Midterm - Dry Trial

Wednesday, April 29, 2020 4:46 PM

1. We are solving a binary classification problem. Two models provide the following results on test data:

actual labels	-	-	+	-	-	-	+	-	-	-	+	-
Model 1	0.2	0.1	0.49	0.26	0.3	0.12	0.31	0.2	0.1	0.32	0.4	0.2
Model 2	0.6	0.51	0.8	0.12	0.54	0.39	0.53	0.46	0.41	0.37	0.49	0.28

The first row contains the actual labels of test data, and the following rows contain the confidence scores (probabilities of positive class). Each row represents a model.

   a. **Choose threshold to be 0.5**. In other words, any test instances whose confidence value is greater than 0.5 will be classified as positive. **Determine the confusion matrix of Model 1!**
   b. Using assumption of part a. determine the **accuracy of Model 1**!
   c. In general, what are the **weaknesses of the accuracy measure**? List three of these weaknesses, and give solutions for them (another measure, modification of accuracy, another procedure).
   d. Construct the **ROC** curve **for both models**! (Note: It is not necessary to fill in the usual table, drawing the curves is enough.)
   e. Calculate **AUC** scores **for both model**!
   f. In general, what is the **probabilistic interpretation of AUC**?
   (25%)

1.a. First we have to sort the test instances according to their confidence value in ascending order.

actual labels — — — — — — — — + — + +

conf score 0.1 0.1 0.12 0.2 0.2 0.2 0.26 0.3 0.31 0.32 0.4 0.49

If threshold = 0.5 => model predicts everything to be negative

Confusion matrix

		Predicted class: P	Predicted class: N
Actual class	P	TP: 0	FN: 3
	N	FP: 0	TN: 9

b. $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{0+9}{0+0+3+9} = \frac{9}{12} = \frac{3}{4}$

c. Weaknesses:

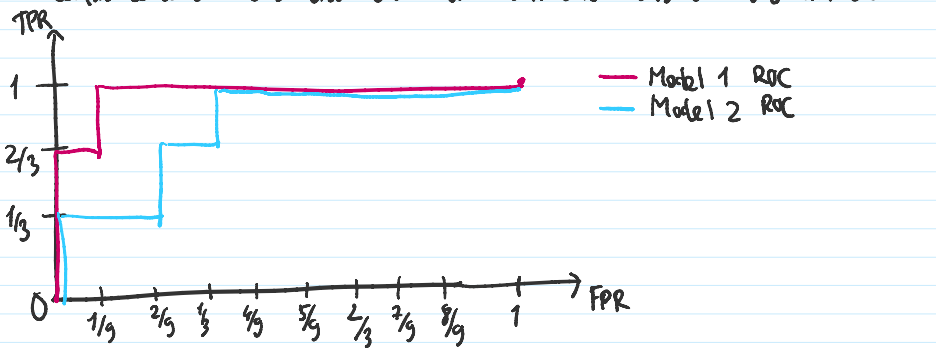
1. Bad for unbalanced classes:
   Fix: use Recall or F-1
2. Does not take in account the cost of false negative vs false positive
   Fix: give weights to them in the calculation
3. Doesn't work super well when you only care about the model's performance in making positive predictions.
   Fix: use Precision

d. Do the same thing in part (a) for Model 2.

Model 2:

actual labels — — — — — — + — + — — +

confidence score 0.12 0.28 0.37 0.39 0.41 0.46 0.49 0.51 0.53 0.54 0.6 0.8



e. Model 1 AUC $= \frac{8}{9} \times 1 + \frac{1}{9} \times \frac{2}{3} = \frac{26}{27}$

Model 2 AUC $= \frac{2}{3} \times 1 + \frac{1}{9} \times \frac{2}{3} + \frac{2}{9} \times \frac{1}{3} = \frac{22}{27}$

f. It is the probability that a randomly-chosen positive record is ranked more highly than a randomly chosen negative record.

2. Calculate the **Jaccard index**, the **Simple Matching Coefficient (SMC)**, the **cosine similarity** and the **Hamming distance** between the following two vectors:

$$p = (1, 0, 0, 1, 1, 0, 0, 0, 1, 0)$$
$$q = (1, 0, 0, 0, 1, 0, 1, 0, 0, 0)$$

Name a situation when Jaccard index defines a more reasonable similarity than SMC!
(15%)

Name a situation when Jaccard index defines a more reasonable similarity than SMC!
(15%)

2.

q \ p	0	1
0	$M_{00}: 5$	$M_{10}: 2$
1	$M_{01}: 1$	$M_{11}: 2$

$p \cdot q = 1 \cdot 1 + 1 \cdot 1 = 2$

$\|p\| = \sqrt{1+1+1+1} = \sqrt{4} = 2$

$\|q\| = \sqrt{1+1+1} = \sqrt{3}$

Jaccard index: $\frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{2}{1+2+2} = \frac{2}{5}$

