Performance Indices

UBCO MDS — DATA 572

Classification Performance

- In most classification tasks, we viewed classification (or confusion) tables and discussed misclassification rates to measure the performance of a classifier
- Let's delve a bit deeper and see cases where misclassification rates could be misleading.
- BIG NOTE: ALL supervised indices should be estimated for the long-run — none are robust to an overfitted model. AKA, we should be using cross-validation and/or train-test setups to estimate these!
- We'll begin with simple, binary classification tasks...

Classification Performance: Binary

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	a	b
	FALSE (0)	c	d

- a = true positive (TP). Correct!
- b = false negative (FN). Incorrect!
- c = false positive (FP). Incorrect!
- d = true negative (TN). Correct!
- $a + b + c + d = n$
- Misclassification rate: $\frac{b+c}{a+b+c+d} = \frac{b+c}{n}$
- Classification accuracy: $\frac{a+d}{a+b+c+d} = \frac{a+d}{n}$
- Misclassification rate + Classification accuracy $= \frac{n}{n} = 1$

Classification Accuracy / Misclassification Rate

- So...is a model that provides 0.9 classification accuracy good?
- Problem: that depends...

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	88	0
	FALSE (0)	10	2

- For unbalanced classes, high classification accuracy (equivalently, low misclassification rates) can be especially deceiving.
- While these measures are easily extended to multi-class scenarios, the problem with unbalanced classes remains.

Classification Accuracy / Misclassification Rate

- Another problem, which will essentially remain regardless of the general performance indices we choose. Is a 0.98 classification accuracy inherently 'better' than 0.82? Well...

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	87	1
	FALSE (0)	1	11

versus

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	79	9
	FALSE (0)	9	3

Classification Accuracy / Misclassification Rate

- But on the other hand...

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	88	0
	FALSE (0)	2	10

versus

		Predicted	
		TRUE (1)	FALSE (0)
Actual	TRUE (1)	70	18
	FALSE (0)	0	12

An aside...

- It's worth noting that the concerns just outlined also occur for continuous responses.
- For example, the mean-squared-error (MSE) is a metric for model performance which is particularly non-robust to outliers (because of the squaring). Meaning that minimizing the MSE (or RSS) explicitly means attempting to avoid ANY large mistakes in prediction as best we can.
- In some scenarios, one might prefer to minimize the mean-absolute-error (MAE), which could choose a model that makes fewer small mistakes and more big ones!
- That choice is, of course, application dependent. Further note that explicitly minimizing metrics other than RSS for fitting a model can often prove difficult (mathematically/statistically).

Using probabilities

- Most classification models provide probabilistic responses for each class, which can be incorporated into useful metrics of a models performance
- Notation-wise, lets use z_{ig} to correspond to the probability that observation i belongs to group/class g
- So for example, if observation 1 belongs to class 3
  - Model 1: $z_1 = (0.10, 0.30, 0.60)$
  - Model 2: $z_1 = (0.05, 0.05, 0.90)$
- Then which model is better?
- Classification accuracy?

here model 2 is better if observation 1 is in class 3 because it classifies observation in group 3 with a higher certainty (0.9), however, their classification accuracy would be the same.

Logloss

- One popular misclassification measure, which is also easily defined in this multi-class scenario, is logloss.
- Logloss is defined by

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{G} I(y_i = g)\log z_{ig}$$

- So for our previous example...where obs 1 belongs to class 3
  - Model 1: $z_1 = (0.10, 0.30, 0.60)$
  - Model 2: $z_1 = (0.05, 0.05, 0.90)$
  - Model 3: $z_1 = (0.55, 0.44, 0.01)$
- Logloss for model 1 $= -\log(0.60) = 0.511$
- Logloss for model 2 $= -\log(0.90) = 0.105$
- Logloss for model 3 $= -\log(0.01) = 4.605$

Logloss

- From the example it's clear that logloss does not have simple upper bound.
- In fact, since -log(0) = Inf, it is technically unbounded by above. The lower bound (perfect probabilistic classifier), would be log(1) = 0 for each observation.
- This means that even just one highly confident misclassification is heavily penalized by this metric.

Benchmarking Unsupervised

- Another problem: suppose a clustering method provides a three group solution, though for benchmarking we know there are only two groups in the data set.

		Predicted		
		1	2	3
Actual	A	10	10	50
	B	90	0	0

- First, how many misclassifications are there here?
- If you say 10 misclass, then that metric would be equivalent to the following, better solution

here, group 1 is not necessarily group "A", because 2 has majority then that is most likely group A and 1 is mostly be for the 2nd row so its mostly B.

