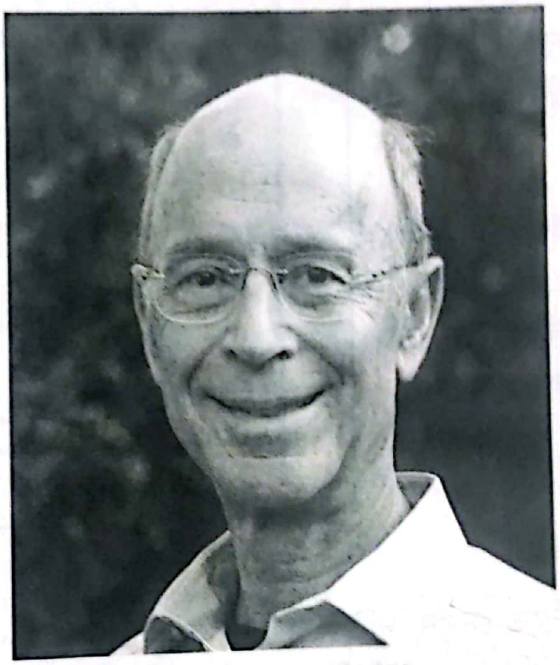
অধ্যায় ৬ : কে-নিয়ারেস্ট নেইবরস (K-Nearest Neighbors)

কে-নিয়ারেস্ট নেইবরস (K-Nearest Neighbors) বা KNN হচ্ছে একটি জনপ্রিয় এবং খুবই সহজ ক্লাসিফিকেশন অ্যালগরিদম। এটি প্রথম ব্যবহার করেন ফিক্স (Evelyn Fix, 1904 – 1965) ও হজেস (Joseph Lawson Hodges Jr., 1922 - 2000) নামের দুজন বিজ্ঞানী 1951 সালে। পরবর্তী সময়ে কভার (Thomas M. Cover, 1938 - 2012) ও হার্ট (Peter E. Hart, 1941) নামে দুজন বিজ্ঞানী এটি উন্নতকরণের কাজ করেন।

Tomas M. Cover (1938 - 2012)



Peter E. Hart (1941)

ছবি 6.1

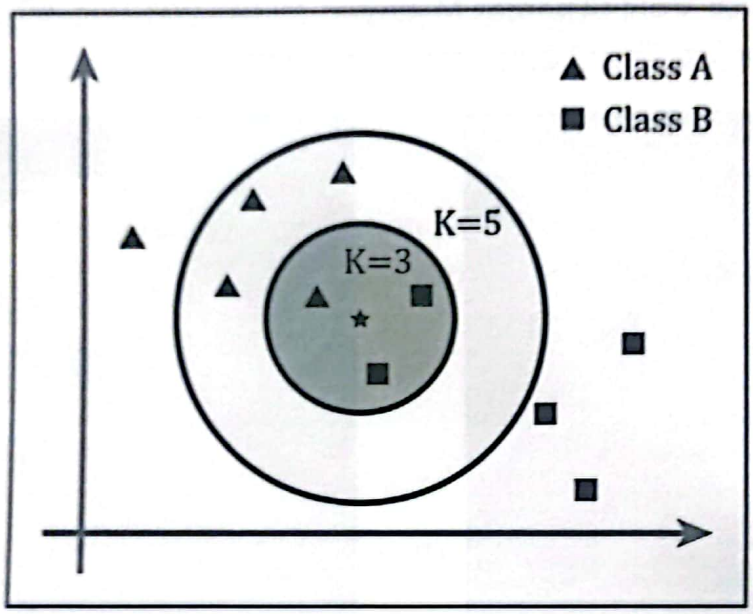
আমি যখন আমার ছাত্রছাত্রীদের মেশিন লার্নিং পড়াই, তারা সবচেয়ে সহজে বুঝতে পারে এই অ্যালগরিদমটি। এটি খুবই সহজ একটি অ্যালগরিদম, কিন্তু এর বাস্তবিক প্রয়োগ (যদি সঠিকভাবে করা যায়) হতে পারে খুবই ফলপ্রসূ। সোজা কথায় যদি বোঝাতে চাই, এটি একটি ভোট গণনা করার মতো অ্যালগরিদম।