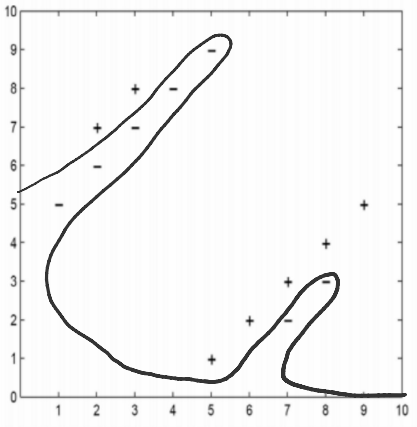
SMC: $\frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{10} + M_{01}} = \frac{5+2}{5+2+2+1} = \frac{7}{10}$

cosine similarity: $\frac{p \cdot q}{\|p\| \cdot \|q\|} = \frac{2}{2 \cdot \sqrt{3}} = \frac{1}{\sqrt{3}}$

Hamming distance: $M_{01} + M_{10} = 1+2 = 3$

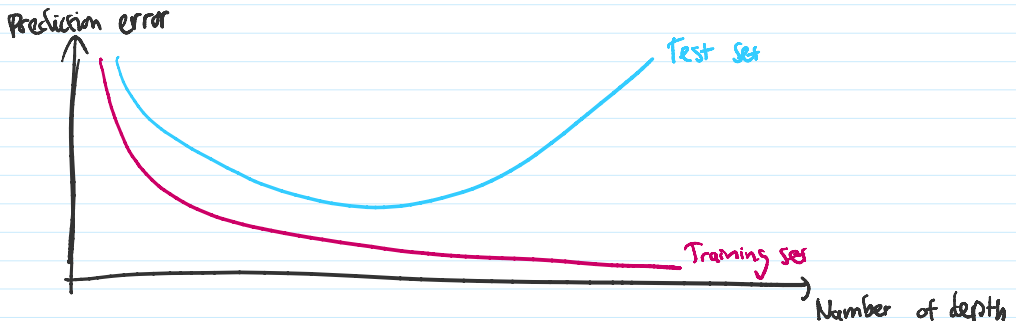
Jaccard index is more reasonable if the common 0s between the two vectors do not carry a role. e.g. two documents do not have the same word does not make them more similar.

3. We consider **a *k*-nearest neighbor classifier** using Euclidean distance metric on a binary classification task.
   a. In the figure, sketch the 1-nearest neighbor **decision boundary** for this dataset!
   b. If you try to classify the entire **training dataset** using a kNN classifier, what value of *k* will **minimize the error** for this dataset? What is the resulting **training error** (error rate on the training set)?
   (10%)



3.b. k=1 will minimize the error for the dataset. The training error will be 0.

4. Draw a graph, which generally depicts **error rate on training test and test set** as function of number of leaves (or depth) of decision tree! What phenomena occur in case of too few leaves (too shallow tree) or too much leaves (too deep tree)?
   (10%)



If the tree is too shallow, our model is underfitting.
If the tree is too deep, our model is overfitting.

5. A restaurant plans to launch a promotion and asked some people if they were interested. In addition, they recorded three features about each person. We use **naïve-Bayes** method to determine who would be interested in the promotion.
   a. Using naïve-Bayes method, decide if an underweight, elderly, sporty person was interested in the promotion!
   b. What is the **core assumption** of the naïve-Bayes classifier?
   c. What is **Laplace-smoothing**? Why can we think that it is a better method than the original estimator?
   (15%)

#	*Age*	*Weight*	*Sporty*	*Interested in promotion*
1	young	Underweight	Yes	Yes
2	elderly	Ideal	No	No
3	middle-aged	Overweight	No	Yes
4	elderly	Ideal	Yes	No
5	young	Overweight	No	Yes
6	middle-aged	Underweight	No	No
7	elderly	Underweight	No	No
8	young	Ideal	No	Yes
9	middle-aged	Overweight	Yes	Yes
10	elderly	Ideal	Yes	No

5.a. $x = (\text{underweight}, \text{elderly}, \text{sporty})$

$P(\text{Yes}) = \frac{5}{10} = \frac{1}{2}$ $\quad P(\text{No}) = \frac{5}{10} = \frac{1}{2}$

With Laplace smoothing

$P(x \mid \text{Yes}) = P(\text{elderly} \mid \text{yes}) \cdot P(\text{underweight} \mid \text{yes}) \cdot P(\text{sporty} \mid \text{yes})$

