BASICS OF NEURAL NETWORKS

Session 4: Model Evaluation

credits

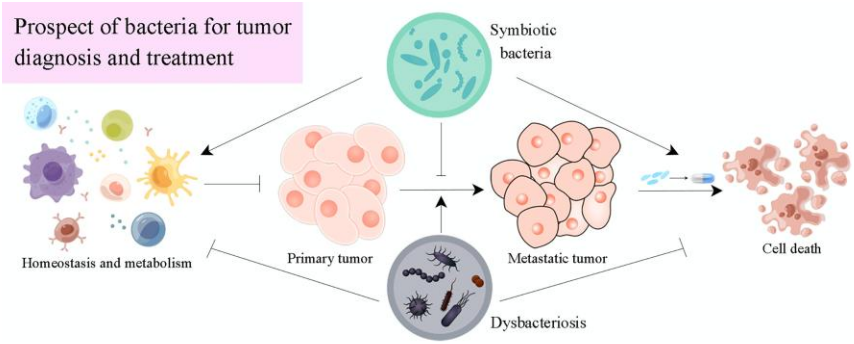
2025 SCHOOL AT THE IAA-CSIC

EDUARDO SÁNCHEZ KARHUNEN

DEPT. ARTIFICIAL INTELLIGENCE. UNIV. SEVILLE. SPAIN

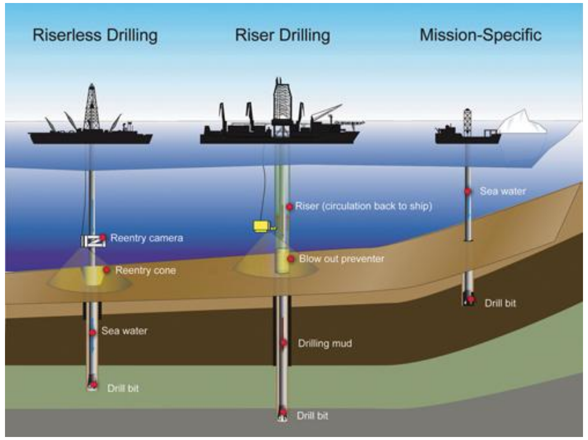
The Accuracy Paradox

- Accuracy can be misleading:
  - Human first reaction: use accuracy.
  - Maybe reliable: balanced dataset.
  - Wrong tool: for imbalanced dataset.



credits

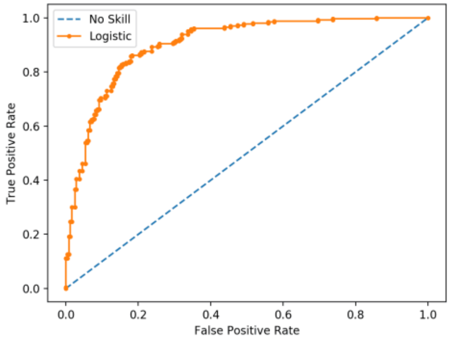
- Rare cases detector:
  - Model to predict disease that affects 1 person in 10,000.
  - Model behavior:
    - Probably model will predict: All no disease.
    - Accuracy = 99,99% but model is completely useless.
    - Balance between accuracy and usefulness.



credits

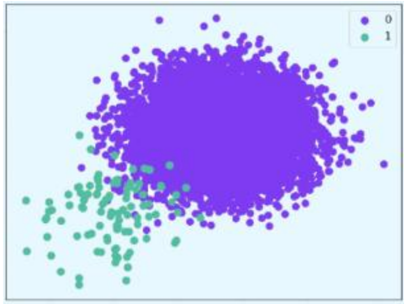
Handling Imbalanced Datasets

- ## Changes in measuring: New Evaluation Metrics
  - Focus on confusion matrix.
  - Prioritize Precision & Recall (F1).
  - Use the ROC Curve.



credits

- ## Changes in Data: Resampling Techniques
  - Augmentation only in the minority class.
  - Undersampling in majority class.

