Naive Bayes Classifier

A simple yet powerful probabilistic model that predicts class membership based on Bayes' theorem, assuming features are independent of each other.

19 OCT, 2025

BAYESIANBRAT*

Naive Bayes

- Naive Bayes classifier is a simple supervised learning model.

  Supervised learning means the model is trained on labeled data (i.e, input features along with their correct output, class).

  Naive Bayes is called Naive because it assumes that all features are independent of each other.

- Example: If we are predicting whether a patient is sick or healthy based on attributes like "fever", "cough", and "headache", Naive Bayes assumes these symptoms are independent of each other given the class (though in reality, they may be co-related).

- Given values of predictor attributes $x = (x_1, x_2, \ldots, x_m)$ predict the class of target attribute Y.

- Y is one of the classes $y \in \{$improved, stable, degraded$\}$ or $y \in \{0, 1\}$

- Classifiers commonly assign to most likely class given a data instance = maximum a posteriori class.

- Given instance data x, assign to class y where $P(Y=y \mid X=x)$ is maximum.
  → Mathematically, we calculate the probability of each class given the features, i.e., $P(Y=y \mid X=x)$
  → Then, we choose the class with the highest probability.
- That is, $\hat{y} = \arg\max_y P(Y=y \mid X=x)$
- Assume that each X_m is categorical, ~~sch~~ such as $x_1 \in \{dry, wet, snow\}$ and $x_2 \in \{medium, rough, smooth\}$
- Direct estimate of $P(Y=y \mid X=x) = P(Y=y \mid X_1=dry, X_2=medium)$ can be calculated if predictor count M is small.

Table : Example weather and surface

Count of delay Weather	surface	delay major	delay minor	delay none
dry	medium	3	6	1
dry	rough	0	3	0
dry	smooth	0	2	1
snow	medium	0	2	0
snow	rough	0	0	0
snow	smooth	0	0	0
wet	medium	1	0	0
wet	rough	1	0	0
wet	smooth	0	0	0
Grand	Total	5	13	2

W = Weather S = surface

Assign to class with maximum class probability estimates.

$P(Y = \text{major} \mid W = \text{dry}, S = \text{medium}) = 3/10$

$P(Y = \text{minor} \mid W = \text{dry}, S = \text{medium}) = 6/10$

$P(Y = \text{none} \mid W = \text{dry}, S = \text{medium}) = 1/10$

However ...

Suppose that each predictor x_m has 4 possible values and $M = 30$

The count of possible unique indicator values is $4^{30} = 2^{60} = (2^{10})^6 \approx (10^3)^6 = 10^{18}$

Try a different approach : Bayes Theorem

$P(Y = \text{major} \mid W = \text{dry}, S = \text{medium}) =$

$$\frac{P(W = \text{dry}, S = \text{medium} \mid Y = \text{major})\, P(Y = \text{major})}{P(W = \text{dry}, S = \text{medium})}$$

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)\, P(Y = y)}{P(X = x)}$$

Assign istance to class with maximum class probability estimate computed Bayes Theorem

$$\hat{Y} = \text{argmax}_y\, P(X = x \mid Y = y)\, P(Y = y)$$

Why is $P(X = x)$ dropped?

- Make a strong assumption: X_m's are conditionaly independent given $Y = y$ so that

$P(W = \text{dry}, S = \text{medium} \mid Y = \text{major}) =$

$P(W = \text{dry} \mid Y = \text{major}) \times$ ~~P(S = major~~
$P(S = \text{medium} \mid Y = \text{major})$

$P(X = x \mid Y = y) = P(X_1 = x_1 \mid Y = y) \times$ ~~$P(X_2 = x_2$~~
$P(X_2 = x_2 \mid Y = y) \times$
$P(X_3 = x_3 \mid Y = y)$

- Each conditional probability $P(X_M = x_m \mid Y=y)$ is easy to estimate from data as

$P(W=dry \mid Y=major) \approx 3/5$
$P(W=wet \mid Y=major) \approx 2/5$
$P(W=snow \mid Y=major) = ?$

$P(W=dry \mid Y=minor) = 11/13$
$P(W=wet \mid Y=minor) = 0/13$
$P(W=snow \mid Y=minor) = 2/13$

$P(W=dry \mid Y=none) = 2/2$
$P(W=Wet \mid Y=none) = ?$
$W(W=snow \mid Y=none) = ?$

