

MACHINE LEARNING WITH PYTHON

NAÏVE BAYES

Themistoklis Diamantopoulos

Bayes Theorem

- Equation created by Thomas Bayes in 1763:

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

where A and B are events and $P(B) \neq 0$

$P(A|B)$: likelihood of event A occurring given that B is true

$P(B|A)$: likelihood of event B occurring given that A is true

Categorical Problem

- Decide whether the traffic is going to be high based on the weather and the day

Weather	Day	HighTraffic
Hot	Work	No
Cold	Vacation	No
Hot	Vacation	Yes
Hot	Work	Yes
Hot	Work	Yes
Cold	Vacation	No
Cold	Vacation	Yes

Naïve Bayes

- Independent features
- Bayes Theorem

$$P(c | x) = \frac{P(x_1 | c) \cdot P(x_2 | c) \cdot \dots \cdot P(x_n | c) \cdot P(c)}{P(x_1) \cdot P(x_2) \cdot \dots \cdot P(x_n)}$$

Weather	Day	HighTraffic
Hot	Work	No
Cold	Vacation	No
Hot	Vacation	Yes
Hot	Work	Yes
Hot	Work	Yes
Cold	Vacation	No
Cold	Vacation	Yes

$$P(\text{Yes} | \text{Hot}, \text{Vacation}) = \frac{P(\text{Hot} | \text{Yes}) \cdot P(\text{Vacation} | \text{Yes}) \cdot P(\text{Yes})}{P(\text{Hot}) \cdot P(\text{Vacation})} = \frac{3/4 \cdot 2/4 \cdot 4/7}{4/7 \cdot 4/7} = 21/32 = 0.65625$$



$$P(\text{No} | \text{Hot}, \text{Vacation}) = \frac{P(\text{Hot} | \text{No}) \cdot P(\text{Vacation} | \text{No}) \cdot P(\text{No})}{P(\text{Hot}) \cdot P(\text{Vacation})} = \frac{1/3 \cdot 2/3 \cdot 3/7}{3/7 \cdot 3/7} = 14/27 = 0.51852$$

When weather is Hot and day is Vacation, traffic is High (prob: $0.65/(0.65+0.51) = 0.56$)

Classification Evaluation

- Confusion Matrix

A confusion matrix diagram. The vertical axis is labeled 'Predicted Class' and has two categories: 'Positive' and 'Negative'. The horizontal axis is labeled 'Actual Class' and has two categories: 'Positive' and 'Negative'. A diagonal line separates the two axes. The matrix cells contain: TP (True Positive) at the intersection of Predicted Positive and Actual Positive; FP (False Positive) at the intersection of Predicted Positive and Actual Negative; FN (False Negative) at the intersection of Predicted Negative and Actual Positive; and TN (True Negative) at the intersection of Predicted Negative and Actual Negative. Below the matrix, the formulas $P = TP + FN$ and $N = FP + TN$ are shown.

Predicted Class \ Actual Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN

$P = TP + FN$ $N = FP + TN$

- Evaluation Metrics

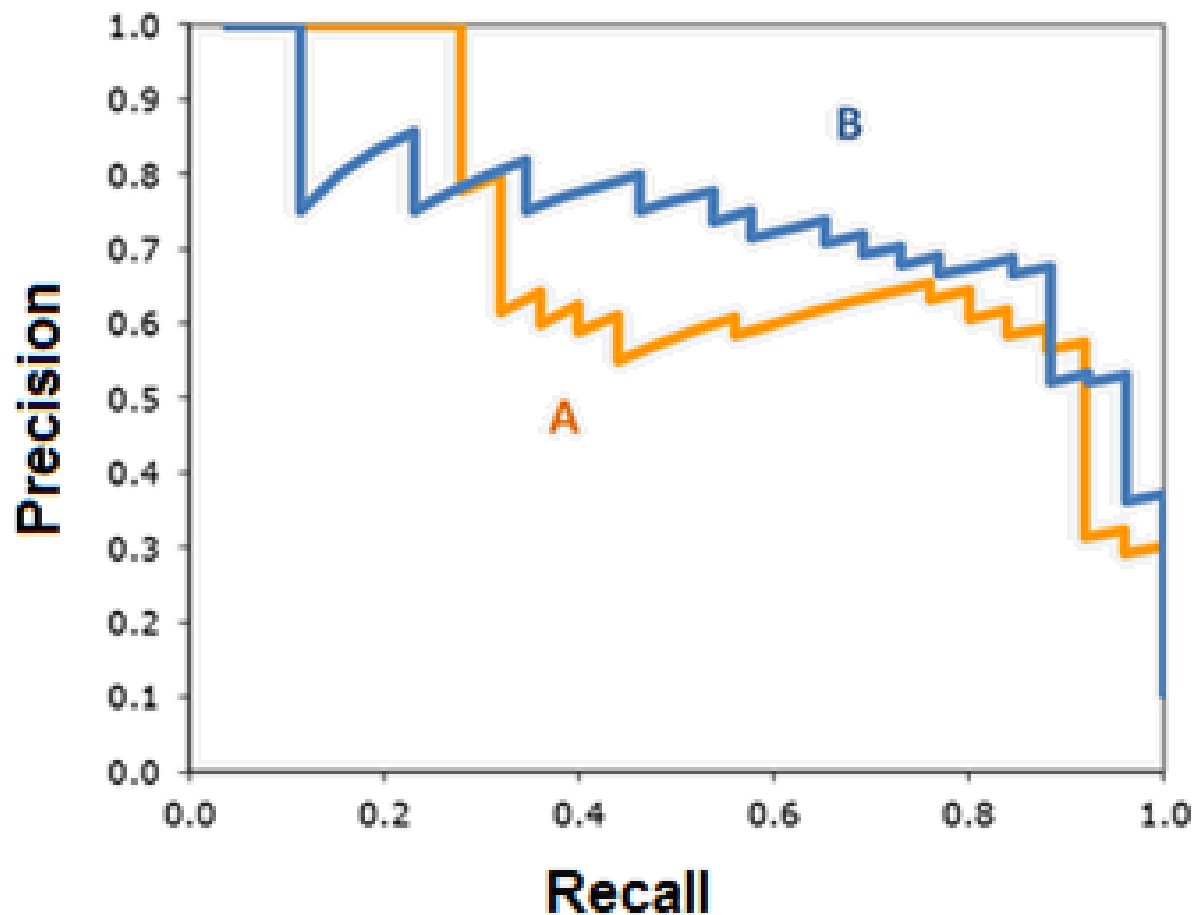
- Accuracy = $(TP + TN) / (P + N)$
- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$

Precision and Recall

- Tradeoff between Precision & Recall

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$



ROC Curve

- True Positive Rate (also known as sensitivity or recall)

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate (also known as specificity)

$$FPR = \frac{FP}{FP + TN}$$

- AUC (Area Under the Curve)