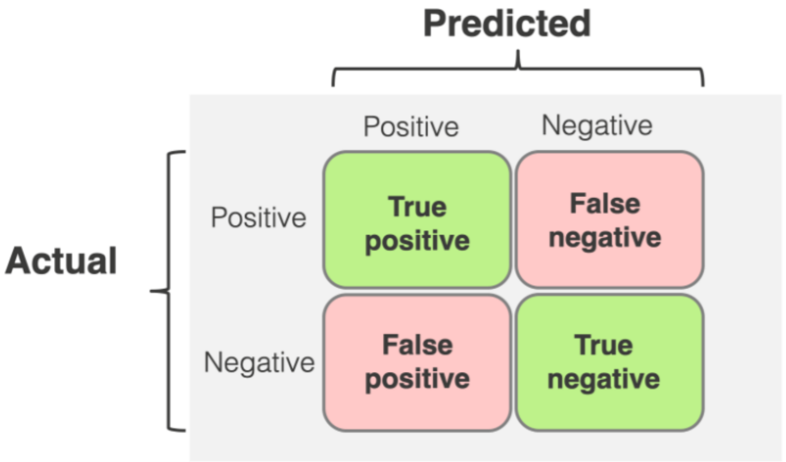


credits

- ## Changes in model learning: Algorithmic approach
  - Class weights.
  - Loss functions adapted for these cases.

The Core Diagnostic Tool: Confusion Matrix

- Foundations of model evaluation:
  - Moves from a single performance number:
    - To a detailed map.
  - Analyze in which classes:
    - Model success and where it fails.
    - It is easier predict a dog as a cat than as a bird.
  - Diagonal cells:
    - Represents the correct predictions (True Positives & True Negatives)
  - Off-diagonal cells:
    - Show which classes are confused easily (False Negative & False Positive)



credits

The Key Diagnostic Metrics

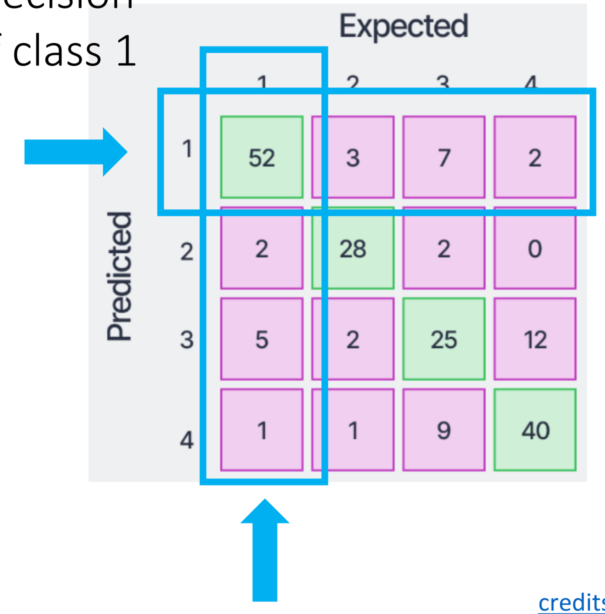
The Precision-Recall trade-off:

- Precision:
  - Measures the precision of the predictions.
  - When model makes a positive prediction, how often is it correct?
- Recall:
  - Measures the completeness of the predictions.
  - Of all the positive actual instances, how many did the model find?
- F1-score:
  - Combines both metrics in a unique measure.
  - It is a harmonic combination, high when both high.

$$F_1\text{-score} = 2 * \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP+FP}$$

