অধ্যায় ৮ : নাইভ বেইজ ক্লাসিফায়ার (Naive Bayes Classifier)

ছবি 8.1 : Thomas Bayes (1702-1761)

আশা করি, আপনারা সবাই উচ্চমাধ্যমিক গণিতে বেইজ থিওরেম পড়েছেন। আমার যতদূর মনে পড়ে, উচ্চমাধ্যমিক গণিত বইয়ে *নষ্ট বল্টু তৈরি করা মেশিন* ইত্যাদি সম্পর্কিত কোনো একটি বিখ্যাত অঙ্ক ছিল, যেটি এই বেইজ থিওরেম দিয়ে করতে হতো। ওই অঙ্ক দেখলেই আমার ঘাম ছুটে যেত। ওই অধ্যায়ের অঙ্কগুলো করতে আমি ভীষণ ভয় পেতাম। আর তার কারণ একটিই, আমি বুঝতাম না কী করছি, কেন করছি। এই অধ্যায়ের অঙ্কগুলো না বুঝে অনেকটা মুখস্থ করার মতো করেছিলাম, সে কারণেই ভয় পেতাম।

আশা করি, আপনারা সবাই উচ্চমাধ্যমিক গণিতে বেইজ থিওরেম পড়েছেন। আমার যতদূর মনে পড়ে, উচ্চমাধ্যমিক গণিত বইয়ে *নষ্ট বল্টু তৈরি করা মেশিন* ইত্যাদি সম্পর্কিত কোনো একটি বিখ্যাত অঙ্ক ছিল, যেটি এই বেইজ থিওরেম দিয়ে করতে হতো। ওই অঙ্ক দেখলেই আমার ঘাম ছুটে যেত। ওই অধ্যায়ের অঙ্কগুলো করতে আমি ভীষণ ভয় পেতাম।

আর তার কারণ একটিই, আমি বুঝতাম না কী করছি, কেন করছি। এই অধ্যায়ের অঙ্কগুলো না বুঝে অনেকটা মুখস্থ করার মতো করেছিলাম, সে কারণেই ভয় পেতাম।

সম্ভাব্যতা (Probability) এখন আমার অনেক প্রিয় একটি বিষয়। এটি আমার পড়তেও ভালো লাগে, পড়াতেও ভালো লাগে। আর এই নাইভ বেইজ ক্লাসিফায়ার বুঝতে হলে আমাদেরকে উচ্চ মাধ্যমিকে পড়ে আসা বেইজ থিওরেম এবং সেই সঙ্গে সম্ভাব্যতার কিছু জিনিস একটু ঝালাই করে নিতে হবে।

কিন্তু তার আগে একটু বলে নিই, বেইজ থিওরেমের জনক রেভারেন্ড থমাস বেইজ (Thomas Bayes, 1702-1761)। তাঁর নামেই এর নামকরণ। থমাসের কিছু অসমাপ্ত গবেষণার হাত ধরেই উদ্ভাবিত হয় এই থিওরেমের।

পরিচ্ছেদ ৮.১ : সম্ভাব্যতার টুকিটাকি

খুব সহজ করে বলতে গেলে, সম্ভাব্যতা হচ্ছে কোনো ঘটনা ঘটার সম্ভাবনার একটি গাণিতিক প্রকাশ। যেখানেই আপনারা সম্ভাব্যতা পাবেন, সেখানেই দেখবেন তার সঙ্গে কোনো সংখ্যা ও গাণিতিক প্রকাশ জড়িত। যদি বলি, আজকে বৃষ্টি হওয়ার সম্ভাব্যতা 70%, সেটি গাণিতিক সম্ভাব্যতার একটি উদাহরণ। এখন, এই সম্ভাব্যতার মান 0 থেকে 1-এর ভেতরে যে-কোনো সংখ্যা হতে পারে। 0 মানে ঘটনাটি ঘটার কোনো সম্ভাবনাই নেই, আর 1 মানে ঘটনাটি ঘটবেই, কোনো নড়চড় হবে না। সম্ভাব্যতা 0.5 মানে অর্ধার্ধি সম্ভাবনা আছে ঘটনাটি ঘটার। যেমন, আপনি যদি একটি কয়েন ছুড়ে মারেন ওপরে টস করার জন্য, 50% সম্ভাবনা আছে Head আসার, আর 50% সম্ভাবনা আছে Tail আসার। অর্থাৎ Head আসার Probability 0.5।

