

Machine Learning Exercises 10



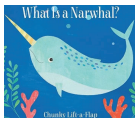



This set of exercises are about getting familiar with performance metrics. Try to do as much as you can by hand, rather than using the computer, so that you get some practice for discussing the concepts at the oral exam.



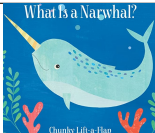
The important things to think about are what each metric actually measures and whether some types of errors are more problematic than others (the latter is naturally extremely application dependent!).

When we specify a loss function/matrix, we exactly consider carefully how to penalise different types of errors. Classifiers can then be compared by the expected loss – this is what we have already often computed as the *test error*. For Bayes classifiers that uses 0-1 loss, the test error is simply the misclassification error (i.e. $1 - \text{Accuracy}$). So long as you are aware of whether the metric captures what you need, there is absolutely nothing wrong with using accuracy (or its complement, the *error rate*) as way of summarising or evaluating the performance of a classifier.

You may also wish to use different metrics for building/training classifiers, selecting between them, and reporting results. In particular, the reporting will typically use a few easily comprehensible summaries of the classifier.

Exercise 1. Fill out the table and compute the macro-F1 score as well as the accuracy.



		Predicted class			
					Total
True class		3	3	0	6
		4	2	2	8
		1	3	2	6
Total		8	8	4	20

	Precision	Recall	F1-score
			
			
			

Exercise 2. Now look at the classification of just ponies and unicorns. Imagine that we have a classifier that leads to the following confusion matrix:

		
	30	60
	20	50

- a) For each of the two classes, find the precision, recall, and F1-score.









	Precision	Recall	F1-score
			
			

An important take-home message from this exercise is that precision and recall consider the performance on a specific class, as does F1-score. So for a binary classifier, switching the two labels (positive/negative) will lead to different results and also different choices of classifiers if you use F1-score to select between them.




- b) Compute the macro-F1 score for the classifier. Is this the same if you switch the labels?
c) What would be the test error (expected loss) for this classifier when loss is evaluated by the following loss matrix:





$$L = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$$

Exercise 3. Imagine that we apply two different classifiers to a balanced dataset with 100 instances in each of two classes. Classifier A has the highest accuracy, and classifier B has the highest F1-score (considering the pony as the “positive” class). Give an example of what the two confusion matrices could look like.


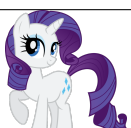

Classifier A				Classifier B			
			Total				Total
			100				100
			100				100


Exercise 4. Now imagine that unicorns are very rare, but that the main class of interest is still ponies, and give two examples of confusion matrices as in Exercise 3.

Classifier C			
			Total
			1000
			10

Classifier D			
			Total
			1000
			10

Exercise 5. What if the unicorns were the main class of interest rather than the ponies?

Classifier E			
			Total
			10
			1000

Classifier F			
			Total
			10
			1000

Exercise 6. For this exercise, you will need Python and the Default dataset from ISLwR.

- Split the Default dataset into test and training datasets, and train an LDA classifier using just one feature, the balance.
- The ROC curve plots true positive rate against true negative rate. For a range of probability thresholds p for the posterior probability of defaulting on the credit card, compute the two rates that the resulting classifier gives on the test set and visualise this as the ROC curve.
- Classifying according to a probability threshold p for posterior probabilities corresponds to choosing a classifier that minimizes expected loss for some loss matrix. What could this loss matrix be, if we classify to “default” whenever its posterior probability is above 0.9? *Note that there are many loss matrices that would lead to the same detection threshold of 0.9*
- If you have more time, fit a logistic regression model and plot the ROC curve and compare to that for LDA.