Precision of class 1



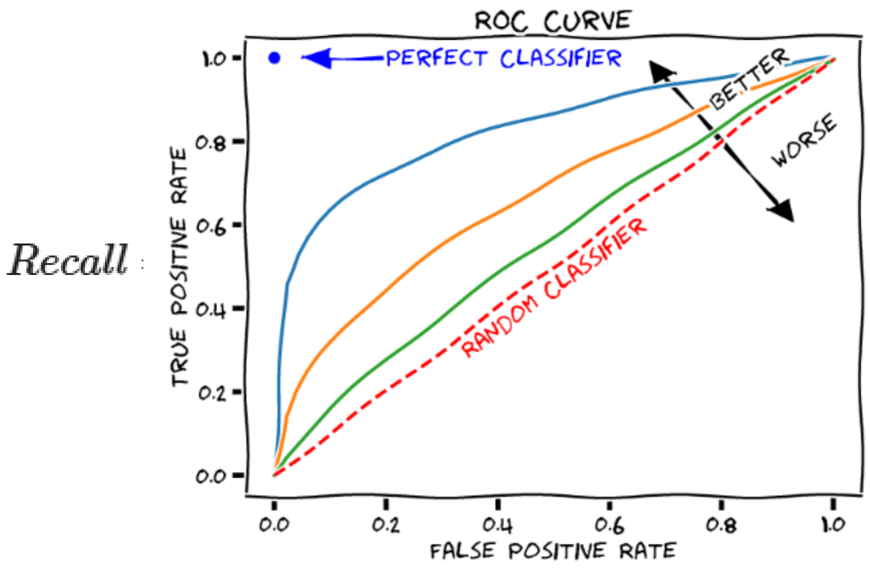
credits

Recall of class 1

$$Recall = \frac{TP}{TP+FN}$$

The ROC Curve

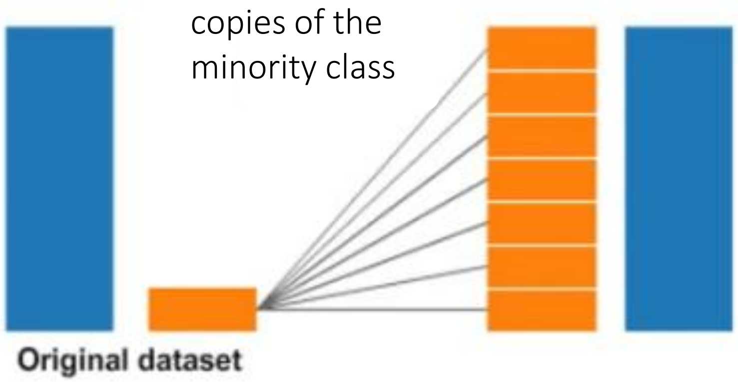
- When threshold moves:
  - Goal: how well model performs at every possible threshold.
  - Measures plotted:
    - Recall (TPR), positives correctly identified.
    - FPR: negatives incorrectly identified as positive.
  - Interpretation:
    - Ideal: TPR = 1 (all positives captured) and FPR = 0.
    - Better performance: the closer the curve to the top-left corner
  - Summarize the curve across all thresholds: Area Under The Curve (AUC)
    - AUC = 1. Perfect classifier
    - AUC < 0.5. Worse than a random guess.



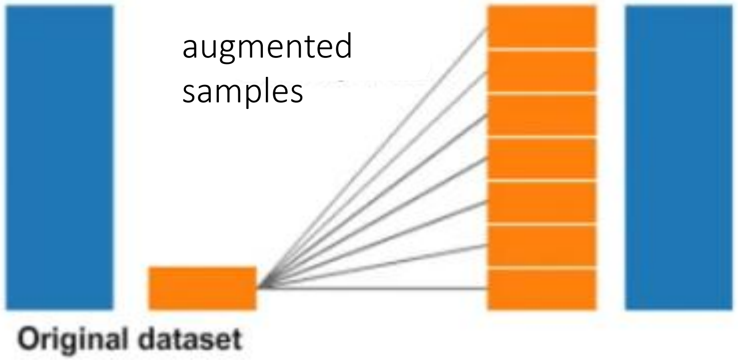
credits

Tackling Imbalance: Upsampling

- Classical ML approach:
  - Random oversampling of the minority class.
    - Exact same photos multiple times in the training.
    - Model sees the exact same images over and over again.
  - Leads to memorizing and contribute to overfitting.



