Classification, Validation and Optimization of Machine Learning Models

Michaël Dell'aiera

ESCAPE Summer School 2022

Laboratoire d'Annecy de Physique des Particules (LAPP) - IN2P3

June 23, 2022

Presentation Outline

Guidelines

I have a classification problem, what should I do ?

I know how to classify, but what about the actual training data ?

I think I correctly trained my classifier, how do I evaluate it ?

How do I optimize my classifier ?

Perceptron

- Developed by Rosenblatt in the 1950s.
- Inspired by nature : Mimic a brain neuron behaviour.
- Classifies linearly separable binary data.
- Trainable parameters : w_i and b.
  - ✓ Learns a linear projection of the data.
  - ✓ Converges toward ONE possible solution.

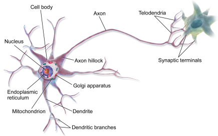

Figure 1: Brain neuron.

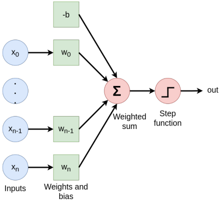

Figure 2: Perceptron.

Perceptron

- Training step : Obtain a decision boundary.
- Test step : Apply on new data.
- Perceptrons are not robust : if the decision boundary is close to the dataset, test samples might be misclassified.
- Doesn't work on non-linearly separable dataset (XOR).
- Improvements → Support Vector Machines (SVM)

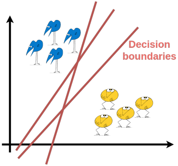

Figure 3: Linearly separable dataset and possible decision boundaries.

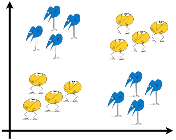

Figure 4: XOR dataset.

Support Vector Machine (SVM)

- **Margin maximization** : Find the hyperplane that separates the two sets by the largest margin between the "support vectors".

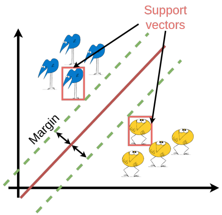

Figure 5: Margin maximization.

- **Kernel trick** : Project to more dimensions which are non-linear functions of the original ones, so that we can now separate linearly.

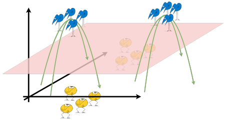

Figure 6: Kernel trick.

Guidelines

I have a classification problem, what should I do ?

I know how to classify, but what about the actual training data ?

I think I correctly trained my classifier, how do I evaluate it ?

How do I optimize my classifier ?

Dataset split

- Classical splitting of the data : Around $\{0.6, 0.2, 0.2\}$.
  - × Not always the best strategy (especially small datasets).
  - ✓ How to make use of the validation dataset in the training procedure ?
    $\rightarrow$ **Cross-validation !**
- Shuffling + Random split $\rightarrow$ Independent observations
  - × Solar panel data : varies depending on the location, the seasons, ...

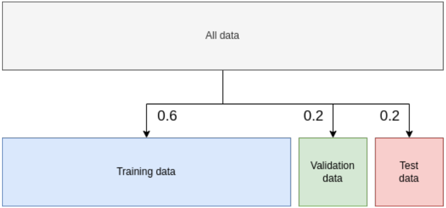

Figure 7: Classical dataset split

Cross-validation

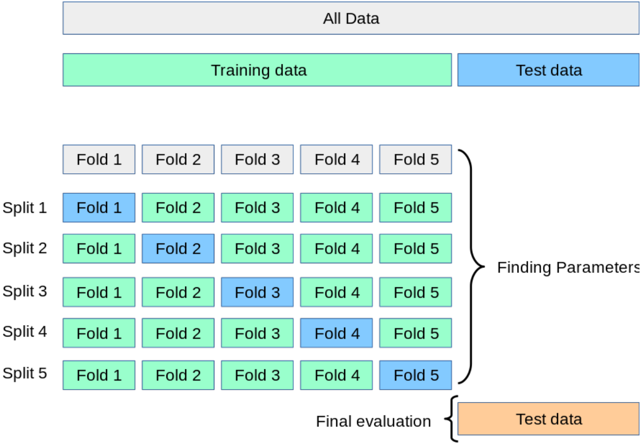

Figure 8: Cross-validation principle. Source : scikit-learn.