এই গেল, সম্ভাব্যতার মান কত থেকে কত হতে পারে, তার একটি বর্ণনা। এখন আসি কীভাবে সম্ভাব্যতা নির্ণয় করবেন সে উপায়ে। সম্ভাব্যতা বের করার সূত্র হচ্ছে =

$$\frac{\textit{Number of Events occured in Favor of Expected Outcome}}{\textit{Number of total events occured irrespective of Expected Outcome}}$$

এ কথার মানে কী? একটি লুডুর ছক্কার কথা চিন্তা করুন। একটি ছক্কা যদি আমরা চালি, তাহলে আমরা কয় ধরনের ভিন্ন ভিন্ন মান পেতে পারি? 6 ধরনের, তাই না? 1, 2, 3, 4, 5 ও 6। তার মানে, একটি ছক্কা চাললে মোট ছয় ধরনের ঘটনা ঘটতে পারে।

এখন যদি আমরা চাই যে আমাদের ছক্কায় 4 উঠুক, সেটি কয়টি ঘটনার জন্য ঘটতে পারে? ছক্কার ওপরে 1, 2, 3 ইত্যাদি থাকলে কি আমরা 4 উঠেছে বলে ধরে নেব? মোটেই না। ছক্কায় ওপরে শুধু 4 উঠলেই কেবল আমরা আমাদের প্রত্যাশিত ফলাফল (Expected Outcome) পেতে পারি। সুতরাং, সূত্র অনুযায়ী একটি ছক্কা চাললে তাতে 4 ওঠার সম্ভাবনা $\frac{1}{6}$। আবার আমরা যদি চাই যে, আমাদের ছক্কায় শুধু বেজোড় মান উঠুক, অর্থাৎ 1, 3 কিংবা 5 উঠুক, তাহলে তার সম্ভাব্যতা হবে $\frac{3}{6} = \frac{1}{2}$।

এই গেল আমাদের সম্ভাব্যতা কীভাবে নির্ণয় করতে হয় তার পদ্ধতি। আরেকটি বিষয় আমাদের জেনে নিতে হবে, সেটি হচ্ছে, শর্তাধীন সম্ভাব্যতা বা কন্ডিশনাল প্রোবাবিলিটি (Conditional Probability)। অর্থাৎ, আপনাকে কোনো একটি ঘটনা (বা, শর্ত) দিয়ে দেওয়া হবে, সেটি ঘটেছে ধরে নিয়ে সেই ঘটনার সাপেক্ষে আপনাকে অন্য কোনো ঘটনা ঘটার সম্ভাব্যতা বিচার করতে হবে।

লুডুর ছক্কা দিয়েই বোঝাই। ধরা যাক, আপনাকে একটি সাধারণ ছক্কা দিয়ে বলল তাতে 5 ওঠার সম্ভাব্যতা কত? আপনি সঙ্গে সঙ্গে উত্তর দিয়ে দিতে পারবেন, $\frac{1}{6}$, তাই না? কিন্তু যদি এখন আপনাকে বলে দেওয়া হয়, আপনি ছক্কা চাললে আপনার কোনো জোড় সংখ্যা উঠবে না। এবার

যদি বলা হয় এই ঘটনার সাপেক্ষে ছক্কা চেলে 5 পাওয়ার সম্ভাবনা কত তা বের করতে, তখন কীভাবে সেটি হিসাব করবেন?

