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)},wewanttoassessagivenclassifierthatlearnshowtopredictyfromx.

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.

r **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:

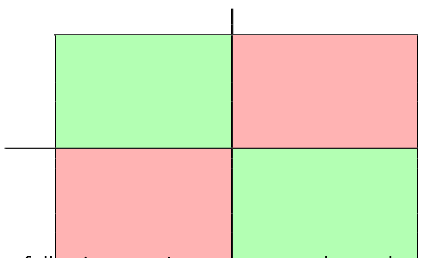
Predicted class

\+ –

TPFN

\+ FalseNegatives
True Positives
Type II error
Actual class
FPTN

– FalsePositives
True Negatives
Type I error



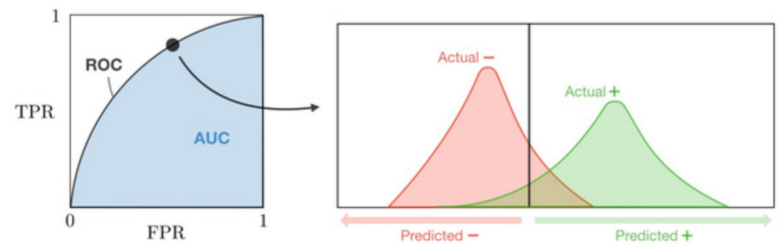
r **Main metrics** – The following metrics are commonly used to assess the performance of classification models:

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP+TN+FP+FN}$	Overallperformanceofmodel
Precision	$\frac{TP}{TP + FP}$	Howaccuratethepositivepredictionsare
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverageofactualpositivesample
Specificity	$\frac{TN}{TN + FP}$	Coverageofactualnegativesample
F1score	$\frac{2TP}{2TP + FP + FN}$	Hybridmetricusefulforunbalancedclasses

r **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
TruePositiveRate TPR	$\frac{TP}{TP + FN}$	Recall,sensitivity
FalsePositiveRate FPR	$\frac{FP}{TN + FP}$	1-specificity

r **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:



Regression

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

Totalsumofsquares	Explainedsumofsquares	Residualsumofsquares
$SS_{tot} = \sum_{i=1}^{m} (y_i - \bar{y})^2$	$SS_{reg} = \sum_{i=1}^{m} (f(x_i) - \bar{y})^2$	$SS_{res} = \sum_{i=1}^{m} (y_i - f(x_i))^2$

r **Coefficient of determination** – The coefficient of determination, often noted R2 or r2, provides a measure of how well the observed outcomes are replicated by the model and is defined as follows:

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

r **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 consid- eration:

Mallow'sCp	AIC	BIC	AdjustedR
$\frac{SS_{res} + 2(n+1)\hat{\sigma}^2}{m}$	$2[(n+2) - \log(L)]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1-R^2)(m-1)}{m-n-1}$

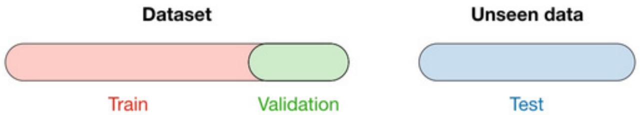
where L is the likelihood and $\hat{\sigma}^2$ is an estimate of the variance associated with each response.

Model selection

r **Vocabulary** – When selecting a model, we distinguish 3 different parts of the data that we have as follows:

Training set	Validation set	Testing set
- Model is trained - Usually 80% of the dataset	- Model is assessed - Usually 20% of the dataset - Also called hold-out or development set	- Model gives predictions - Unseen data

Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:

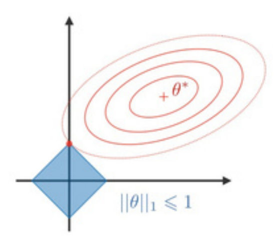
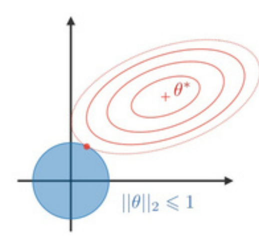
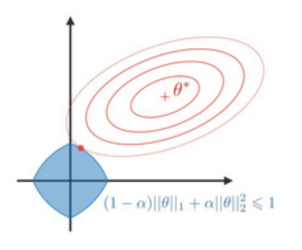


r **Cross-validation** – Cross-validation, also noted CV, is a method that is used to select a model that does not rely too much on the initial training set. The different types are summed up in the table below:

k-fold	Leave-p-out
- Training on $k-1$ folds and assessment on the remaining one - Generally $k = 5$ or 10	- Training on $n-p$ observations and assessment on the p remaining ones - Case $p = 1$ is called leave-one-out

The most commonly used method is called k-fold cross-validation and splits the training data into k folds to validate the model on one fold while training the model on the $k-1$ other folds, all of this k times. The error is then averaged over the k folds and is named cross-validation error.

r **Regularization** – The regularization procedure aims at avoiding the model to overfit the data and thus deals with high variance issues. The following table sums up the different types of commonly used regularization techniques:

LASSO	Ridge	Elastic Net
- Shrinks coefficients to 0 - Good for variable selection	Makes coefficients smaller	Tradeoff between variable selection and small coefficients
		
$... + \lambda\|\|\theta\|\|_1$ $\lambda \in \mathbb{R}$	$... + \lambda\|\|\theta\|\|_2^2$ $\lambda \in \mathbb{R}$	$... + \lambda\Big[(1-\alpha)\|\|\theta\|\|_1 + \alpha\|\|\theta\|\|_2^2\Big]$ $\lambda \in \mathbb{R}, \quad \alpha \in [0,1]$

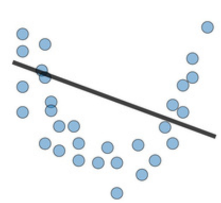
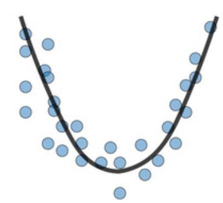
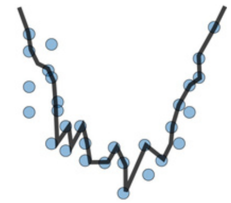
r **Model selection** – Train model on training set, then evaluate on the development set, then pick best performance model on the development set, and retrain all of that model on the whole training set.

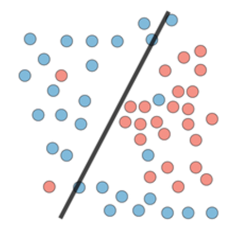
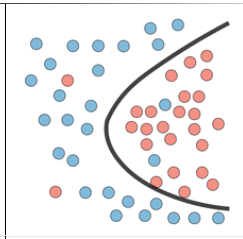
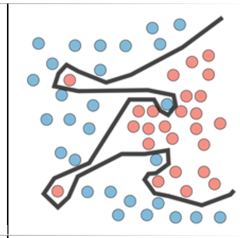
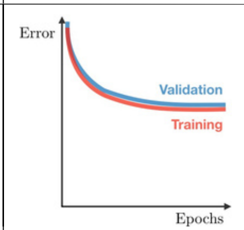
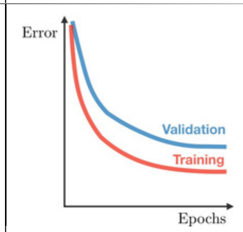
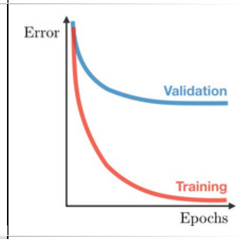
Diagnostics

r **Bias** – The bias of a model is the difference between the expected prediction and the correct model that we try to predict for given data points.

r **Variance** – The variance of a model is the variability of the model prediction for given data points.

r **Bias/variance tradeoff** – The simpler the model, the higher the bias, and the more complex the model, the higher the variance.

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			

Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

r **Error analysis** – Error analysis is analyzing the root cause of the difference in performance between the current and the perfect models.

r **Ablative analysis** – Ablative analysis is analyzing the root cause of the difference in performance between the current and the baseline models.