে ওই ক্লাস্টারে যতগুলো পয়েন্ট আছে সবগুলোর গড় মান।

আমরা শুরুতে প্রতিটি সেন্ট্রয়েডকে র্যানডমলি একটি করে মান দিয়ে দেব। এরপর, আমাদের ট্রেনিং ডেটা থেকে আমরা একটি করে ডেটা পয়েন্ট নেব এবং K-সংখ্যক সেন্ট্রয়েডের প্রত্যেকটি থেকে সেই ডেটা পয়েন্টের দূরত্ব মাপব। যেই সেন্ট্রয়েড সবচেয়ে কাছে থাকবে, ওই ডেটা থেকে সেই সেন্ট্রয়েডের জন্য যে ক্লাস্টার আছে তাতে অ্যাসাইন করব। এভাবে আমরা একটি পয়েন্টকে সেই সেন্ট্রয়েডের জন্য যে ক্লাস্টারে আছে তাতে অ্যাসাইন করব। করে ডেটা পয়েন্ট নেব এবং তাকে কোনো-না-কোনো ক্লাস্টারে অ্যাসাইন করব।

সেই সঙ্গে আরো একটি কাজ চলবে, সেটি হচ্ছে, প্রতিবার ক্লাস্টারে একটি করে পয়েন্ট যুক্ত হওয়ার পর, আবার নতুন করে ওই ক্লাস্টারে থাকে সবগুলো পয়েন্টের গড় নিতে হবে। এই নতুন গড়-ই হবে তখন ওই ক্লাস্টারের জন্য নতুন সেন্ট্রয়েড, যেটি আমাদের প্রতিবার হিসাব করতে হবে।

অর্থাৎ, আমাদের কাজ হবে দুটি – ক্লাস্টার অ্যাসাইনমেন্ট এবং সেন্ট্রয়েড আপডেট।

ধরা যাক, আমরা m-সংখ্যক ট্রেনিং ডেটার প্রতিটি পয়েন্ট একটি করে ক্লাস্টারে অ্যাসাইন করব। আমাদের K-সংখ্যক ক্লাস্টার আছে। আমরা একটি অ্যারের কথা চিন্তা করি, যার নাম C এবং সাইজ হচ্ছে m। এই অ্যারেতে m-সংখ্যক উপাদান থাকবে। অ্যারের i-তম উপাদান হবে i-তম ট্রেনিং ডেটাকে 1 থেকে K-এর মধ্যে কত নম্বর ক্লাস্টারে অ্যাসাইন করা হয়েছে সেই সংখ্যাটি।

ধরা যাক, আমাদের প্রথম ডেটা পয়েন্ট হলো a , যাকে 2 নম্বর ক্লাস্টারে অ্যাসাইন করা হয়েছে। একইভাবে সুতরাং, $C^{(1)}$ হবে 2 যেখানে $C^{(1)}$ দিয়ে প্রথম ডেটা পয়েন্টকে নির্দেশ করা হচ্ছে। একইভাবে যদি দ্বিতীয় ডেটা পয়েন্টকে 4 নম্বর ক্লাস্টারে অ্যাসাইন করা হয়, তাহলে $C^{(2)}$ হবে 4। কোনো ডেটা পয়েন্টকে K-সংখ্যক ক্লাস্টারের মধ্যে সেই ক্লাস্টারকেই অ্যাসাইন করা হবে, যে ক্লাস্টারের সেন্ট্রয়েড (ওই ক্লাস্টারে অ্যাসাইন করা সব পয়েন্টের মানের গড়মান) থেকে ওই ডেটা পয়েন্টের দূরত্ব অন্য সব ক্লাস্টারের সেন্ট্রয়েডের তুলনায় সবচেয়ে কম হবে। আমরা যদি k-তম ক্লাস্টারের সেন্ট্রয়েডকে μ_k দিয়ে চিহ্নিত করি, তাহলে i-তম ট্রেনিং ডেটা $x^{(i)}$ এবং k-তম ক্লাস্টারের সেন্ট্রয়েডকে μ_k দিয়ে চিহ্নিত করা হবে। $\|x^{(i)} - \mu_k\|$ যেখানে $1 \leq i \leq m$ ।