Cross-validation

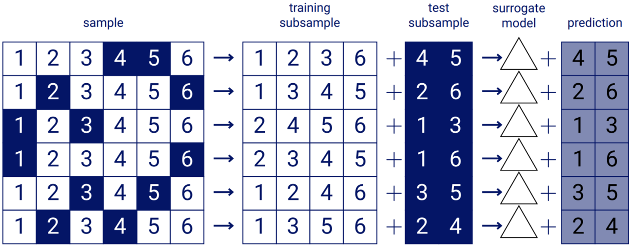

Figure 9: Cross-validation principle.

- Each surrogate model sees a part of the training dataset.
- The whole training dataset is used.
- Score = Mean(surrogate$_1$, ..., surrogate$_k$)

Under-fitting and Over-fitting

- **Under-fitting :** Input variables are not significant enough or the model has not trained enough.
- **Over-fitting :** The model fits exactly the training data.
  - Early-stopping
  - Data augmentation (rotation, noise, ...)
  - Feature selection
  - Regularization (SVM robust to over-fitting)

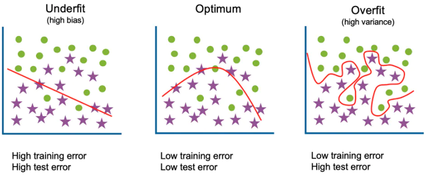

Figure 10: Under-fitting and over-fitting. Source IMB.

We want **generalization**.

Structured data : example

- Let's consider an image $I \in \{0,1\}^{20\times 20}$ made of pixel $p_{i,j} \sim \mathcal{U}\{0,1\}$.
  There are $2^{400} > 10^{80}$ possible outcomes.
  Image $I \in ([0,255]^3)^{10^6}$: No chance to randomly draw the picture of a dog, bird, scene ...

Figure 11: Examples of some pixel space observations.

- Natural images have *structures*, so are gathered into *clusters*.

Statistical Independence

× Test cluster $\neq$ Train cluster.

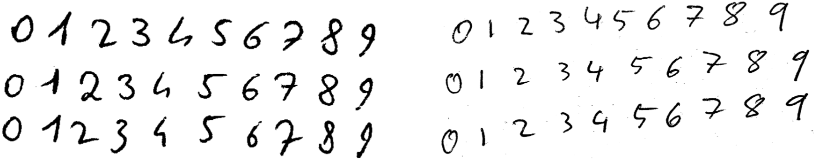

Figure 12: On the left : the train digit dataset. On the right : the test digit dataset. The test and train domains are slightly different (look at the "1").

✓ The train dataset must be representative enough to promote generality.

✓ The test dataset must be independent from the train dataset to avoid bias.

Guidelines

I have a classification problem, what should I do ?

I know how to classify, but what about the actual training data ?

I think I correctly trained my classifier, how do I evaluate it ?

How do I optimize my classifier ?

Figures of merit

- Tools to measure model's performances.
- A variety of figures of merit : Accuracy, Sensitivity, Predictive values, MSE, RMSE, R^2, ...
- Usually you will characterize your model's performances using several figures of merit but choose the most relevant ones.
- Figures of merit are measured and then prone to biases and variances.

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

Mean Squared Error

✓ Regression

- $MSE = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$
- $RMSE = \sqrt{MSE}$ same scale has y

✓ Classification (Brier's score)

- $BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - p_i)^2$

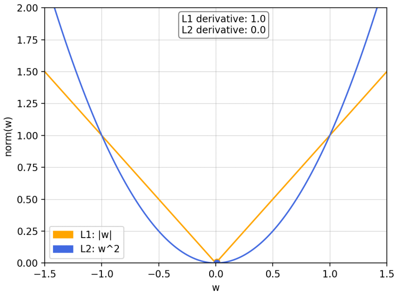

Figure 13: L1 and L2 penalties. L2 penalizes more large deviations.

Confusion matrix

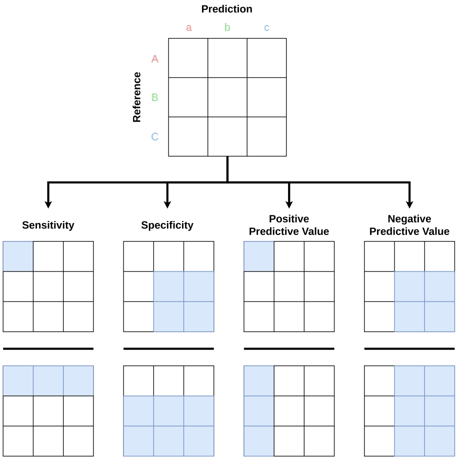

Figure 14: Confusion matrix.

Definitions

- **Sensitivity (Recall)** : of all truly class A cases, which fraction is correctly recognized as class A?

