Evaluating Classifiers

How do we know if a classifier is good?

- We want to compare our model to some baseline model
- We need to come up with a baseline model

How do we know which classifier is better?

- Accuracy (Ratio of correctly classified observations) - m.score in logistic regression
  - Can fail if we have very imbalanced classes
- What we see as a good classifier also depends on the preferences towards the result

Cancer Example

- Cancer Cases: 50
- Non-Cancer Cases: 950
- How many of the cancer cases do we correctly classify?:
- How many of the cancer cases do we incorrectly classify?:
- How many of the non-cancer cases do we correctly classify?:
- How many of the non-cancer cases do we incorrectly classify?:

Machine Learning Workflow

1. Define the Problem/Come up with a question (Hypothesis, Definition of Data that you need, Expected outcome ...)

2. Get data

3. Train-Test split

4. Exploratory Data Analysis

5. Feature Engineering

6. Train Model

7. Evaluate Model

   ...