Naïve Bayes

Author - Kanav Bansal
You can find me on **Linkedin** -
www.linkedin.com/in/kanavbansal
Or http://www.thataiguy.com

Building a Machine Learning Model

1. Training

X_train_transformed , y_train → [Algorithm] → (Model)

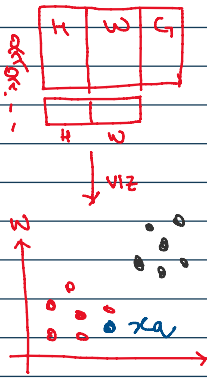
2. Predicting using unseen data

X_test_transformed → (Model) → y_test_pred

3. Evaluation

y_test , y_test_pred → [Evaluation Metric] → Score

- Given Data containing Height, Weight & Gender of some individuals.



* Predict the Gender for x_q

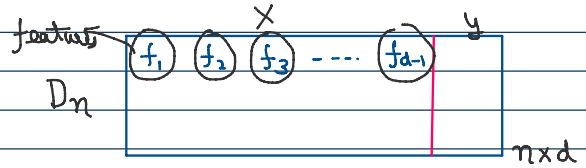
Type → Task
? ?

Evaluation Metric → ?

1. Target Variable → Gender (y)
   Gender $\in \{M, F\}$
2. Task → Classification
3. Evaluation Metrics → Accuracy
4. Algorithms → (How?) Probabilistic Approach → Naive Bayes Algorithm

Naive Bayes Algorithm (AKA Maximum A Posteriori)

Type – Supervised Learning
Task – Classification.

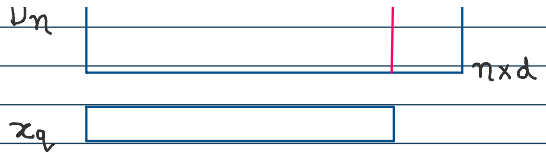


Given → $D_n = \{(x_i, y_i)_{i=1}^n \mid x_i \in \mathbb{R}^{d-1}, y_i \in \{C_1, C_2 \dots C_k\}\}$

$x_q = f_1 \cap f_2 \cap \dots \cap f_{d-1}$

Task → Find y_q

How? → Given x_q, compute the prob of x_q belonging to each of the classes. Whichever is maximum, assign that class to x_q.



Whichever is maximum, assign that class to x_q.

ie $P(C_1|x_q)$ $P(C_2|x_q)$ --- $P(C_K|x_q)$

$$y_q = \underset{C_1, C_2 \dots C_K}{\arg\max} \{ P(C_i|x_q) \}$$

Posteriori Prob.

From Bayes Theorem ⤵

$$\underbrace{P(C_i|x_q)}_{\text{Posterior}} = \frac{\overbrace{P(x_q|C_i)}^{\text{Likelyhood}} * \overbrace{P(C_i)}^{\text{Prior}}}{P(x_q)} \quad \text{Marginal}$$

Key Concepts ⤵

→ Probability

$$P(\text{Event}) = \frac{\text{No. of favourable Outcomes}}{\text{Total no. of Outcomes}} = \frac{|\text{Event}|}{|\text{Sample Space}|}$$

→ In case of two or more events ⤵

Event A & B

- Marginal Probability i.e $P(A)$ & $P(B)$
- Joint Probability i.e $P(A \cap B)$
- Conditional Probability i.e $P(A|B)$

Bayes Theorem →

Acc^{n} to Conditional Prob → $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$P(B|A)$: Likelyhood; $P(A)$: Prior; $P(B)$: Marginal; $P(A|B)$: Posteriori

Independent Events ⤵

If given events A & B are know to be independent:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Similarly:

$$P(A \cap B) = P(A) * P(B)$$

Data →

$$D_n = \left\{ (x_i, y_i)_{i=1}^{n} \;\middle|\; x_i \in \mathbb{R}^{d-1},\ y_i \in \{0,1\} \right\}$$

f_1	f_2	----	f_{d-1}	y
← x_1 →				y_1
← x_2 →				y_2
⋮				⋮
← x_n →				y_n

$$x_i = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{d-1} \end{bmatrix}$$

Naive Bayes Algorithm

Task → Classification

How?

Given x_q, compute $P(C_1|x_q)$, $P(C_2|x_q)$, ---- $P(C_K|x_q)$.

Which ever probability is maximum, prediction would be the corresponding class.

$$\hat{y}_q = \underset{i \in \{1,2 \cdots k\}}{\arg\max} \{P(C_i|x_q)\} \quad \text{i.e. maximum a posteriori}$$

$$\rightarrow P(C_1|x_q) = \frac{P(x_q|C_1)\,P(C_1)}{P(x_q)} \Rightarrow P(C_1|x_q) \propto P(x_q|C_1)\,P(C_1)$$

$$\rightarrow P(C_2|x_q) = \frac{P(x_q|C_2)\,P(C_2)}{P(x_q)} \Rightarrow P(C_2|x_q) \propto P(x_q|C_2)\,P(C_2)$$

⋮ ⋮

$$\rightarrow P(C_K|x_q) = \frac{P(x_q|C_K)\,P(C_K)}{P(x_q)} \Rightarrow P(C_K|x_q) \propto P(x_q|C_K)\,P(C_K)$$

Naive Bayes Derivation

Given $x_q = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{d-1} \end{bmatrix}$ i.e. x_q is $f_1 \cap f_2 \cap f_3 \cap \cdots \cap f_{d-1}$

Using Bayes Theorem:

$$P(C_K|x_q) = \frac{P(x_q|C_K) * P(C_K)}{P(x_q)}$$

Compute → $\boxed{P(C_K|x_q) \propto P(x_q|C_K)\,P(C_K)}$

$$\hat{y}_q = \underset{k \in \{1,2 \ldots K\}}{\arg\max} \left\{ P(C_k \mid X_q) \right\}$$

$$\hat{y}_q = \underset{k \in \{1,2 \ldots K\}}{\arg\max} \left\{ \boxed{P(X_q \mid C_k)\, P(C_k)} \right\}$$

↳ Lets zoom inside this term and understand how to compute it step by step.

