MACHINE LEARNING WITH PYTHON

NAÏVE BAYES

Themistoklis Diamantopoulos

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

$$P(c \mid x) = \frac{P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot \ldots \cdot P(x_n \mid c) \cdot P(c)}{P(x_1) \cdot P(x_2) \cdot \ldots \cdot P(x_n)}$$

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

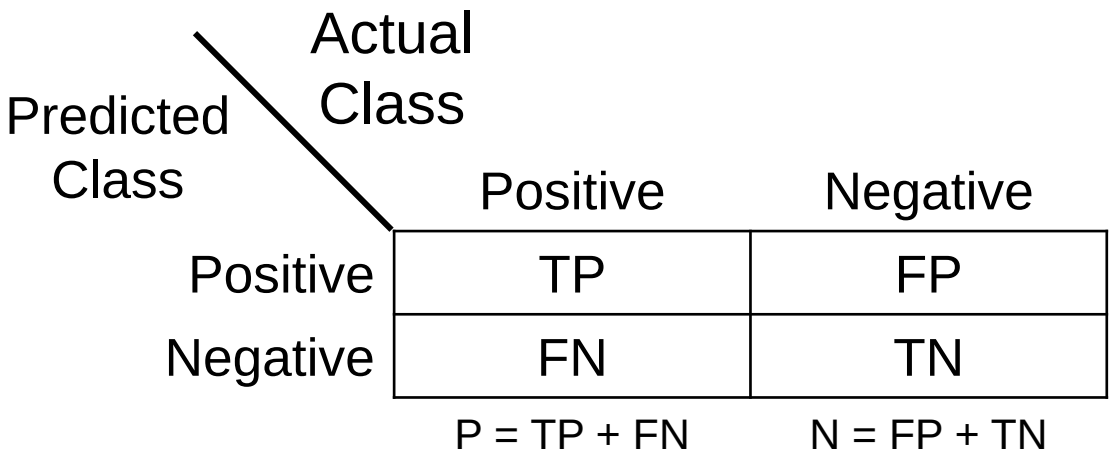
$$P(Yes \mid Hot, Vacation) = \frac{P(Hot \mid Yes) \cdot P(Vacation \mid Yes) \cdot P(Yes)}{P(Hot) \cdot P(Vacation)} = \frac{3/4 \cdot 2/4 \cdot 4/7}{4/7 \cdot 4/7} = 21/32 = 0.65625$$

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

When the weather is Hot and the day is Vacation, the traffic is High (prob: 0.56)

Classification Evaluation

- Confusion Matrix



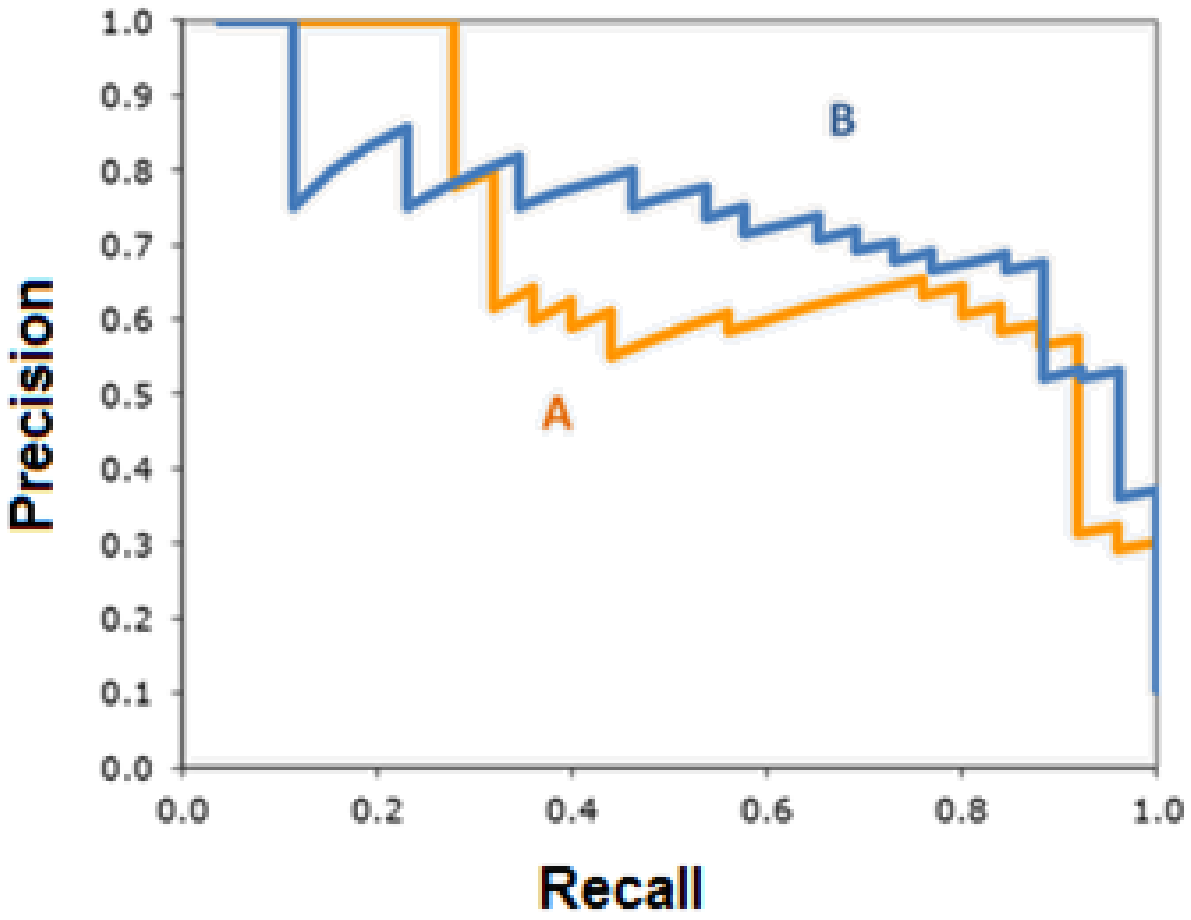
- Evaluation Metrics
  - Accuracy = (TP + TN) / (P + N)
  - Precision = TP / (TP + FP)
  - Recall = TP / (TP + FN)