সেন্ট্রয়েড μ_k -এর মধ্যে দূরত্ব হবে $\|x^{(i)} - \mu_k\|$ যেখানে $1 \leq i \leq m$ । এই গেল আমাদের প্রথম ধাপ – ক্লাস্টার অ্যাসাইনমেন্ট। এরই সঙ্গে আরেকটি ধাপ আছে, সেটি হচ্ছে সেন্ট্রয়েডের মান আপডেট করা। প্রতিবার একটি করে নতুন ডেটা পয়েন্ট $x^{(i)}$ কোনো একটি ক্লাস্টার k-তে অ্যাসাইন করার পর, সেই ক্লাস্টারের সেন্ট্রয়েড আপডেট করতে হবে। নতুন

সেন্ট্রয়েডের মান হবে ওই ক্লাস্টারে অ্যাসাইন করা সব ট্রেনিং ডেটার মানের (সদৃ অ্যাসাইন করা নতুন ডেটা পয়েন্টসহ) গড়।

আমরা যদি এখন এর জন্য কস্ট ফাংশন লিখতে যাই, সেটি তাহলে হবে এরকম –

$$J(\underbrace{C^{(1)}, C^{(2)}, C^{(3)}, \dots C^{(m)}}_{m \text{ number of training examples}}, \underbrace{\mu_1, \mu_2, \mu_3, \dots \mu_K}_{k \text{ number of cluster centroids}}) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_k\|^2$$

এবং অবজেকটিভ ফাংশন হবে,

$$\underset{(c^{(i)}, \mu)}{\text{minimize}} J(C^{(1)}, C^{(2)}, C^{(3)}, \dots C^{(m)}, \mu_1, \mu_2, \mu_3, \dots \mu_K)$$

পরবর্তী পরিচ্ছেদে উদাহরণ দিয়ে কে-মিনস ক্লাস্টারিং আরে বিস্তারিত আলোচনা করা হয়েছে। উদাহরণটি দেখলেই অনেকখানিই পরিক্ষার ধারণা পাওয়া যাবে।

পরিচ্ছেদ ৭.২ : উদাহরণ

যাক, অনেক গণিত হলো, এখন চলুন একটি ছোটো উদাহরণ দিয়ে দেখি কীভাবে এই কে-মিনস ক্লাস্টারিং কাজ করে।