দেখুন, আপনাকে যেহেতু বলেই দেওয়া হচ্ছে যে আপনার কোনো জোড় সংখ্যা উঠবে না, তার মানে দাঁড়াচ্ছে ছক্কা চেলে 2, 4 ও 6 পাওয়ার কোনো সম্ভাবনা-ই নেই। তাহলে আপনার ছক্কা চেলে উঠতে পারে 1, 3 কিংবা 5। তাহলে এই ক্ষেত্রে আপনার 5 পাওয়ার সম্ভাব্যতা বেড়ে দাঁড়াবে $\frac{1}{3}$-তে। একইভাবে যদি আপনাকে বলে দেওয়া হতো যে, আপনি ছক্কা চাললে আপনার 4 বাদে অন্য যে-কোনো সংখ্যা উঠতে পারে, তাহলে সে ক্ষেত্রে আপনার 5 পাওয়ার সম্ভাবনা হতো $\frac{1}{5}$। এটিই হচ্ছে শর্তাধীন সম্ভাব্যতা।

যদি x কোনো ঘটনা বোঝায়, যার মানে হচ্ছে, 'c is Not Even' এবং c দিয়ে ছক্কার ওপরে কোন মান উঠবে সেটি নির্দেশ করা হয়, তাহলে আমাদের ছক্কায় জোড় সংখ্যা উঠবে না, এই শর্তসাপেক্ষে ছক্কায় 5 পড়ার সম্ভাব্যতাকে লেখা হবে $P(c = 5 \mid x)$; যেখানে x হচ্ছে আমাদের আগে থেকে দিয়ে দেওয়া শর্ত এবং $P(c = 5 \mid x)$ মানে বোঝাচ্ছে এই শর্তের অধীন থাকা অবস্থায় $c = 5$ হওয়া সম্ভাব্যতা। সুতরাং,

$$P(c = 5|x) = \frac{1}{3}$$

পরিচ্ছেদ ৮.২ : বেইজ থিওরেম হাতে-কলমে

এখন আসি বেইজ থিওরেমে। সাধারণত, শর্তাধীন সম্ভাব্যতার ক্ষেত্রে আমরা আগে কোনো একটি ঘটনা ঘটে গেছে, তার সাপেক্ষে পরের ঘটনা ঘটার সম্ভাব্যতা কত সেটি বের করি।

যেমন, *বৃষ্টি হয়েছে, তার সাপেক্ষে রাস্তায় কাদা-পানি থাকার সম্ভাবনা কত?* কিংবা, *ক্যানসার হয়েছে, তার সাপেক্ষে রোগীর মৃত্যুর সম্ভাব্যতা কত?* ইত্যাদি।

কিন্তু, বেইজ থিওরেমে এর বিপরীত সম্ভাব্যতা বের করা হয়। পরের যে ঘটনাটি ঘটেছে, তার সম্ভাব্যতা আমরা জানি, তার সাপেক্ষে আগের কোনো একটি ঘটনা ঘটার সম্ভাব্যতা কত (অর্থাৎ, পরের ঘটনাটি আগের কোনো একটি নির্দিষ্ট ঘটনার কারণে ঘটছে তার সম্ভাব্যতা কত), সেটি আমাদের বের করতে হয়। অর্থাৎ, রাস্তায় কাদা-পানি দেখা যাচ্ছে, এখন এই কাদা-পানি যে বৃষ্টির কারণেই হয়েছে, রাস্তায় কেউ এক বালতি পানি ছুড়ে মারার কারণে হয়নি, তার সম্ভাব্যতা কত; কিংবা, রোগী মারা গিয়েছে, কিন্তু রোগী যে ক্যানসারজনিত কারণেই মারা গিয়েছে, অন্য কোনো রোগের কারণে নয়, তার সম্ভাব্যতা কত, এরকম একটি বিষয়।

আমরা আগের ঘটনাকে যদি x ধরি এবং পরের ঘটনাকে যদি c ধরি এবং বেইজ থিওরেম প্রয়োগ করে যদি পরের ঘটনাটি যে আগের ঘটনার কারণেই ঘটেছে তার সম্ভাব্যতা বের করতে চাই, তাহলে সূত্র হবে এরকম –