$= \frac{1}{7} \cdot \frac{2}{7} \cdot \frac{3}{7} = \frac{6}{343}$

$P(x \mid \text{No}) = P(\text{elderly} \mid \text{No}) \cdot P(\text{underweight} \mid \text{No}) \cdot P(\text{sporty} \mid \text{No})$

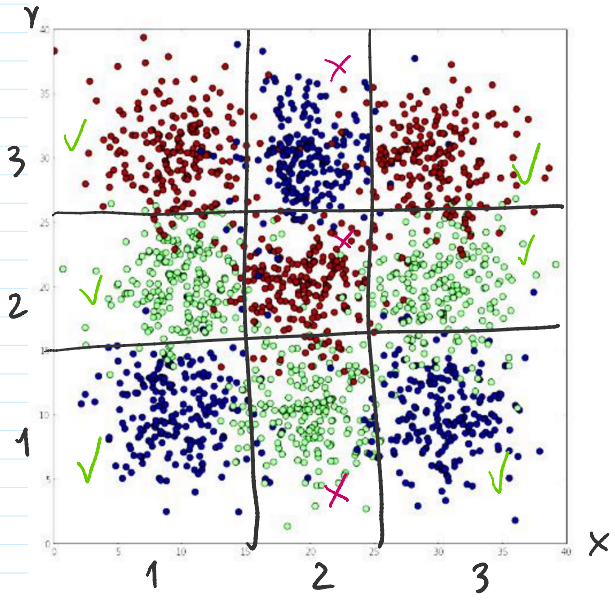
$= \frac{5}{7} \cdot \frac{3}{7} \cdot \frac{3}{7} = \frac{45}{343}$

Since $P(X \mid \text{No}) \cdot P(\text{No}) > P(X \mid \text{Yes}) \cdot P(\text{Yes})$, $P(\text{No} \mid X) > P(\text{Yes} \mid X)$. Hence, the predicted label is No.

b. The core assumption is the features are conditionally independent from each other given the class variable.

c. Laplace smoothing is when we add 1 to the numerator and add the number of class labels to the denominator when calculating the conditional probabilities. We do this to avoid having any conditional probabilities turn out to be 0. It is better since just by having 1 conditional probability to be 0 can make the whole expression to become 0 despite all the other features. Hence, we are allowing 1 single feature dictates our whole prediction. Using Laplace Smoothing will help avoid this as we only going to be given very little weight to that conditional probability that was previously 0.

6. We have a two-dimensional data set illustrated in the figure below. Colors represent actual class labels (red, blue, green). How would the following algorithms perform on this data: **kNN, Decision Tree, naïve-Bayes**? Outline the steps and results of each algorithm. Which algorithm would misclassify certain parts of the two-dimensional data? Which parts? In case of naïve-Bayes discretize the data (make three intervals: 0<=x<15; 15<=x<25; 25<=x<40 and perform the same discretization for y)!
(25%)



1. kNN would perform generally well on this data. The labels form cluster with high density, so kNN would be able to make pretty good predictions based on the surrounding neighbors of a record. kNN would have the most trouble classifying records at the edge of each cluster where the density is not as high.

2. Decision tree will divide the space using rectilinear boundaries around each clusters. I suspect it will perform reasonably well, but worse than kNN due to the fact that the clusters are not rectangular in nature. It will have errors near the edge and corners of each clusters.

3. Assuming each cluster has a similar amount of records, we have:
$P(Red) = P(Green) = P(Blue) = \frac{1}{3}$

We have the following conditional probabilities:

$P(X=1 \mid Red) = 1/3$	$P(X=1 \mid Blue) = 1/3$	$P(X=1 \mid Green) = \frac{1}{3}$
$P(X=2 \mid Red) = 1/3$	$P(X=2 \mid Blue) = \frac{1}{3}$	$P(X=2 \mid Green) = \frac{1}{3}$
$P(X=3 \mid Red) = 1/3$	$P(X=3 \mid Blue) = \frac{1}{3}$	$P(X=3 \mid Green) = \frac{1}{3}$
$P(Y=1 \mid Red) = 0$	$P(Y=1 \mid Blue) = \frac{2}{3}$	$P(Y=1 \mid Green) = 1/3$
$P(Y=2 \mid Red) = 1/3$	$P(Y=2 \mid Blue) = 0$	$P(Y=2 \mid Green) = 2/3$
$P(Y=3 \mid Red) = \frac{2}{3}$	$P(Y=3 \mid Blue) = \frac{1}{3}$	$P(Y=3 \mid Green) = 0$

We can observe that X does not play a role in deciding the label for the record, but only Y. For the (X=1) and (X=3) area, we will identify correctly since the label for those clusters are also the majority label of its respective Y interval.

However, we will make classification error in the whole (X=2) area due to the majority label for each cluster in the (X=2) area does not match the majority label of each respective Y interval.