220515

Agenda: ① Naive Bayes Algorithm (Classification Algorithm)
② KNN

1. Naive Bayes Algorithm

Rolling a dice {1, 2, 3, 4, 5, 6} → Independent events

$P(1) = 1/6$ $P(3) = 1/6$ $P(5) = 1/6$

Picking marbles from box → Dependent events

1. $P(B) = 3/6 = 1/2$

Conditional Probability : Probability of Green if Blue is already taken once before

2. $P(G/B) = 3/5$

$P(B \text{ and } G) = P(B) \times P(G/B)$

$= 1/2 \times 3/5 = 3/10 = 0.3_{4}$

$P(A \text{ and } B) = P(A) \times P(B/A)$

$P(A \text{ and } B) = P(B \text{ and } A)$

$P(A) \times P(B/A) = P(B) \times P(A/B)$

$$P(B/A) = \frac{P(B) \times P(A/B)}{P(A)}$$ Bayes' Theorem

Dataset : $\underbrace{x_1, x_2, \dots, x_n}_{\text{input features}}$ $\quad$ Y (output)

$$P(Y/(x_1, x_2, \dots, x_n)) = \frac{P(Y) \times P((x_1, x_2, \dots, x_n)/Y)}{P((x_1, x_2, \dots, x_n))}$$

$$= \frac{P(Y) \cdot P(x_1/Y) \cdot P(x_2/Y) \cdot \dots \cdot P(x_n/Y)}{P(x_1) \cdot P(x_2) \cdot \dots \cdot P(x_n)}$$

Binary Dataset

x_1	x_2	x_3	o/p
			Yes
			No

$$P(y = Yes/x_i) = \frac{P(Yes) \cdot P(x_1/Yes) \cdot P(x_2/Yes) \cdot P(x_3/Yes)}{P(x_1) \cdot P(x_2) \cdot P(x_3)}$$ → $P((x_1, x_2, x_3)/Yes)$

$$P(y = No/x_i) = \frac{P(No) \cdot P(x_1/No) \cdot P(x_2/No) \cdot P(x_3/No)}{P(x_1) \cdot P(x_2) \cdot P(x_3)}$$

Let's say $P(Yes/x_i) = 0.13$

$P(No/x_i) = 0.05$

$P(Yes/x_i) = \frac{0.13}{0.13+0.05} = \frac{0.13}{0.18} \approx \%72$

$P(No/x_i) = \frac{0.05}{0.13+0.05} = \frac{0.05}{0.18} \approx \%28$

$+$ } $\%100$ $\geq \%50$ $< \%50$

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Outlook

	Yes	No	P(Y)	P(N)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rain	3	2	3/9	2/5
	9	5		

Temperature

	Yes	No	P(Y)	P(N)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cold	3	1	3/9	1/5
	9	5		

Play

Yes	9	P(Yes)	P(No)
No	5	9/14	5/14

(Sunny, Hot) → O/P Naive Bayes → ?

$$P(Yes/(sunny,hot)) = \frac{P(Yes)\cdot P(sunny/Yes)\cdot P(hot/Yes)}{P(sunny)\cdot P(hot)} = 9/14 \cdot \frac{2}{9} \cdot \frac{2}{9} = \frac{2}{63} = 0.031$$

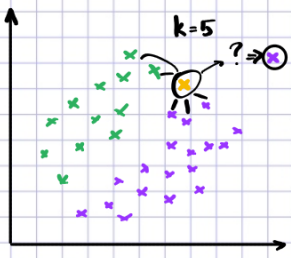
$$P(No/(sunny,hot)) = \frac{P(No)\cdot P(sunny/No)\cdot P(hot/No)}{P(sunny)\cdot P(hot)} = 5/14 \cdot 3/5 \cdot 2/5 = \frac{3}{35} = 0.0857$$

Real Probability

$P(Yes/(sunny,hot)) = 0.031$ $= \frac{0.031}{0.031+0.0857} = \%27$

$P(No/(sunny,hot)) = 0.0857$ $= 1 - \%27 = \%73$

2. K-Nearest Neighbors (KNN) [Classification & Regression]

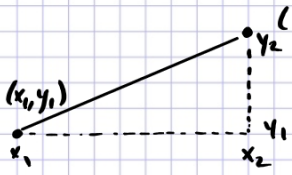


Calculate shortest distance

① Euclidean Distance

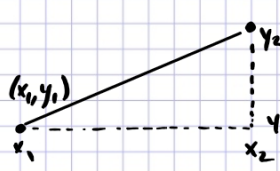
② Manhattan Distance

1. Euclidean Distance



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

2. Manhattan Distance



$$|(x_2 - x_1) + (y_2 - y_1)|$$

The same idea is used in regression problem. Output is calculated as average of k-nearest data points' y-values.