$$P(X \mid C) = \frac{P(C \mid X) \cdot P(X)}{P(C)}$$

খুব ছোট্ট একটি উদাহরণ দিয়ে বেইজ থিওরেম বোঝা শেষ করি। ধরা যাক, কোনো এলাকায় এক সমীক্ষা থেকে জানা যায়, সেখানকার এলাকাবাসীর ক্যানসার হওয়ার ঝুঁকি রয়েছে এবং প্রত্যেকের ক্যানসার হওয়ার সম্ভাব্যতা শতকরা 60 ভাগ। এ ছাড়াও, আরো জানা যায় যে ক্যানসারজনিত কারণে আগামী দুই বছরের মধ্যে ওই এলাকার কারো মারা যাওয়ার সম্ভাবনা শতকরা 85 ভাগ এবং ক্যানসার না হয়ে মারা যাওয়ার সম্ভাবনা শতকরা 45 ভাগ।

এখন, ধরা যাক, ওই এলাকায় 'ক' নামে এক বয়স্ক ভদ্রলোক থাকতেন, যিনি ওই সমীক্ষা চালানোর ঠিক এক বছরের মাথায় মারা যান। এখন আপনাকে বের করতে হবে যে ওই ভদ্রলোক ক্যানসারজনিত কারণে মারা গিয়েছেন, তার সম্ভাব্যতা কত?

এটি কীভাবে বের করবেন? প্রথমে বের করতে হবে যে আগের এবং পরের ঘটনা কোনটি? এখানে ঘটনা আছে দুটি – ক্যানসার হওয়া এবং মারা যাওয়া। কোনটি আগে হবে বলুন তো? ক্যানসার হয়ে মানুষ মারা যাবে, নাকি মানুষ মারা যাওয়ার পরে তাঁর ক্যানসার হবে? অবশ্যই প্রথমটি, তাই না?

তার মানে, এ ক্ষেত্রে,

আগের ঘটনা, x = ক্যানসার হওয়া, এবং

পরের ঘটনা, c = মারা যাওয়া।

আমাদেরকে বের করতে হবে $P(X \mid C)$ কত?

প্রথমে চলুন বের করি, $P(C \mid X)$ কত?

$P(C \mid X)$ হচ্ছে ক্যানসার হওয়ার পরে মারা যাওয়ার সম্ভাব্যতা = 85% = 0.85।

এর পরে, $P(X)$ = ক্যানসার হওয়ার সম্ভাব্যতা = 60% = 0.60।

আর সবশেষে $P(C)$ = মারা যাওয়ার সম্ভব্যতা

= (ক্যানসার হওয়ার সম্ভাব্যতা × ক্যানসারের কারণে মৃত্যুর সম্ভাব্যতা)

+ (ক্যানসার না হওয়ার সম্ভাব্যতা × ক্যানসার না হয়েই মৃত্যুর সম্ভাব্যতা)

$= (0.6 \times 0.85) + ((1 - 0.6) \times 0.45)$

$= 0.51 + 0.18$

$= 0.69$

সুতরাং, 'ক'-এর মৃত্যু যে ক্যানসারেই হয়েছে, তার সম্ভাব্যতা,

$$P(X \mid C) = \frac{0.85 \times 0.6}{0.69} = 0.739 = 73.9\%$$

যাক, আশা করি আপনারা বেইজ থিওরেম কীভাবে কাজ করে এবং কীভাবে প্রয়োগ করতে হয় সবাই কমবেশি বুঝতে পেরেছেন। এখন আমরা দেখব কীভাবে বেইজ থিওরেম ব্যবহার করে নাইভ বেইজ ক্লাসিফায়ার তৈরি ও ডেটাসেটের ওপরে ব্যবহার করতে হয়।

পরিচ্ছেদ ৮.৩ : হাতে কলমে নাইভ বেইজ ক্লাসিফায়ার

আমরা প্রথমে একটি ছোটো ডেটাসেট নিই :

Day	Outlook	Temperature	Routine	Wear Coat?
D_1	Sunny	Cold	Indoor	No
D_2	Sunny	Warm	Outdoor	No
D_3	Cloudy	Warm	Indoor	No
D_4	Sunny	Warm	Indoor	No
D_5	Cloudy	Cold	Indoor	Yes
D_6	Cloudy	Cold	Outdoor	Yes
D_7	Sunny	Cold	Outdoor	Yes