  If I have covid, what is the probability of being tested + ?

- **Specificity** : of all cases truly not belonging to class A, which fraction is correctly recognized as not belonging to class A?

  If I don't have covid, what is the probability of being detected - ?

- **Positive Predictive Value (Precision)** : of all cases predicted to belong to class A, which fraction does truly belong to class A?

  If I am tested +, what is the probability of having covid ?

- **Negative Predictive Value** : of all cases predicted not to belong to class A, which fraction does truly not belong to class A?

  If I am tested -, what is the probability of not having covid ?

Confidence Intervals for Sensitivity

- $x \in A$ or $x \notin A$?
  - ✓ Bernoulli trial of parameter p
- Sum of independent Bernoulli trials $\rightarrow$ Binomial distribution
  - ✓ Appropriate only for $np \geq 5$ and $n(1-p) \geq 5$
- Uncertainty $var(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n_{test}}$
- Think about flipping a coin.

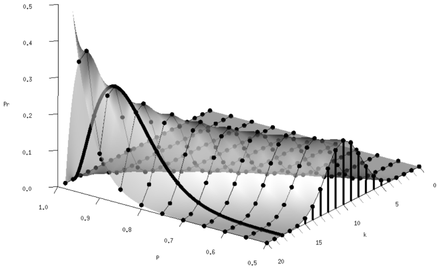

Figure 15: Binomial distribution.

Confidence Intervals for Sensitivity

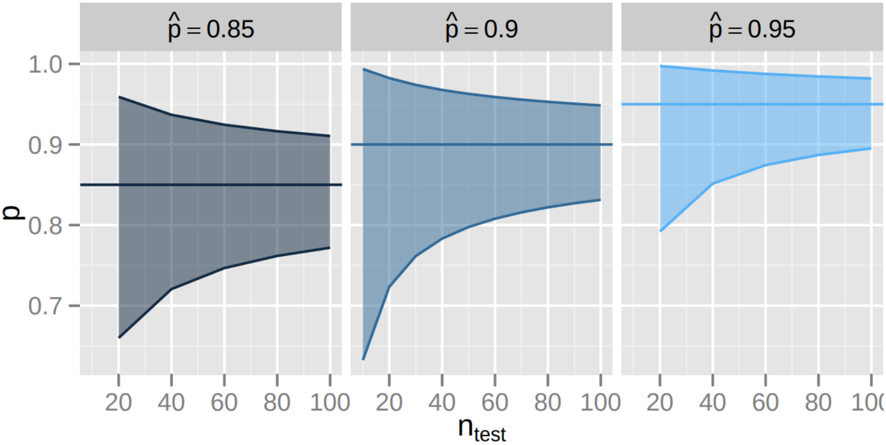

Figure 16: Source [1]. Proportions have bad variance properties.

Receiver Operating Characteristic / Specificity-Sensitivity-Diagram

- Sheds light on the diagnostic ability of a binary classifier at various threshold.
- Plots the True Positive Rate (TPR = sensitivity) against the False Positive Rate (FPR = 1 - specificity).
- Allows to compare different tests.

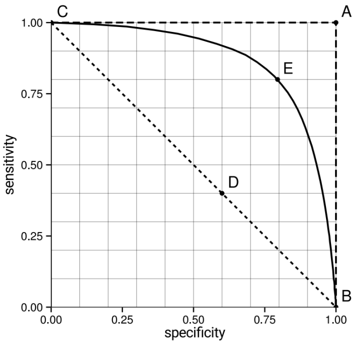

Figure 17: ROC curve.

ROC computation procedure

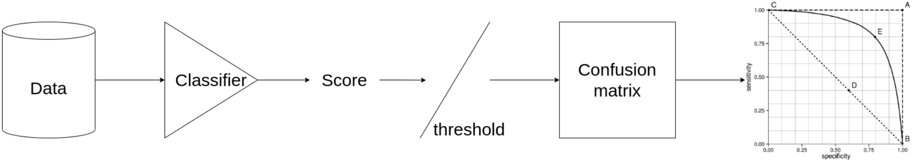

Figure 18: Scores go through different threshold.

- Set of classifiers.
- The best classifier depends on the application.
  - ✓ For a covid test, you may want higher **sensitivity**.