ধরা যাক, আপনি কাচ্চি বিরিয়ানি খেতে চাইছেন হুট করে। আপনি আপনার তিনজন বন্ধুকে জিজ্ঞাসা করলেন কোথায় কাচ্চি বিরিয়ানি ভালো হবে? পুরান ঢাকার কোনো দোকানে, নাকি স্টার কাবাবের কাচ্চি? দুজন বন্ধু পুরান ঢাকার পক্ষে মতামত দিল, আর একজন দিল স্টারের পক্ষে। এখন তাহলে আপনি কোনটায় যাবেন? যেটায় বেশি মানুষ আপনাকে যেতে বলছে, অর্থাৎ পুরান

ঢাকাতে, সেটাতেই তো যাবেন, তাই নয় কি? যদি এটি বুঝে থাকেন, তাহলে এই অ্যালগরিদম আপনি বুঝে গেছেন অনেকখানিই।

পরিচ্ছেদ ৬.১ : KNN-এর সাধারণ ধারণা

নিচের ছবিটি দিয়ে বোঝা শুরু করি :



ছবি 6.1.1

ছবিতে দেখুন, দুটি ভিন্ন ভিন্ন আকৃতির ডেটা পয়েন্ট আছে, ত্রিভুজ ও বর্গ। ধরা যাক, এই দুই আকৃতি দিয়ে দুটি ভিন্ন ভিন্ন ক্লাস বোঝানো হয়েছে (মনে করি ত্রিভুজাকৃতি দিয়ে কুকুরের ছবি আর বর্গাকৃতি দিয়ে বিড়ালের ছবি)। ছবি অনুযায়ী, প্রতিটি ডেটার জন্য দুটি করে ফিচার নেওয়া হয়েছে, x_1 ও x_2। মাঝখানে একটি তারকা চিহ্ন আঁকা আছে, এটি হচ্ছে আমাদের অজানা ডেটা পয়েন্ট, অর্থাৎ আমরা জানি না, এটি কুকুর নাকি বিড়ালের ছবি। এখন, আমাদের যেটি করতে হবে, সেটি হচ্ছে, এই অজানা ডেটা পয়েন্টটিকে ক্লাসিফাই করতে হবে, অর্থাৎ বলতে হবে এটি কুকুরের ছবি নাকি বিড়ালের ছবি।

কীভাবে পুরো কাজটি করব, সেটি বিস্তারিত আলোচনার আগে আমি খুব সংক্ষেপে বলে দিই, কী করব আমরা। প্রথমে, আমরা K-এর একটি মান ধরে নেব। সাধারণত K-এর মান বেজোড় সংখ্যা ধরা হয় (1, 3, 5, 7... ইত্যাদি)। এর কারণ হচ্ছে, ধরুন যদি, $K = 4$ নিই, অর্থাৎ চারজন মানুষ ভোট দিচ্ছে। যদি দুজন ভোট দেয় যে নতুন ডেটা পয়েন্টটি কুকুরের ছবি, আর বাকি দুজন ভোট দেয় যে নতুন ডেটা পয়েন্টটি বিড়ালের ছবি, তাহলে সমান সমান ভোট হয়ে গেল না? তখন কে

আমরা মুশকিলে পড়ে যাব – নতুন পয়েন্টটিকে তখন আমরা কী হিসেবে মানব? কুকুর, নাকি বিড়াল? এই মুসিবত থেকে যাতে আমরা বিরত থাকতে পারি, তাই K-এর মান সাধারণত বেজোড় সংখ্যা নেওয়া হয়।