		Predicted	
		1	2
Actual	A	10	60
	B	90	0

Benchmarking Unsupervised

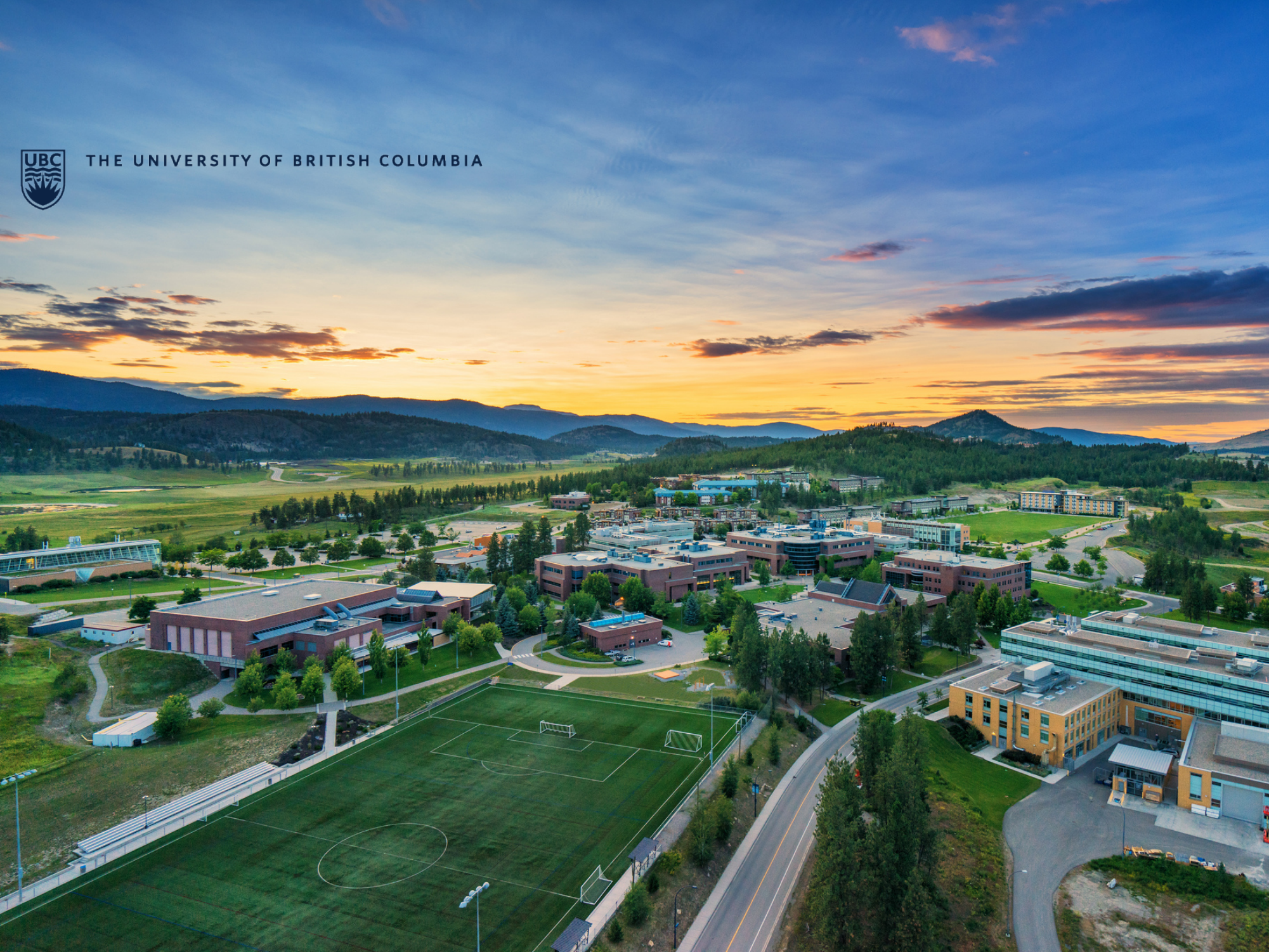
- One common metric used in unsupervised benchmarking (to either other model predictions, or in some cases known values for true benchmarking) is the Rand index.
- It compares two classification vectors
- Rand index $= \frac{\#\text{ of pairwise agreements}}{\#\text{ of pairs}}$
- Where number of pairwise agreements means, for example,
  1. Observations a and b are grouped together in model 1, and also in model 2
  2. Observations y and z are NOT grouped together in model 1, and also NOT together in model 2
- Rand of 1 means perfect agreement between models, buuuttttt....

Benchmarking Unsupervised

- One major problem with the Rand index is that even if one is performing two random groupings, the expected value is something other than 0 (and depends on both the sample size as well as the number of groups).
- This makes the Rand index difficult to interpret for intermediate values (a Rand of 0.6 could be quite good or quite terrible).
- Luckily, there is an Adjusted Rand Index (ARI) which standardizes the Rand index (details out of scope)
- For the ARI, a value of 0 suggests no better than randomly classifying. A value of 1 still suggests perfect agreement.
- Importantly (especially for 572), the ARI can be used in the context of supervised learning as well. Also importantly, it suffers from similar problems as (mis)class rates for unbalanced data (it essentially treats groups as equivalently important).

Final note

- For measuring classification/clustering performance, it is generally good practice to report the classification table alongside any chosen metrics
- This can help flag any strangities in results that might be missed by looking at just summary metrics.


THE UNIVERSITY OF BRITISH COLUMBIA