X	Y
2	4
2	3
5	2
6	2
5	2.5
2.5	3.5

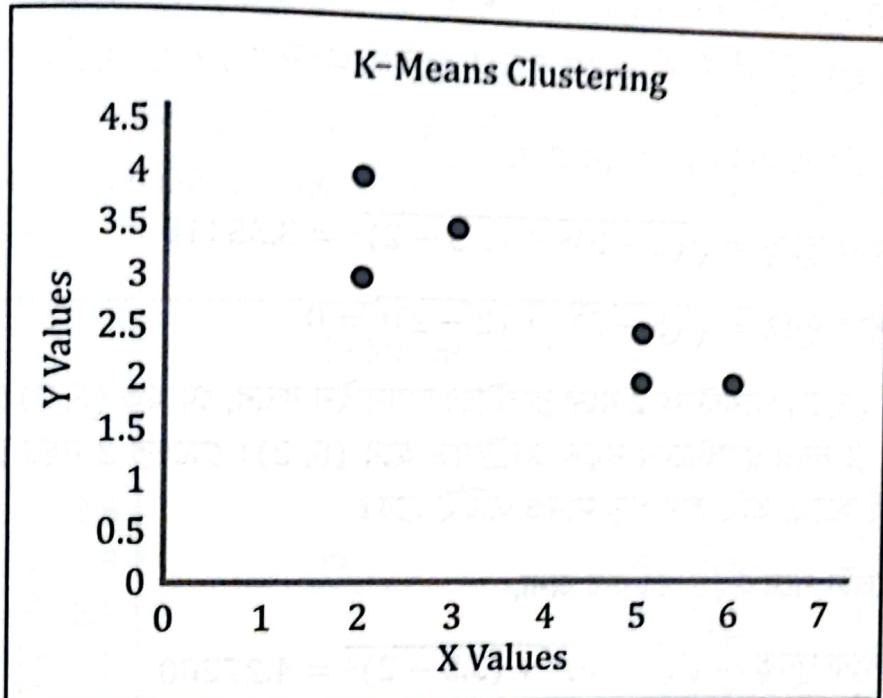
টেবিল 7.2.1

টেবিল 7.2.1-এর ডেটা একটি গ্রাফে প্লট করলে অনেকটা গ্রাফ 7.2.1-এর মতো দেখাবে। এখন দেখুন, আমরা যদি এই ডেটা সেটের ওপরে কে-মিনস ক্লাস্টারিং অ্যালগরিদম প্রয়োগ করতে যাই, তাহলে প্রথমেই আমাদের ঠিক করতে হবে আমরা কয়টি ক্লাস্টার নেব, অর্থাৎ K-এর মান কত হবে? আমরা যে-কোনো মান দিয়েই শুরু করতে পারি। ধরে নিই K = 2 (চিত্র দেখেই বোকা

অধ্যায় ৭ : কে-মিনস ক্লাস্টারিং (K-Means Clustering)

যাচ্ছে ক্লাস্টার ২টি হবে, তাই দুটি ক্লাস্টার দিয়েই দেখাচ্ছি। তবে K-এর মান কত হলে সবচেয়ে ভালো হবে সেটি বের করার একটি উপায় আছে, যেটি পরে আলোচনা করব।

যদি $K = 2$ নিই, তাহলে আমাদের দুটি ক্লাস্টারের জন্য দুটি সেন্ট্রয়েড হবে μ_1 ও μ_2 । এদের মান হিসেবে শুরুতে আমরা যে-কোনো আনুমানিক মান ধরে নিতে পারি। তবে ভালো বৃক্ষ হচ্ছে আমাদের ট্রেনিং ডেটা পয়েন্টগুলো মধ্যে থেকে র্যানডমলি যে-কোনো দুটি পয়েন্টকে সেন্ট্রয়েড হিসেবে ধরে নেওয়া।



আফ 7.2.1

তাই, হিসাবের সুবিধার্থে আমরা ধরে নিই যে আমাদের $\mu_1 = (2, 4)$ এবং $\mu_2 = (5, 2)$ । এখন, প্রতিটি ডেটা পয়েন্ট থেকে আমরা এই দুটি সেন্ট্রয়েডের দূরত্ব বের করব এবং যেই সেন্ট্রয়েড থেকে দূরত্ব কম, আমরা ওই পয়েন্টকে সেই সেন্ট্রয়েডের ক্লাস্টারে অ্যাসাইন করব এবং সেই সেন্ট্রয়েডের মান আপডেট করব।

এখন, প্রথম ট্রেনিং ডেটা পয়েন্ট $(2, 4)$ -এর জন্য,

✓ μ_1 থেকে দূরত্ব = 0

✓ μ_2 থেকে দূরত্ব = $\sqrt{(5 - 2)^2 + (2 - 4)^2} = 3.60555$

সুতরাং আমরা $(2, 4)$ পয়েন্টকে 1 নম্বর ক্লাস্টারে অ্যাসাইন করব এবং নতুন সেন্ট্রয়েড হবে $(2, 4)$ । যেহেতু 1 নম্বর ক্লাস্টারে মাত্র একটি পয়েন্টই আছে, তাই তার গড় মানও এটিই হবে।

এখন, তৃতীয় ট্রেনিং ডেটা পয়েন্ট (2, 3)-এর জন্য,

$$\checkmark \mu_1 \text{ থেকে দূরত্ব} = \sqrt{(2-2)^2 + (4-3)^2} = 1$$

$$\checkmark \mu_2 \text{ থেকে দূরত্ব} = \sqrt{(5-2)^2 + (2-3)^2} = 3.16$$

সুতরাং আমরা (2, 3) পয়েন্টকেও 1 নম্বর ক্লাস্টারে অ্যাসাইন করব, যেহেতু (2, 3) থেকে μ_1 -এর দূরত্ব সর্বনিম্ন। সুতরাং 1 নম্বর ক্লাস্টারে পয়েন্ট আছে এখন দুটি: (2, 4) ও (2, 3)। সুতরাং নতুন সেন্ট্রয়েড হবে, $(\frac{2+2}{2}, \frac{4+3}{2}) = (2, 3.5)$ । সুতরাং এখন থেকে $\mu_1 = (2, 3.5)$ ।

