MTAT.03.227 Machine Learning
Practice session 7
F1-measure & ROC curves

October 28-30, 2019

Exercise 1.

In the following confusion matrix (contingency table), please replace each '??' by one of the following labels: TP, FP, FN, TN, PPos, PNeg, Neg, Pos, N. Try to do this without look at the materials. Note that the goal is not to memorize this table, but to be able to do this by deduction / reasoning.

ACTUAL \ PREDICTED	PREDICTED (+)	PREDICTED (-)	TOTAL
ACTUAL (+)	??	??	??
ACTUAL (-)	??	??	??
TOTAL	??	??	??

Exercise 2.

From the lecture material, we know the following formulas:

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$
$$\text{precision} = \frac{TP}{TP+FP} = \frac{TP}{PPos}$$
$$\text{recall} = \frac{TP}{TP+FN} = \frac{TP}{Pos}$$
$$\text{F1-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Why is F1-measure defined as the harmonic mean of precision and recall and not just as the usual arithmetic mean (average)? To find the answer to this, imagine a dataset with 10 positives and 10 000 negatives (for example, positive could mean 'disease' and negative could mean 'no disease'. Consider the following 2 classifiers:

1. A classifier that predicts all instances to be positive;
2. A classifier that predicts only 10 of the instances as positive, and among them there are 5 actual positives and 5 actual negatives.

Write out the contingency tables for these classifiers. Which of the two classifiers seems more useful? Calculate precision and recall for both classifiers. Which classifier has a higher arithmetic mean of precision and recall? Which has a higher harmonic mean of precision and recall (that is, higher F1-measure)?

Exercise 3.

We have 280 patients, 80 have a disease (positive class) and 200 don't (negative class). Please write out the contingency table in the following 4 cases:

1. A Classifier that predicts only 'no disease';
2. A Classifier that predicts only 'disease';

3. A Classifier that randomly predicts 'disease' with 50% probability;

4. A Classifier that randomly predicts 'disease' with 75% probability.

Exercise 4.

Calculate accuracy, error rate, precision, recall and F1-measure for all 4 classifiers from the previous task and fill in the table:

Classifier	Accuracy	Error rate	Precision	Recall	F1
C1	?	?	?	?	?
C2	?	?	?	?	?
C3	?	?	?	?	?
C4	?	?	?	?	?

They are all random (dumb) classifiers, do not use any features, but accuracy is different between them. Why? Similarly, compare precision, recall and F1-score of these classifiers and discuss the results.

Exercise 5.

Plot the 4 classifiers from the previous task in the ROC space.

Exercise 6.

Consider the following data which lists the predictions of 2 scoring classifiers S1 and S2 on 10 instances, as well as the true class of these instances:

S1	S2	Actual class
3	3	+
3	2	+
3	2	+
1	2	+
3	2	−
1	2	−
1	2	−
1	1	−
1	1	−
1	1	−

Draw a ROC curve for both of these scoring classifiers S1 and S2. What is the AUC of these classifiers?

Exercise 7.

Which scoring threshold returns the highest accuracy on these instances for S1? For S2? Which of the two models achieves a better accuracy? Find the best threshold using two alternative methods:

1. Calculate accuracy at different thresholds and find out the best threshold;

2. Take the line of slope $\frac{Neg}{Pos}$, let it fall down from the ROC heaven and see which point is touched first on each of the ROC curves.

Exercise 8.

Increase the size of the dataset by making 10 more copies of the last instance. What is the best threshold that maximizes the accuracy for S1? For S2? Which of the two models achieves a better accuracy now?

Exercise 9.

Watch a video about ROC https://www.youtube.com/watch?v=OAl6eAyP-yo If you are interested, then the tool used in the video is available here: http://www.navan.name/roc