4 Machine Learning Tips and Tricks

4.1 Metrics

Given a set of data points $\{x^{(1)},...,x^{(m)}\}$, where each $x^{(i)}$ has n features, associated to a set of outcomes $\{y^{(1)},...,y^{(m)}\}$, we want to assess a given classifier that learns how to predict y from x

4.1.1 Classification

In a context of a binary classification, here are the main metrics that are important to track to assess the performance of the model.

□ Confusion matrix – The confusion matrix is used to have a more complete picture when assessing the performance of a model. It is defined as follows:

		Predicted class		
		+	_	
Actual class	+	TP True Positives	FN False Negatives Type II error	
	_	FP False Positives Type I error	TN True Negatives	

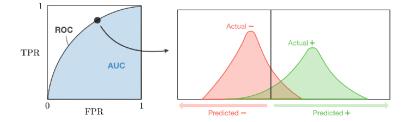
☐ Main metrics – The following metrics are commonly used to assess the performance of classification models:

Metric	Formula	Interpretation	
Accuracy	$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	Overall performance of model	
Precision	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$	How accurate the positive predictions are	
Recall Sensitivity	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	Coverage of actual positive sample	
Specificity	$\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$	Coverage of actual negative sample	
F1 score	$\frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$	Hybrid metric useful for unbalanced classes	

 \square ROC – The receiver operating curve, also noted ROC, is the plot of TPR versus FPR by varying the threshold. These metrics are are summed up in the table below:

Metric	Formula	Equivalent
True Positive Rate TPR	$\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$	Recall, sensitivity
False Positive Rate FPR	$\frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$	1-specificity

□ AUC – The area under the receiving operating curve, also noted AUC or AUROC, is the area below the ROC as shown in the following figure:



4.1.2 Regression

 \square Basic metrics – Given a regression model f, the following metrics are commonly used to assess the performance of the model:

Total sum of squares	Explained sum of squares	Residual sum of squares	
$SS_{tot} = \sum_{i=1}^{m} (y_i - \overline{y})^2$	$SS_{reg} = \sum_{i=1}^{m} (f(x_i) - \overline{y})^2$	$SS_{res} = \sum_{i=1}^{m} (y_i - f(x_i))^2$	

 \Box Coefficient of determination – The coefficient of determination, often noted R^2 or r^2 , provides a measure of how well the observed outcomes are replicated by the model and is defined as follows:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

 \square Main metrics – The following metrics are commonly used to assess the performance of regression models, by taking into account the number of variables n that they take into consideration:

Mallow's Cp	AIC	BIC	Adjusted R^2
$\frac{\mathrm{SS}_{\mathrm{res}} + 2(n+1)\widehat{\sigma}^2}{m}$	$2\Big[(n+2)-\log(L)\Big]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1 - R^2)(m - 1)}{m - n - 1}$

where L is the likelihood and $\hat{\sigma}^2$ is an estimate of the variance associated with each response.