এরপরে আমাদের অজানা ডেটা পয়েন্টের আশপাশের সব জানা ডেটা পয়েন্ট থেকে সবচেয়ে কাছের K-সংখ্যক ডেটা পয়েন্ট আমরা বিবেচনা করব আমাদের পরবর্তী ধাপের জন্য। ওপরের চিত্রে দেখুন, $K = 3$-এর জন্য ভেতরের ছোটো বৃত্ত এবং $K = 5$-এর জন্য বাইরের বড়ো বৃত্তটি আঁকা হয়েছে। বড়ো বৃত্তের ভেতরে একটি বিন্দু অতিরিক্ত আছে, ওটি নিয়ে আপাতত মাথা ঘামানোর দরকার নেই।

এখন ছোটো বৃত্তের ভেতরে দেখুন তিনটি ডেটা পয়েন্ট রয়েছে, যেগুলোর মধ্যে দুটি বর্গ ও একটি ত্রিভুজ। সুতরাং বর্গের সংখ্যা বেশি, অর্থাৎ বিড়ালের সংখ্যা বেশি। এর মানে হচ্ছে $K = 3$ নিয়ে আমরা দেখতে পাই যে অজানা ডেটা পয়েন্টটির সঙ্গে বিড়ালের সামঞ্জস্য বেশি, তাই এর আকৃতি তারকা থেকে বর্গ করে দিয়ে একে বিড়াল বলে বিবেচনা করা হবে।

একইভাবে, বড়ো বৃত্তের ভেতরে দেখুন ছয়টি ডেটা পয়েন্ট রয়েছে, যার মধ্যে চারটি ত্রিভুজ এবং দুটি বর্গ। সুতরাং ত্রিভুজের সংখ্যা বেশি অর্থাৎ কুকুরের সংখ্যা বেশি। এর মানে হচ্ছে $K = 5$ নিয়ে আমরা দেখতে পাই যে অজানা ডেটা পয়েন্টটির সঙ্গে কুকুরের সামঞ্জস্য বেশি, তাই এটির আকৃতি তারকা থেকে ত্রিভুজ করে দিয়ে একে কুকুর বলে বিবেচনা করা হবে। এভাবেই মূলত KNN অ্যালগরিদম কাজ করে।

তাহলে এর মূল ধাপ চারটি –

- অজানা ডেটা পয়েন্ট থেকে বাকি সব ডেটা পয়েন্টের দূরত্ব বের করতে হবে।
- দূরত্বের মান অনুযায়ী ছোটো থেকে বড়ো আকার (বা, Ascending Order)-এ ডেটা পয়েন্টগুলো সর্ট (sort) করে নিতে হবে।
- সর্ট করা ডেটা পয়েন্ট থেকে প্রথম K-সংখ্যক পয়েন্ট নিতে হবে।
- এই K-সংখ্যক ডেটা পয়েন্টের মধ্যে যে ক্লাসের পয়েন্ট সবচেয়ে বেশি সংখ্যকবার আছে, অজানা ডেটা পয়েন্টটিকে সেই ক্লাসে হিসেবে চিহ্নিত করতে হবে।

পরিচ্ছেদ ৬.২ : উদাহরণ

এখন আমরা একটি উদাহরণ দিয়ে পুরো অ্যালগরিদমটির বিস্তারিত বুঝব। নিচের চার্টটি দেখি (টেবিল 6.2.1)। এই চার্টে ওপরের চিত্রের ডেটা পয়েন্টগুলোর (ত্রিভুজ ও বর্গ) x_1 এবং x_2 ফিচার ভ্যালুগুলোর মান দেওয়া আছে। সেই সঙ্গে আমরা কোন ডেটা পয়েন্ট কোন ক্লাসের

অন্তর্ভুক্ত, সেটিও শেষ কলামে লিখে দিয়েছি। এখানে 1 মানে কুকুর (ত্রিভুজ), 0 মানে বিড়াল (বর্গ)।

X_1	X_2	Class
4.2	2.8	1
4.0	2.0	1
3.8	0.5	1
2.0	1.5	1
2.7	2.5	1
1.7	3.2	0
2.7	4.0	0
1.2	5.2	0
2.2	6.2	0
0.3	6.2	0