Precision and Recall

- Tradeoff between Precision & Recall

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

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



Source: https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

ROC Curve

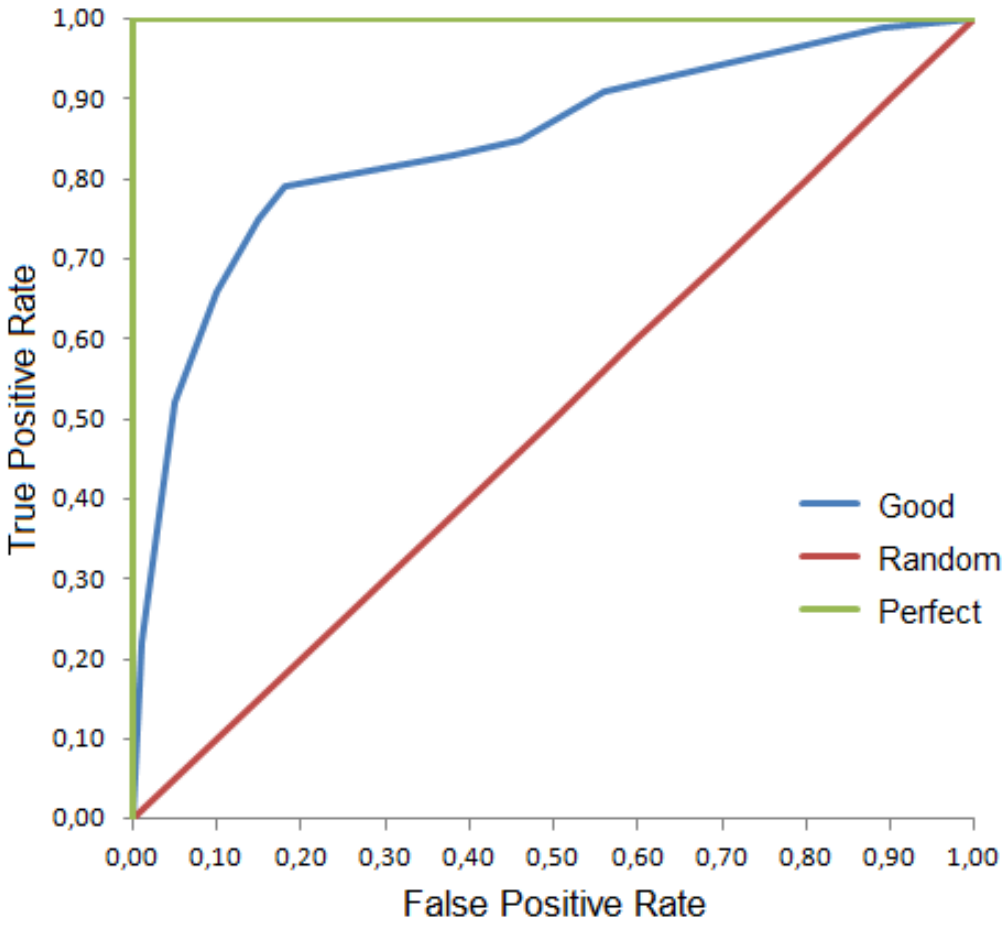
- True Positive Rate (also known as sensitivity or recall)

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

- False Positive Rate (also known as specificity)

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

- AUC (Area Under the Curve)



Source: https://en.wikipedia.org/wiki/Receiver_operating_characteristic