AI Academy: Introduction to Data Mining
Week 5 Seminar

Seminar 5 contains 2 questions.

1. Consider two models, A and B, for the classification of 100 patients. ” + ” indicates positive while ”−” label indicates negative. Both models have the accuracy of 80%. Model A reports 30 positive patients correctly but incorrectly predicted 10 positive patients as negative. Model B reports 25 positive patients correctly but incorrectly predicted 5 positive patients as negative.

(a) (2 points) Fill in the Confusion Matrix for BOTH Model A and Model B.
SOLUTION:

Model A		*Predicted*	
		+	-
	+	30	10
Actual			
	-	10	50

Model B		*Predicted*	
		+	-
	+	25	5
Actual			
	-	15	55

Model A		*Predicted*	
		+	-
	+		
Actual			
	-		

Model B		*Predicted*	
		+	-
	+		
Actual			
	-		

(b) (2 points) Compute **precision** and **recall** for Model A and Model B respectively.
SOLUTION:

For **model A**:
precision = 30/(30+10) = 0.75
recall = 30/(30+10) = 0.75
f1 = 2(.75)(.75)/(.75+.75) = .75
For **model B**:
precision = 25/(25+15) = 0.625
recall = 25/(25+5) = 0.83
f1 = 2*.625*.83/(.625+.83) = .71

(c) (2 points) Produce a **cost (earning)** matrix assuming that: a true positive *earns* \$50; a true negative *earns* \$10; a false positive *loses* \$10; and a *false* negative *loses* \$100.
SOLUTION:

$ Value		*Predicted*	
		+	-
	+	50	-100
Actual			
	-	-10	10

$ Value		*Predicted*	
		+	-
Actual	+		
	-		

(d) (2 points) Given the cost (earning) matrix above, compute **the expected earning** for models A and B respectively. **Which** model is better w.r.t. the expected earning?

SOLUTION:

Expected rewards for **model A**:
30*50+10*(-100)+10*(-10)+50*10 = 900

Expected rewards for **model B**:
25*50+5*(-100)+15*(-10)+55*10 = 1150

According to the expected rewards of two models, **model B** is better, since it has higher expected earning.

2 Decision Stumps, & Cross Validation

Consider the following dataset (9 instances) with **2 binary attributes** (x_1 and x_2), and a **class attribute** y, shown in Table 1. For this question, we will consider a **Decision Stump** classifier. A Decision Stump is a decision tree with a max depth of 1 (i.e. only one split before classification).

Table 1: Data

ID	x1	x2	Class
1	True	False	+
2	False	False	-
3	True	True	-
4	False	False	+
5	True	True	-
6	False	True	-
7	False	False	+
8	False	True	+
9	True	False	-

1. By hand, evaluate the Decision Stump classifier, calculating the confusion matrix and testing accuracy (show your work by labeling each data object with the predicted class). You should be able to eyeball which split is best, but if you can't, use the GINI index. If the two splits end up being identical, go with x_1. If you end up with a split where one class has an even distribution, default to a class assignment of positive.

   Use the following evaluation methods:

   (a) A holdout test dataset consisting of last 4 instances

   (b) 3-fold cross-validation, using the following folds with IDs: [1,2,3], [4,5,6], [7,8,9] respectively.

   (c) Leave one out cross validation (LOOCV)

1. (a) By considering the first 5 instances, the best split would be splitting on x_2. Where if $x_2 = F, C = +$ and $x_2 = T, C = -$.

	Predicted (+)	Predicted (-)
Actual (+)	1	1
Actual (-)	1	1

   Accuracy = 2/4

   (b) i. Training on [4-9], we get that the bets split would be x_1 is the best split, and if $x_1 = F, C = +$ and $x_1 = T, C = -$.

   We now evaluate on [1,2,3]

	Predicted (+)	Predicted (-)
Actual (+)	0	1
Actual (-)	1	1

Accuracy = 1/3

ii. Training on [1,2,3] and [7,8,9], we get that x_1 is the best split, with $x_1 = F, C = +$ and $x_1 = T, C = -$.
We now evaluate on [4,5,6]

	Predicted (+)	Predicted (-)
Actual (+)	1	0
Actual (-)	1	1

Accuracy = 2/3

iii. Training on [1-6], we get that x_2 is the best split with $x_2 = F, C = +$ and $x_1 = T, C = -$.
We now evaluate on [7,8,9]

	Predicted (+)	Predicted (-)
Actual (+)	1	1
Actual (-)	1	0

Accuracy = 1/3
We can take the total average across all of the three folds, which is 4/9.
The total confusion matrix would be:

	Predicted (+)	Predicted (-)
Actual (+)	2	2
Actual (-)	3	2

(c) LOOCV

ID	Best Split	Predicted
1	x_1	-
2	x_2	+
3	x_2	-
4	x_1	+
5	x_1	-
6	x_1	-
7	x_1	+
8	x_2	-
9	x_2	+

	Predicted (+)	Predicted (-)
Actual (+)	3	3
Actual (-)	1	2

Accuracy = 5/9