টেবিল 6.2.1

অজানা ডেটা পয়েন্টের ফিচারে দুটির মান হচ্ছে (2.2, 3)। আমাদের এখন বের করতে হবে যে এই অজানা ডেটা পয়েন্ট কুকুর হবে, নাকি বিড়াল।

সেজন্য আমরা এখন যেটি করব, সবার প্রথমে (2.2, 3) পয়েন্ট থেকে টেবিলের সব ডেটা পয়েন্টের মধ্যেকার দূরত্ব বের করব। দূরত্ব বের করার জন্য আমরা ইউক্লিডীয় দূরত্বের সূত্র প্রয়োগ করব। দুটি পয়েন্ট (x_1, y_1) ও (x_2, y_2)-এর মধ্যেকার ইউক্লিডীয় দূরত্ব হচ্ছে,

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

এখন তাহলে আমরা সব ডেটা পয়েন্ট থেকে (2.2, 3) পয়েন্টের ইউক্লিডীয় দূরত্ব বের করে ফেলি:

X_1	X_2	Distance Formula	Distance Value	Sort Rank
4.2	2.8	$\sqrt{(4.2-2.2)^2 + (2.8-3)^2}$	2.00998	5
4.0	2.0	$\sqrt{(4-2.2)^2 + (2-3)^2}$	2.05913	6
3.8	0.5	$\sqrt{(3.8-2.2)^2 + (0.5-3)^2}$	2.96816	8

2.0	1.5	$\sqrt{(2-2.2)^2+(1.5-3)^2}$	1.51327	4
2.7	2.5	$\sqrt{(2.7-2.2)^2+(2.5-3)^2}$	0.70711	2
1.7	3.2	$\sqrt{(1.7-2.2)^2+(3.2-3)^2}$	0.53852	1
2.7	4.0	$\sqrt{(2.7-2.2)^2+(4-3)^2}$	1.11803	3
1.2	5.2	$\sqrt{(1.2-2.2)^2+(5.2-3)^2}$	2.41661	7
2.2	6.2	$\sqrt{(2.2-2.2)^2+(6.2-3)^2}$	3.2	9
0.3	6.2	$\sqrt{(0.3-2.2)^2+(6.2-3)^2}$	3.72156	10

টেবিল 6.2.2

এখন আমরা এই ডেটা পয়েন্টগুলোকে দূরত্বের মানে ছোটো থেকে বড়ো ক্রম বা অ্যাসেন্ডিং অর্ডার (ascending order)-এ সর্ট করি। সর্ট করার পরে টেবিলে ডেটা পয়েন্টগুলোর অবস্থান ওপরের টেবিলে Sort Rank হিসেবে দেওয়া আছে। সেখান থেকে ডেটা পয়েন্টগুলো নিয়ে যদি আমরা একটি আলাদা টেবিলে একই ক্রমে সাজাই, তাহলে নিচের টেবিলের মতো দাঁড়াবে :

X_1	X_2	Class
1.7	3.2	0
2.7	2.5	1
2.7	4.0	0
2.0	1.5	1
4.2	2.8	1
4.0	2.0	1
1.2	5.2	0
3.8	0.5	1
2.2	6.2	0
0.3	6.2	0

টেবিল 6.2.3

এখন আমরা K = 3 ধরে নিলে, আমরা সবচেয়ে কাছের তিনটি বিন্দু নিয়ে কাজ করব। ওপরের টেবিল থেকে আমরা প্রথম তিন সারির ডেটা পয়েন্ট নিলেই সবচেয়ে কাছের তিনটি বিন্দুর ক্লাস আমরা পেয়ে যাচ্ছি, যা যথাক্রমে হলো 0, 1, 0 অর্থাৎ দুটি বিড়াল ও একটি কুকুর (ধরে নিয়েছিলাম

1 = কুকুর , 0 = বিড়াল)। যেহেতু বিড়ালের সংখ্যা বেশি, তাই আমাদের (2.2, 3) পয়েন্টটির ক্লাস হিসেবে আমরা 0 অর্থাৎ বিড়াল হিসেবে চিহ্নিত করব।