- $P(Y=y)$ estimates is the proportion of data ~~ta~~ instances with $Y=y$

$$P(X_1 = x_1 \mid Y=y) \ldots P(X_M = x_m \mid Y=y) \underline{P(Y=y)}$$

Underline a reminder to include the a priori probability that $Y=y$

$P(Y=major \mid W=wet, S=medium) \propto$
$P(W=wet \mid Y=major) \times P(S=medium \mid Y=major)$
$\times P(Y=major)$

$$= \frac{2}{5} \times \frac{4}{5} \times \frac{5}{20}$$

$$= \frac{40}{500}$$

$P(Y = \text{minor} \mid W = \text{wet}, S = \text{medium}) \propto$
$P(W = \text{wet} \mid Y = \text{minor}) \times P(S = \text{medium} \mid Y = \text{minor}) \times P(Y = \text{minor})$

$$= \frac{0}{13} \times \frac{8}{13} \times \frac{13}{20}$$

$$= 0$$

$P(Y = \text{none} \mid W = \text{wet}, S = \text{medium}) \propto$ ~~P(W~~
$P(W = \text{wet} \mid Y = \text{none})\, P(S = \text{medium} \mid Y = \text{none})\, P(Y = \text{none})$
$= 0$

Assign to class $\hat{y} = \text{major}$

→ The problem here is, when we have a single predictor that does the whole probability zero, puts a huge dent in the calculation.

→ There's a need of some smoothing ~~tech~~ technique

Applying law of total probability and, using denominator.

$$P(X = x) = \sum_{y'} P(X = x \mid Y = y')\, P(Y = y')$$

$$P(Y = \text{major} \mid X) = \frac{0.08}{P(X = x)} = \frac{0.08}{\sum_{y'} P(X = x \mid Y = y')\, P(Y = y')}$$

$P(Y = \text{minor} \mid X) = 0$
$P(Y = \text{none} \mid X) = 0$

$$= \frac{0.08}{0.08} = 1$$

Laplase Smoothing:

-) What happens if the estimate of $P(X_m = x_m \mid Y = y) = 0$? More likely with many predictors.

-) Software commonly adjusts with Laplace smoothing. Suppose that X has V unique values $x_1, x_2, \ldots, x_V$.

-) Laplace smoothing adds α to each count $N(X = x_V)$ to remove zero counts.

-) In simpler terms:

If one of these probabilities $= 0$, we then multiply

$$P(X_1 \mid Y) \cdot P(X_2 \mid Y) \cdots P(Y)$$

The whole product becomes zero.

That class can never be chosen even if the class is actually possible.

This is called the zero frequency problem.

To fix this, we pretend we have seen at each possible category at least once.

We add small constant α (often 1) to each count

Then we re-normalize

Suppose feature X has V possible values $(x_1, \ldots, x_V)$

$$P(X = x_v \mid Y = y) = \frac{N(X = x_v, Y = y)}{N(Y = y)}$$

With laplace smoothing:

$$P(X = x_v \mid Y = y) = \frac{N(X = x_v, Y = y) + \alpha}{N(Y = y) + \alpha V}$$

- Numerator: We add α
- Denominator: We add αV (because we added α to each V possible values).

Example:

For class = "none"

Without smoothing: $P(W = \text{wet} \mid Y = \text{none})$

$$= \frac{0}{2}$$

With Laplace ($\alpha = 1$, and Weather it has 3 categories: dry, wet, snow)

$$P(W = \text{wet} \mid Y = \text{none}) = \frac{0+1}{2+3}$$

$$= \frac{1}{5} = 0.2$$

Smoothing prevents classes from being completely wiped out.

Especially important when you have many features and categories, since some combinations will be missing in training data.

Example:

1. Without smoothing (raw counts)

Imagine a categorical feature X with 3 possible values

a, b, c

Suppose for class $Y=y$ in training data we saw

$N(X=a, Y=y) = 2$
$N(X=b, Y=y) = 0$
$N(X=c, Y=y) = 1$
Total $N(Y=y) = 3$

$P(a|y) = \frac{2}{3} = 0.667$

$P(b|y) = \frac{0}{3} = 0$

$P(c|y) = \frac{1}{3} = 0.333$

2. Doing a Laplace smoothing:

$$P(X = x_v \mid Y = y) = \frac{N(X = x_v, Y = y) + \alpha}{N(Y = y) + \alpha V}$$

Here $V = 3$, $N(Y = y) = 3$ $\alpha = 1$

$a = \frac{2+1}{3+3} = \frac{3}{6} = 0.5$

$b : \frac{0+1}{3+3} = \frac{1}{6} = 0.167$

$c : \frac{1+1}{3+3} = \frac{2}{6} = 0.333$

Not all probabilities are nonzero and still sum to 1.