Sensitivity - Specificity Trade-off

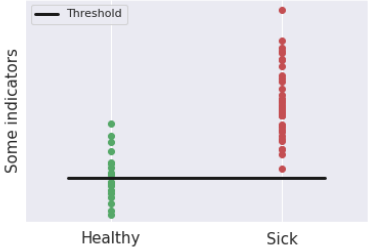

Figure 19: Trade-off between sensitivity and specificity. When the threshold is 0.1, the sensitivity is 0.55 and the specificity is 1.0.

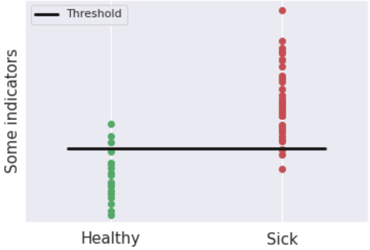

Figure 20: Trade-off between sensitivity and specificity. When the threshold is 0.5, the sensitivity is 0.85 and the specificity is 0.9.

ROC curves : example 1

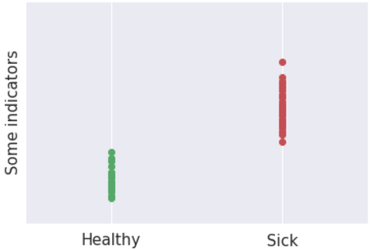

Figure 21: A perfect test would discriminate between the healthy and sick with a sensitivity of 1 and a specificity of 1.

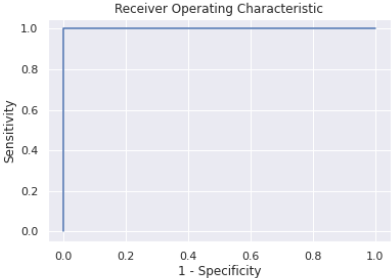

Figure 22: The ROC curve passes through the upper left corner. The area under the ROC curve is 1.

ROC curves : example 2

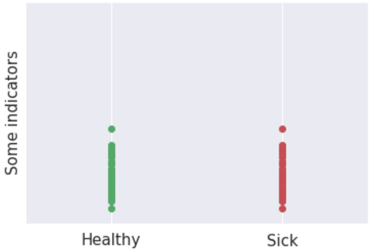

Figure 23: Complete overlap between healthy and sick. This is equivalent to flipping a coin.

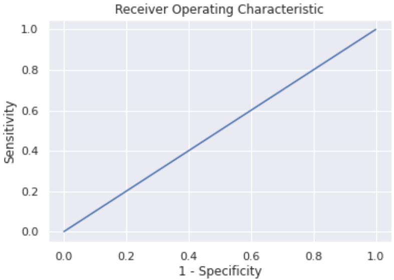

Figure 24: The ROC curve falls on the diagonal line. The area under the ROC curve is 0.5.

Guidelines

I have a classification problem, what should I do ?

I know how to classify, but what about the actual training data ?

I think I correctly trained my classifier, how do I evaluate it ?

How do I optimize my classifier ?

Hyper-parameter optimization

- Parameters vs Hyper-parameters
- A search consists in :
  - A parameter space.
  - A searching procedure (Regular grid, Monte Carlo, Pseudo Monte Carlo).
  - A cross-validation scheme.
  - A function to compute the score.

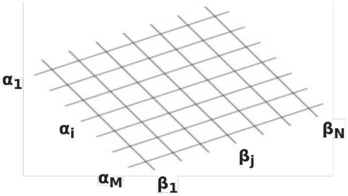

Figure 25: Grid search for two hyper-parameters on a regular grid.

Bias–Variance trade-off

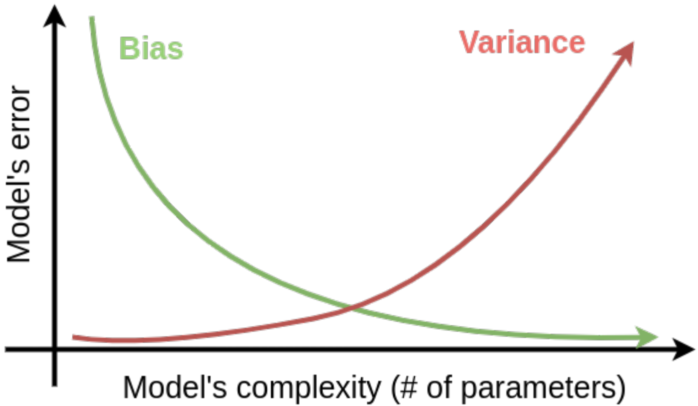

Figure 26: Bias–Variance trade-off.

References

[1] Beleites C et al. "Sample size planning for classification models". In: *Anal Chim Acta* (2012).

Thank You for you attention!