Classifiers: Exercises

The below questions all reference this hypothetical data set:

	Weather	**Traffic**	**Accident**
1	Rainy	Light	Yes
2	Rainy	Light	No
3	Rainy	Heavy	Yes
4	Sunny	Light	No
5	Sunny	Heavy	No
6	Sunny	Heavy	No
7	Sunny	Heavy	Yes
8	Rainy	Light	No
9	Rainy	Light	Yes
10	Rainy	Heavy	Yes

1. Suppose we build a classifier that predicts the following values for `Accident`:

Sample #	1	2	3	4	5	6	7	8	9	10
Predicted	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes
Actual	Yes	No	Yes	No	No	No	Yes	No	Yes	Yes

   Fill out the below confusion matrix.

	Predicted Negative	Predicted Positive
Actual Negative		
Actual Positive		

2. Calculate the accuracy, F_1 score, and φ coefficient for the above data set. Are any of these values potentially misleading?
3. Using Naive Bayes, predict the value of `Accident` for a data point $(\text{Weather} = \text{Rainy}, \text{Traffic} = \text{Heavy})$.

Reference

Classifier Accuracy

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{Total}}$$

$$F_1 = \frac{2 * \text{True positives}}{2 * \text{True positives} + \text{False positives} + \text{False negatives}}$$

$$\varphi = \text{MCC} = \frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Bayes' Theorem

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

Naive Bayes

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y) * p(y)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|y) * p(y) \tag{1}$$

$$p(y) = \frac{\text{\# of } y}{\text{Total}} = \frac{|y|}{|X|} \tag{2}$$

$$\begin{aligned} p(\boldsymbol{x}|y) &= p((a_1, a_2, ..., a_n)|y) \\ &= p(a_1|y) * p(a_2|y) * ... * p(a_n|y) \\ &= \prod_{i=1}^{n} p(a_i|y) \end{aligned} \tag{3}$$

To predict y for a data point $\boldsymbol{x} = (a_1, a_2, ..., a_n)$, calculate $p(y|\boldsymbol{x})$ for each possible y and choose the one that is the largest:

$$\begin{aligned} \text{Predicted } y = \hat{y} &= \max_y p(y|\boldsymbol{x}) \\ &= \max_y p(y) * p(\boldsymbol{x}|y) && \text{By Equation 1} \\ &= \max_y \frac{|y|}{|X|} * p(\boldsymbol{x}|y) && \text{By Equation 2} \\ &= \max_y \frac{|y|}{|X|} * \prod_{i=1}^{n} p(a_i|y) && \text{By Equation 3} \end{aligned}$$

$$p(a_i|y) = \frac{\text{\# of data points with attribute } a_i \text{ and } y}{\text{Total number of data points with } y}$$

$$|X| = \text{Size of data set}$$