এরপর তৃতীয় ডেটা পয়েন্ট (5, 2) এর জন্য,

$$\checkmark \mu_1 \text{ থেকে দূরত্ব} = \sqrt{(2-5)^2 + (3.5-2)^2} = 3.35410$$

$$\checkmark \mu_2 \text{ থেকে দূরত্ব} = \sqrt{(5-5)^2 + (2-2)^2} = 0$$

সুতরাং আমরা (5, 2) পয়েন্টকে 2 নম্বর ক্লাস্টারে অ্যাসাইন করব, যেহেতু (5, 2) থেকে μ_2 -এর দূরত্ব সর্বনিম্ন। 2 নম্বর ক্লাস্টারের নতুন সেন্ট্রয়েড হবে (5, 2)। যেহেতু 2 নম্বর ক্লাস্টারে মাত্র একটি পয়েন্টই আছে, তাই তার গড় মানও এটিই হবে।

এরপর চতুর্থ ডেটা পয়েন্ট (6, 2)-এর জন্য,

$$\checkmark \mu_1 \text{ থেকে দূরত্ব} = \sqrt{(2-6)^2 + (3.5-2)^2} = 4.27200$$

$$\checkmark \mu_2 \text{ থেকে দূরত্ব} = \sqrt{(5-6)^2 + (2-2)^2} = 1$$

সুতরাং আমরা (6, 2) পয়েন্টকেও 2 নম্বর ক্লাস্টারে অ্যাসাইন করব, যেহেতু (6, 2) থেকে μ_2 -এর দূরত্ব সর্বনিম্ন। সুতরাং, 2 নম্বর ক্লাস্টারে পয়েন্ট আছে এখন দুটি: (5, 2) ও (6, 2)। সুতরাং নতুন সেন্ট্রয়েড হবে, $(\frac{5+6}{2}, \frac{2+2}{2}) = (5.5, 2)$ । সুতরাং এখন থেকে $\mu_2 = (5.5, 2)$ ।

এরপর পঞ্চম ডেটা পয়েন্ট (5, 2.5)-এর জন্য,

$$\checkmark \mu_1 \text{ থেকে দূরত্ব} = \sqrt{(2-5)^2 + (3.5-2.5)^2} = 3.16228$$

$$\checkmark \mu_2 \text{ থেকে দূরত্ব} = \sqrt{(5.5-5)^2 + (2-2.5)^2} = 0.70710$$

সুতরাং আমরা (5, 2.5) পয়েন্টকেও 2 নম্বর ক্লাস্টারে অ্যাসাইন করব, যেহেতু (5, 2.5) থেকে μ_2 -এর দূরত্ব সর্বনিম্ন। সুতরাং, 2 নম্বর ক্লাস্টারে পয়েন্ট আছে এখন তিনটি: (5, 2), (6, 2) ও

অধ্যায় ৭ : কে-মিনস ক্লাস্টারিং (K-Means Clustering)

