MA679 Lab2

Classification Evaluation Model Metrics

- **Confusion matrix**

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

- **Precision, recall, F1 score**

precision=TP/(TP+FP),
recall=TP/(TP+FN),

$$F1-score=\frac{2\cdot precision\cdot recall}{precision+recall}=\frac{2TP}{2TP+FN+FP}$$

  - Remarks:
    - Precision: the fraction of retrieved instances that are relevant.
    - Recall: accuracy on the positive class.
    - F1-score ranges in [0,1]. F1-score=0 indicates all the positive samples are misclassified. F1-score=1 indicates a perfect classifier.
    - F1-score is independent from TN.
    - F1-score is not symmetric for class swapping.

- **Mattews Correlation Coefficient (MCC)**

$$MCC=\frac{TP\cdot TN-FP\cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}}$$

  - Remarks:
    - MCC ranges in [-1,1]. -1 and 1 indicate perfect misclassification and perfect classification respectively. 0 indicates a random classifier.
    - MCC is invariant to class swapping.
    - When working with imbalanced data, use MCC and F1-score instead of accuracy.
    - MCC is a **robust** metric: it generates a high quality score **only** if the prediction correctly classified a high percentage of negative data instances and a high percentage of positive data instances, with any class balance or imbalance.

Reference: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. David Chicco & Giuseppe Jurman.

Table 1

	Balance		Confusion matrix				Accuracy [0, 1]	F_1 score [0, 1]	MCC [−1, +1]	Figure	Informative Response
	Pos	Neg	TP	FN	TN	FP					
Use case A1 Positively imbalanced dataset	91	9	90	1	0	9	0.90	0.95	**−0.03**	Figure 2	**MCC**
Use case A2 Positively imbalanced dataset	75	25	5	70	19	6	**0.24**	**0.12**	**−0.24**	Suppl. Additional file 1	**Accuracy, F_1 score, MCC**
Use case B1 Balanced dataset	50	50	47	3	5	45	**0.52**	0.66	**+0.07**	Suppl. Additional file 2	**Accuracy, MCC**
Use case B2 Balanced dataset	50	50	10	40	46	4	**0.56**	**0.31**	**+0.17**	Suppl. Additional file 3	**accuracy, F_1 score, MCC**
Use case C1 Negatively imbalanced dataset	10	90	9	1	1	89	**0.10**	**0.17**	**−0.19**	Suppl. Additional file 4	**accuracy, F_1 score, MCC**
Use case C2 Negatively imbalanced dataset	11	89	2	9	88	1	0.90	**0.29**	**+0.31**	Suppl. Additional file 5	**F_1 score, MCC**

For the Use case A1, MCC is the only statistical rate able to truthfully inform the readership about the poor performance of the classifier. For the Use case B1, MCC and accuracy are able to inform about the poor performance of the classifier in the prediction of negative data instances, while for the Use case A2, B2, C1, all the three rates (accuracy, F_1, and MCC) are able to show this information. For the Use case C2, the MCC and F_1 are able to recognize the weak performance of the algorithm in predicting one of the two original dataset classes. pos: number of positives. neg: number of negatives. TP: true positives. FN: false negatives. TN: true negatives. FP: false positives. Informative response: list of confusion matrix rates able to reflect the poor performance of the classifier in the prediction task. We highlighted in bold the informative response of each use case

Table 2: Colon cancer gene expression data (35.48% nagatives and 64.52% positives)

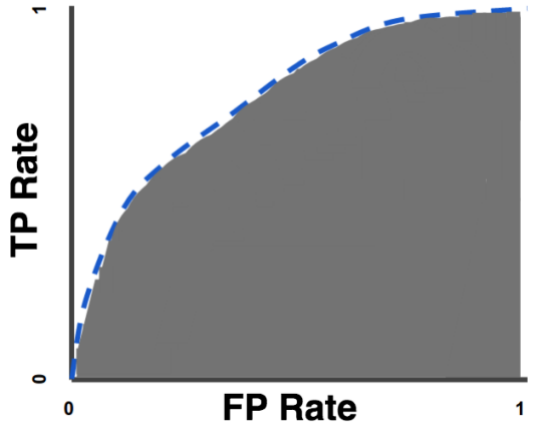
Classifier	MCC	F_1 score	Accuracy	TP rate	TN rate
MCC ranking:					
Gradient boosting	**+0.55**	0.81	0.78	0.85	0.69
Decision tree	**+0.53**	0.82	0.77	0.88	0.58
k-nearest neighbors	**+0.48**	0.87	0.80	0.92	0.52
Linear SVM	**+0.41**	0.82	0.76	0.86	0.53
Radial SVM	**+0.29**	0.75	0.67	0.86	0.40
F_1 score ranking:					
k-nearest neighbors	+0.48	**0.87**	0.80	0.92	0.52
Linear SVM	+0.41	**0.82**	0.76	0.86	0.53
Decision tree	+0.53	**0.82**	0.77	0.88	0.58
Gradient boosting	+0.55	**0.81**	0.78	0.85	0.69
Radial SVM	+0.29	**0.75**	0.67	0.86	0.40
Accuracy ranking:					
k-nearest neighbors	+0.48	0.87	**0.80**	0.92	0.52
Gradient boosting	+0.55	0.81	**0.78**	0.85	0.69
Decision tree	+0.53	0.82	**0.77**	0.88	0.58
Linear SVM	+0.41	0.82	**0.76**	0.86	0.53
Radial SVM	+0.29	0.75	**0.67**	0.86	0.40

Prediction results on colon cancer gene expression dataset, based on MCC, F_1 score, and accuracy. linear SVM: support vector machines with linear kernel. MCC: worst value −1 and best value +1. F_1 score, accuracy, TP rate, and TN rate: worst value 0 and best value 1. To avoid additional complexity and keep this table simple to read, we prefered to exclude the standard deviation of each result metric. We highlighted in bold the ranking of each rate

- **Logloss**
  For logistic regression:

$$\textbf{Logloss} = -\frac{1}{n}\sum(y_i \ln(p_i) + (1-y_i)\ln(1-p_i))$$

- **AUC (Area under ROC curve)**
  - TP rate/recall=TP/(TP+FN), FP rate=FP/(FP+TN)



  - Remarks:
    - An AUC of 1 indicates a perfect classifier, while an AUC of 0.5 indicates a poor classifier.
    - Displays two types of errors for all possible thresholds.

Imbalanced Data

How to deal with imbalanced data?

- Be careful when choosing evaluation metrics. Use MCC, F1 instead of accuracy.
- Oversampling: SMOTE (synthetic minority over-sampling technique): interpolate with nearest neighbors.
- Downsampling: sample without replacement.