আবার যদি, $K = 5$ ধরে নিই, তাহলে আমরা সবচেয়ে কাছের পাঁচটি বিন্দু নিয়ে কাজ করব। ওপরের টেবিল থেকে আমরা প্রথম পাঁচ সারির ডেটা পয়েন্ট নিলেই সবচেয়ে কাছের ছয়টি বিন্দুর ক্লাস আমরা পেয়ে যাচ্ছি, যা যথাক্রমে হলো 0, 1, 0, 1, 1 অর্থাৎ দুটি বিড়াল ও তিনটি কুকুর। সুতরাং, এই ক্ষেত্রে তাই আমাদের (2.2, 3) পয়েন্টটির ক্লাস হিসেবে আমরা 1 অর্থাৎ কুকুর অ্যাসাইন করব।

শেষ করার আগে, একটি বিষয় জেনে রাখা ভালো। যদি K-এর মান খুব ছোটো হয়, তখন ব্যাপারটি এরকম দাঁড়ায় যে কাছাকাছি অল্প দু-তিনটি ডেটা পয়েন্ট দেখেই আমরা সিদ্ধান্তে পৌঁছে যাই, এর ফলে বিভিন্ন ধরনের noise (বাজে ডেটা, যেগুলো আমাদের ডেটাসেটের কোনো অংশ নয়, ভুলভ্রান্তি) দিয়ে আমাদের সিদ্ধান্ত বায়াসড (biased) বা পক্ষপাতদুষ্ট হয়ে যেতে পারে। খুব সহজভাবে চিন্তা করুন, যদি আপনি মাত্র তিনজন মানুষকে জিজ্ঞাসা করেন যে, 'চুরি করা কি ভালো?' এবং দুর্ভাগ্যক্রমে ওই তিনজনের মধ্যে দুজন যদি হয় চোর এবং তারা আপনাকে 'হ্যাঁ' বলে এবং বাকি একজন 'না' বলে, তখন আপনি কী করবেন? যেহেতু দুজন 'হ্যাঁ' বলেছে, সেহেতু সেটিই আপনি মেনে নেবেন এবং হয়তো চুরি করা ভালো কাজ মনে করবেন, তাই না? তার মানে কী দাঁড়াল? এখানে চোর দুজনকে আমরা noise ডেটার সঙ্গে তুলনা করতে পারি। তাই, k-এর মান খুব ছোটো হলে noise ডেটা দিয়ে প্রভাবিত হওয়ার সম্ভাবনা বেশি থাকে। ফলে ভ্যারিয়েন্স বেড়ে যায়, বায়াস কমে যায়।

একই ভাবে, যদি k-এর মান আমরা অনেক বেশি নিই, (ধরুন $k = 1000$), তাহলে আবার সমস্যা হলো আগের ঘটনার উলটো ঘটনা ঘটবে, অর্থাৎ ভ্যারিয়েন্স কমে গিয়ে বায়াস বেড়ে যাবে। আর তাই, k-এর মান নির্ধারণ করার সময় ভারসাম্য বজায় রাখতে হয়, যাতে খুব বড়োও না হয় আবার খুব ছোটোও না হয়।

K-এর অপটিমাল (optimal) মান নেওয়ার একটি উপায় হচ্ছে, মোট যত ডেটা আছে, (ধরি n সংখ্যক) তার বর্গমূলের মানকে আমরা K-এর অপটিমাল মান হিসেবে নিতে পারি। যদি, ধরা যাক $n = 150$ হয়, তবে $K \cong \sqrt{150} \cong 12.28 \cong 13$ নিতে পারি (যেহেতু K-এর মান বেজোড় হবে তাই আমরা 12 নেব না, 13 নেব)।

এই ছিল KNN অ্যালগরিদমের বিবরণ। এটি আসলে খুবই ছোটো এবং সহজ একটি ক্লাসিফিকেশন অ্যালগরিদম। আশা করি, সবাই বুঝতে পেরেছেন কীভাবে KNN অ্যালগরিদমের সাহায্যে আমরা ক্লাসিফিকেশন করতে পারি।