CIS 635 Knowledge Discovery & Data Mining

Predictive modeling: Classification: Metrics and Imbalanced Data

Classification Matrices

- Accuray

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

Classification Matrices

- Accuray

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?

Classification Matrices

- Accuray

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?
- Not always

Classification Matrices

- Accuray

$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?
- Not always
- Let's analyze the confusion matrix of our Breast cancer prediction problem
  - Accuray metric can be catastrophic

Classification Matrices

- Accuray

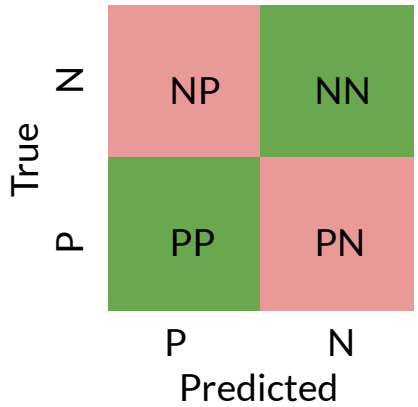
$$\text{Accuracy} = \frac{\text{Nb of correct predictions}}{\text{Nb of (correct + incorrect) predictions}}$$

- Is accuracy a good metric?
- Not always
- Let's analyze the confusion matrix of our Breast cancer prediction problem
  - Accuray metric can be catastrophic
- What other metrics we may use?

Metrics

- Accuray

$$\text{Accuracy} = \frac{PP + NN}{PP + NN + NP + PN}$$



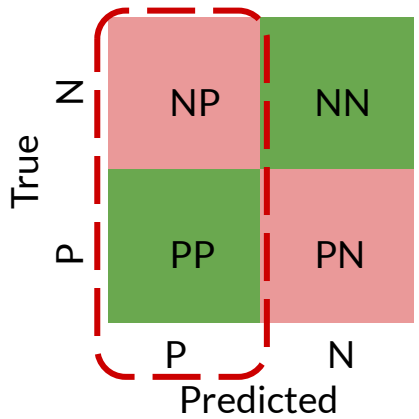
Other important classification metrics

- Precision (also called **Positive Predictive Value)**
- Recall (also called **Sensitivity)**
- F1 Score

Metrics

- Precision (also called **Positive Predictive Value**)

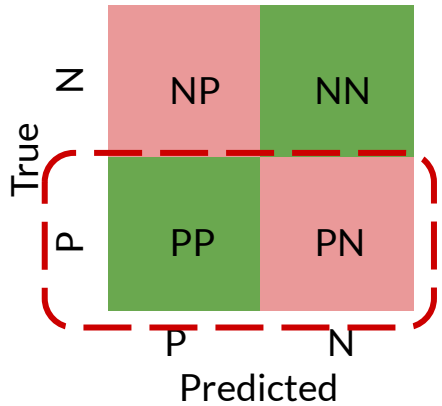
$$\text{Precision} = \frac{PP}{PP + NP}$$



Metrics

- Recall (also called **Sensitivity**)

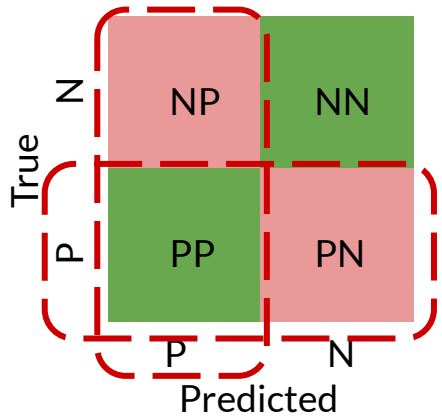
$$\text{Recall} = \frac{PP}{PP + PN}$$



Metrics

- F1 Score

$$\text{F1 Score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$



Data Imbalance Problem

- Demonstration through a practical example
- How to deal with data imbalance?
  - Through sampling bias
    - Undersampling
    - Oversampling
  - Redefining model loss function
    - Applying higher penalty to data points from the less occurring class.