টেবিল 8.3.1

ডেটাসেটটি যদি আপনারা ভালো করে লক্ষ করেন, তাহলে দেখবেন যে মধ্যের তিনটি কলাম হচ্ছে আমাদের ফিচার ডেটা, আর শেষের কলামটি হচ্ছে আমাদের সর্বশেষ যে ফলাফল হবে সেটি।

এখন, আমাদের বের করতে হবে যে, যদি আমাদের ফিচার সেট এরকম হয় – {Cloudy, Warm, Outdoor} তাহলে ফলাফল কী হবে? বাইরে বের হওয়ার সময় কোট পরব, নাকি পরব না?

এটি বের করার জন্য আমরা যদি নাইভ বেইজ ক্লাসিফায়ার প্রয়োগ করতে চাই, তাহলে আমাদেরকে দুটি মান বের করতে হবে –

1. $P(C = Yes \mid X = cloudy, warm, outdoor)$ এবং
2. $P(C = No \mid X = cloudy, warm, outdoor)$।

এখানে $C = yes / no$ দিয়ে *Wear Coat*-এর মান *yes / no* বোঝানো হয়েছে।

প্রথমটির ক্ষেত্রে,

$$P(C = \text{Yes} \mid X) = \frac{P(X \mid C=\text{Yes}) \times P(C=\text{Yes})}{P(X)}$$

$$= \frac{P(Cloudy \mid C=\text{Yes}) \cdot P(Warm \mid C=\text{Yes}) \cdot P(Outdoor \mid C=\text{Yes}) \cdot P(C=\text{Yes})}{P(Cloudy) \cdot P(Warm) \cdot P(Outdoor)}$$

$$= \frac{\frac{2}{3} \cdot \frac{0}{3} \cdot \frac{2}{3} \cdot \frac{3}{7}}{\frac{3}{7} \cdot \frac{3}{7} \cdot \frac{3}{7}}$$

$$= 0$$

এখন, অনেকেই হয়তো বুঝতে পারেননি, $P(Cloudy \mid C = \text{Yes})$-এর মান কেন $\frac{2}{3}$ হলো। ওপরের টেবিল 8.3.1-এ প্রথমেই দেখুন, *Wear Coat*-এর মান yes আছে মোট 3 জায়গায় (শেষের তিন সারিতে)। এই তিন সারিতে দেখুন, *Outlook* কলামে Cloudy আছে মোট 2টি আর Sunny আছে 1টি। অর্থাৎ তিনটি C = Yes-এর মধ্যে দুটির জন্য আমরা *Outlook* কলামে Cloudy পাচ্ছি।

তাই $P(Cloudy \mid C = \text{Yes})$-এর মান দাঁড়ায়, $\frac{\text{মোট কয়টি } Cloudy \text{ আমরা } C = yes \text{এর জন্য পেয়েছি}}{\text{মোট কয়টি } Cloudy \text{ আছে } outlook \text{ কলামে}} = \frac{2}{3}$

এইভাবে বাকিগুলোও হিসাব করা হয়েছে।

এখন, একইভাবে,

$$P(C = \text{No} \mid X) = \frac{P(X \mid C=\text{No}) \times P(C=\text{No})}{P(X)}$$

$$= \frac{P(Cloudy \mid C=\text{No}) \cdot P(Warm \mid C=\text{No}) \cdot P(Outdoor \mid C=\text{No}) \cdot P(C=\text{No})}{P(Cloudy) \cdot P(Warm) \cdot P(Outdoor)}$$

$$= \frac{\frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{4}{7}}{\frac{3}{7} \cdot \frac{3}{7} \cdot \frac{3}{7}}$$

$$= \frac{49}{144}$$

$$= 0.340$$

যেহেতু $P(C = No \mid X) > P(C = Yes \mid X)$, সুতরাং, {Cloudy, Warm, Outdoor} এই ডেটা পয়েন্টের জন্য ফলাফল হবে No। এভাবেই একটি নাইভ বেইজ ক্লাসিফায়ার কাজ করে। আশা করি সবাই বুঝতে পেরেছেন।