- CNNs Approach:
  - Aggressive data augmentation to the minority class.
  - This is the preferred technique for upsampling the minority class.
  - Images are not duplicated, learn more robust features.
  - Not contributing to overfitting.



credits

Tackling Imbalance: Class Weights

- Source of problem:
  - Loss function is "dominated" by the majority class.
- Mitigate:
  - Increases the importance of the minority class, forcing the model to pay attention.
  - Making errors in the minority class more painful for the model.
  - Applying different weights for each class in the loss function.
- Advantages:
  - No overhead in the training loop
  - It does not contribute to the overfitting, does not duplicate images, learn more robust features

Tackling Imbalance: Loss Functions

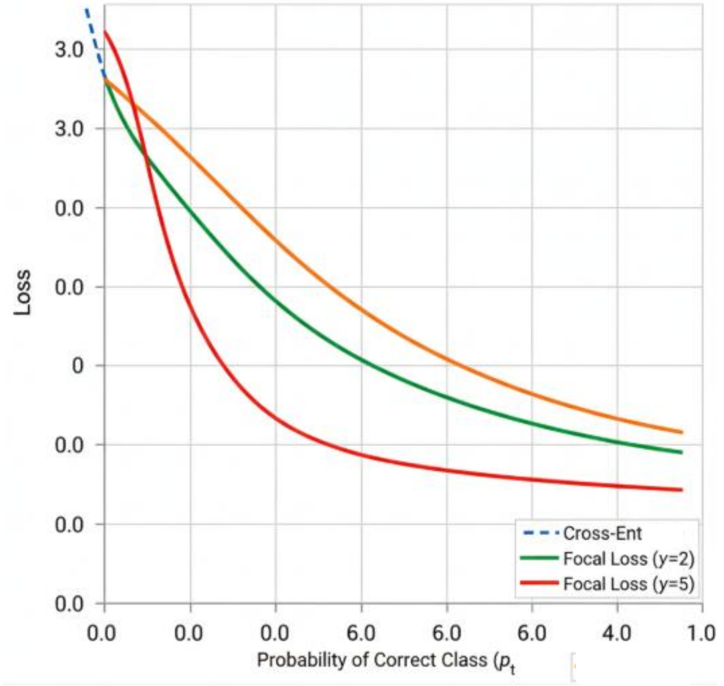
Focal Loss Function:

- Loss functions specifically designed for class imbalance.
- Focal loss function: modification of cross-entropy.

$$\text{Focal Loss} = -(1 - p_t)^{\gamma} \log(p_t)$$

Idea:

- Dynamically reduces the influence of easy examples.
- Parameter γ controls how aggressively model down-weight easy examples
  - In an easy example, p_t is near 1: $(1-p_t)^{\gamma}$ is near zero.
  - In a hard example, p_t is near 0: $(1-p_t)^{\gamma}$ is near one.



- Individual mistakes:
  - Find your worst failures (based on loss)
  - Visualize and search for patterns:
    - Is the model always failing on blurry images?
    - Images with unusual lightning?
    - Images where the object is partially obscured?
  - Error analysis gives you actionable feedback.
    - More diverse training data is needed
    - Or a better preprocessing .
    - Or a different model architecture.

credits

Long-term Performance

- Model drift:
  - Degradation of a model predictive performance over time.
  - Why: Real world is not static (e.g. environmental conditions).
  - Result:
    - A model trained on historical data becomes less accurate as data drifts from patterns it was trained on.
  - Impact:
    - Poor predictions, unreliable insights and loss of trust in the NN.
  - Solution: continuously monitoring and regular retraining.