According to Conditional Probability],

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$\Rightarrow P(A \cap B) = P(A \mid B)\, P(B)$$

$P(X_q \mid C_k)\, P(C_k)$ can be written as $P(X_q \cap C_k)$

We have already seen that X_q is $f_1 \cap f_2 \cdots \cap f_{d-1}$

→ $P(X_q \mid C_k)\, P(C_k)$

→ $P(X_q \cap C_k)$

→ $P(f_1 \cap f_2 \cap f_3 \cdots f_{d-1} \cap C_k)$

$P(\underbrace{f_1}_{A} \cap \underbrace{f_2 \cap f_3 \cdots f_{d-1} \cap C_k}_{B})$

Using $P(A \cap B) = P(A \mid B)\, P(B)$

→ $P(f_1 \mid f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k) *$
$P(\underbrace{f_2}_{A} \cap \underbrace{f_3 \ldots f_{d-1} \cap C_k}_{B})$

Expanding this expressing using Chain Rule of conditional Probability

→ $P(f_1 \mid f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k) *$
$P(f_2 \mid f_3 \cap f_4 \cdots \cap f_{d-1} \cap C_k) *$
$\vdots$
$P(f_{d-1} \mid C_k) * P(C_k)$

∴ Above expression is extremly hard to compute.

To simplify the computation lets take a NAIVE assumption that:
Features are conditionally independent of each other.
ie if $f_1, f_2, f_3 \dots f_{d-1}$ are independent of each other,

$$P(f_1 \mid f_2 \cap f_3 \dots \cap f_{d-1} \cap C_k) = P(f_1 \mid C_k)$$

$$\rightarrow P(f_1 \mid C_k) * P(f_2 \mid C_k) * \dots * P(f_{d-1} \mid C_k) * P(C_k)$$

$$\rightarrow P(C_k) * \prod_{i=1}^{d-1} P(f_i \mid C_k)$$

$$\hat{y}_q = \underset{J \in \{1,2 \dots k\}}{\arg\max} \left\{ P(X_q \mid C_J)\, P(C_J) \right\}$$

$$\downarrow$$

$$P(C_J) * \prod_{i=1}^{d-1} P(f_i \mid C_J)$$

ie $$\hat{y}_q = \underset{J \in \{1,2,\dots k\}}{\arg\max} \left\{ P(C_J) * \prod_{i=1}^{d-1} P(f_i \mid C_J) \right\}$$

Flavours of Naive Bayes

1. Gaussian Naive Bayes

   If data contains numerical features use GaussianNB.

   Assumptions:
   - (a) X contains features which are conditionally independent
   - (b) All the numerical features follows Normal Distribution

$$P(f_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} * \exp\left\{ -\frac{(f_i - \mu_y)^2}{2\sigma_y^2} \right\}$$

2. Multinomial Naive Bayes

   If data contains text use MultinomialNB.

3. Bernoulli Naive Bayes

If data contains categorical features use Bernoulli NB.

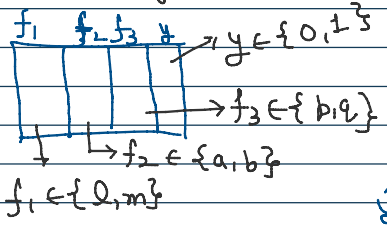
Issues with Naive Bayes :

1. Numerical Instability (For all flavours of NB)

   * Coming Soon * → Log transformation

2. Zero probability issue. (For MultinomialNB & BernoulliNB)

In x_q, if a feature contains a value that was never observed in the historical data, we will encounter a zero probability issue.



$x_q \rightarrow f_1 = m$, $f_2 = c$, $f_3 = p$

Acc^n to NB :

$$\hat{y}_q = \underset{i \in \{1,2 \ldots K\}}{\arg\max} \left\{ P(C_i) * \prod_{j=1}^{d-1} P(f_j | C_i) \right\}$$

i.e Find $P(C_1 = 0 \mid x_q) \propto P(C_1=0) * P(f_1|C_1) * P(f_2|C_1) * P(f_3|C_1)$

& $P(C_2 = 1 \mid x_q) \propto P(C_2=1) * P(f_1|C_2) * P(f_2|C_2) * P(f_3|C_2)$

Observe that → $P(f_2 \mid C_1) = 0$

& $P(f_2 \mid C_2) = 0$

which means → $P(C_1=0 \mid x_q) = 0$

& $P(C_2=1 \mid x_q) = 0$

Now, how do we decide the class of x_q ?

Solution → Apply regularization technique.

{ Here we will use → Laplace Smoothing }

ie Whenever the $P(f_j | C_i) = 0$, apply a smoothening technique so that the value is not exactly equal to 0.

Laplace Smoothing $\Rightarrow \dfrac{P(f_j \mid C_i) + \alpha}{n + k * \alpha}$

Here, α = Hyperparameter

$\alpha \in [0, +\infty]$

$n = P(C_i)$

k = # of unique values in f_j