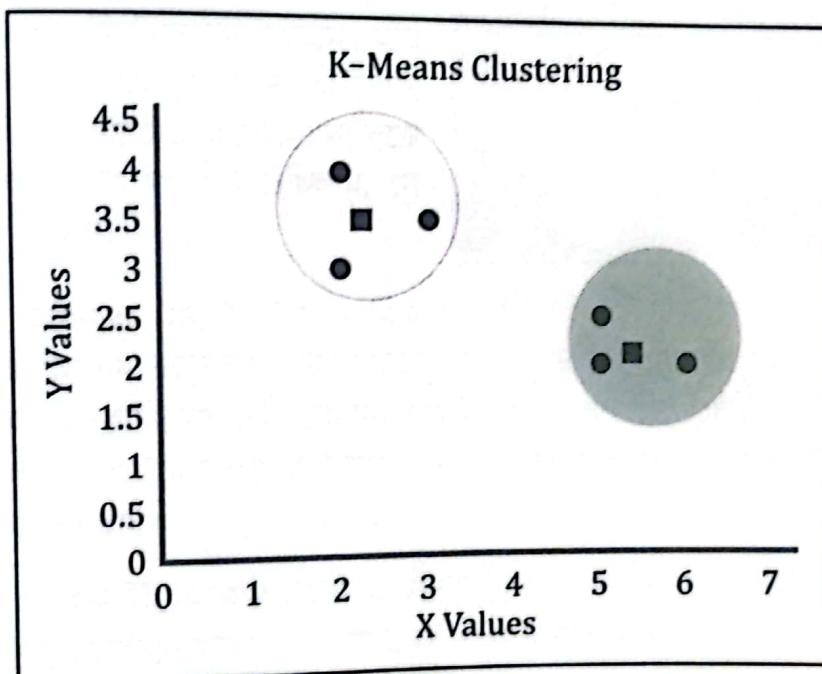
(5, 2.5)। সুতরাং, নতুন সেন্ট্রয়েড হবে, $\left(\frac{5+6+5}{3}, \frac{2+2+2.5}{3}\right) = (5.33, 2.17)$ । সুতরাং, এখন $\mu_2 = (5.33, 2.17)$ ।

সবশেষে, ঘষ্ট ডেটা পয়েন্ট (2.5, 3.5)-এর জন্য,

$$\checkmark \mu_1 \text{ থেকে দূরত্ব} = \sqrt{(2 - 2.5)^2 + (3.5 - 3.5)^2} = 0.5$$

$$\checkmark \mu_2 \text{ থেকে দূরত্ব} = \sqrt{(5.33 - 2.5)^2 + (2.17 - 3.5)^2} = 2.49799$$

সুতরাং (2.5, 3.5) পয়েন্টকে আমরা 1 নম্বর ক্লাস্টারে অ্যাসাইন করব, যেহেতু μ_1 থেকে (2.5, 3.5)-এর দূরত্ব সর্বনিম্ন। সুতরাং এখন, 1 নম্বর ক্লাস্টারে পয়েন্ট আছে তিনটি : (2, 4), (2, 3) ও (2.5, 3.5)। সুতরাং নতুন সেন্ট্রয়েড হবে, $\left(\frac{2+2+2.5}{3}, \frac{4+3+3.5}{3}\right) = (2.17, 3.5)$ । সুতরাং, এখন $\mu_1 = (2.17, 3.5)$ ।



গ্রাফ 7.2.2

এখন যদি ওপরের গ্রাফটি দেখি, তাহলে যে বর্গাকৃতির চিহ্ন দুটি দেখব, সে দুটি হচ্ছে আমাদের দুটি ক্লাস্টারের নতুন সেন্ট্রয়েড। বাঁ দিকের ক্লাস্টারটি হচ্ছে 1 নম্বর ক্লাস্টার এবং ডান দিকের ক্লাস্টারটি হচ্ছে 2 নম্বর ক্লাস্টার।

আগে বলেছিলাম যে, K-এর মান কীভাবে নির্ণয় করতে হয় সে সম্পর্কে বলব। ব্যাপারটি হচ্ছে, K-এর মান হিসেবে কোনটি ব্যবহার করলে কষ্ট সবচেয়ে কম হবে, সেটি সরাসরি বের করার জন্য কোনো গাণিতিক সূত্র নেই। তাই যেটি করতে হবে, K-এর মান 2, 3, 4, 5... ইত্যাদি ধরে দেখতে

হবে যে K-এর কোন মানের জন্য কল্পটি ফাঁশের কমতার্জ করতে, অর্থাৎ সর্বনিম্ন মান দিয়ে। এটি একটি ট্রিয়াল অ্যাস্ট এর (Trial and error) পদ্ধতি হবে, এই আর কি। পুর ভালো হয়, যখন আপনার K-সময়-কল্পটি-এর একটি গোফ তৈরি করে নিতে পারেন, তাহলে আপনাদের ব্যাপকভাৱে উপরাক্ষি করতে সুবিধা হবে।

অশা করি, সবাই দুবাতে পেরেছেন দীভাবে কে-বিনস ড্রাপ্টারিং অ্যাপ্লারিসমের সাহায্যে অমৃত ড্রাপ্টারিং করতে পারি।