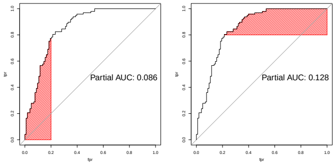
Introduction to Machine Learning

Evaluation: AUC Extensions

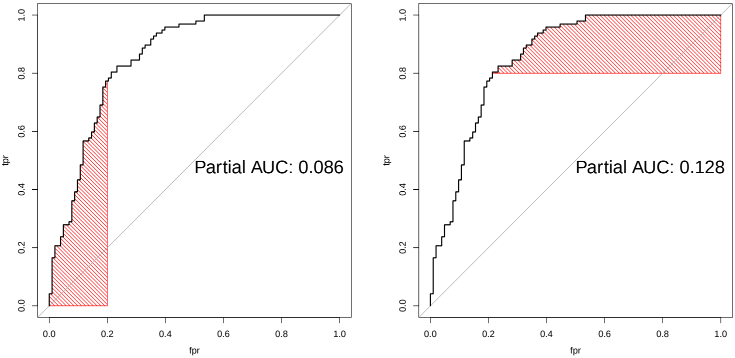


Learning goals

- Understand why pAUC is a reasonable metric in some contexts
- Know how pAUC is computed and normalized
- Understand multi-class AUC

PARTIAL AUC

- Sometimes it can be useful to look at a specific region under the ROC curve ⇒ partial AUC (pAUC).
- For example, we might focus on a region with low FPR or a region with high TPR:



PARTIAL AUC – EXAMPLE

- Applications where sensitivity and specificity are treated asymmetrically often occur in biomedical contexts.
- For example, Wild et al. (2010) used pAUC in their study of biomarkers for the detection of colorectal cancer.
- Sensitivity, i.e., being able to correctly detect present diseases, is crucial in this setting.
- At the same time, high sensitivity is only useful if the classifier also achieves high specificity.
  $\rightarrow$ Otherwise, healthy patients might receive costly and entirely unnecessary treatment.
- It is therefore reasonable to demand a certain level of specificity and evaluate/optimize learners on the resulting pAUC.

CORRECTED PARTIAL AUC

- The scale of the partial AUC depends on the FPR cut-off values used to determine the region of interest $\Rightarrow$ pAUC $\in [0, c_2 - c_1]$.
- For standard AUC, we have $c_1 = 0$ and $c_2 = 1$.
- We can scale pAUC to take on values in $[0, 1]$ again:

$$\text{pAUC}_{\text{corrected}} = \frac{1}{2}\left(1 + \frac{\text{pAUC} - \text{AUC}_{\text{min}}}{\text{AUC}_{\text{max}} - \text{AUC}_{\text{min}}}\right),$$

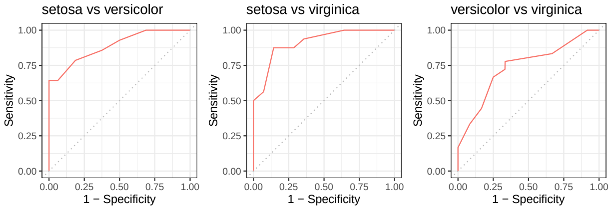
  where

  - AUC_{min} is the value of the non-discriminant AUC, and
  - AUC_{max} is the maximum possible AUC in the region.

- NB: using pAUC means casting aside parts of the information deliberately.

MULTI-CLASS AUC

- In its original form, AUC (as the other ROC metrics) is defined for the binary-class case.
- We can extend AUC to **multi-class** classification, where estimating the area under the ROC curve evolves into estimating the hypervolume under the ROC surface.
- This can be achieved by considering a set of two-dimensional curves, resulting from binry comparisons, and subsequent aggregation.
  $\rightarrow$ In principle, we have the choice between one-vs-one and one-vs-rest comparisons.



One-vs-one comparisons between classes for classification of `iris` species with LDA according to sepal width.

MULTI-CLASS AUC

- For the first possibility, Hand and Till (2001) proposed to average the AUC of respective pairwise comparisons between two classes.
  - First, compute for all pairs of classes $k, \ell \in \{1, \dots, g\}$ the probability $\text{AUC}(k \mid \ell)$ of a randomly drawn member of class k having a lower probability of belonging to class ℓ than a randomly drawn member of class ℓ.
  - For $g = 2$, we have $\text{AUC}(k \mid \ell) = \text{AUC}(\ell \mid k)$, but not necessarily so for $g > 2$.
  - However, since class identifiability is immune to any bijective transformation of the labels, we cannot distinguish $\text{AUC}(k \mid \ell)$ from $\text{AUC}(\ell \mid k)$, so we set $\text{AUC}(k, \ell) = \frac{1}{2} \cdot [\text{AUC}(k \mid \ell) + \text{AUC}(\ell \mid k)]$.
  - Averaging over all pairs of classes yields the overall AUC_{MC} as a multi-class performance metric:

$$\text{AUC}_{MC} = \frac{2}{g(g+1)} \sum_{k<\ell} \text{AUC}(k, \ell) \in [0, 1].$$

  - This reduces to the standard AUC for the binary case.