Chapter 7 What is a Good Model?

12/28/20 3:04 PM

Evaluating Classifiers

1. *Accuracy*

$$\text{accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

- Problem 1: imbalance class (predicting only major class could still result a high accuracy)
- Problem 2: can't incorporate unequal costs and benefits

2. *Confusion Matrix*

	p	n
Y	True positives	False positives
N	False negatives	True negatives

Y & N (row indicators): prediction; p & n (column indicators): true classes

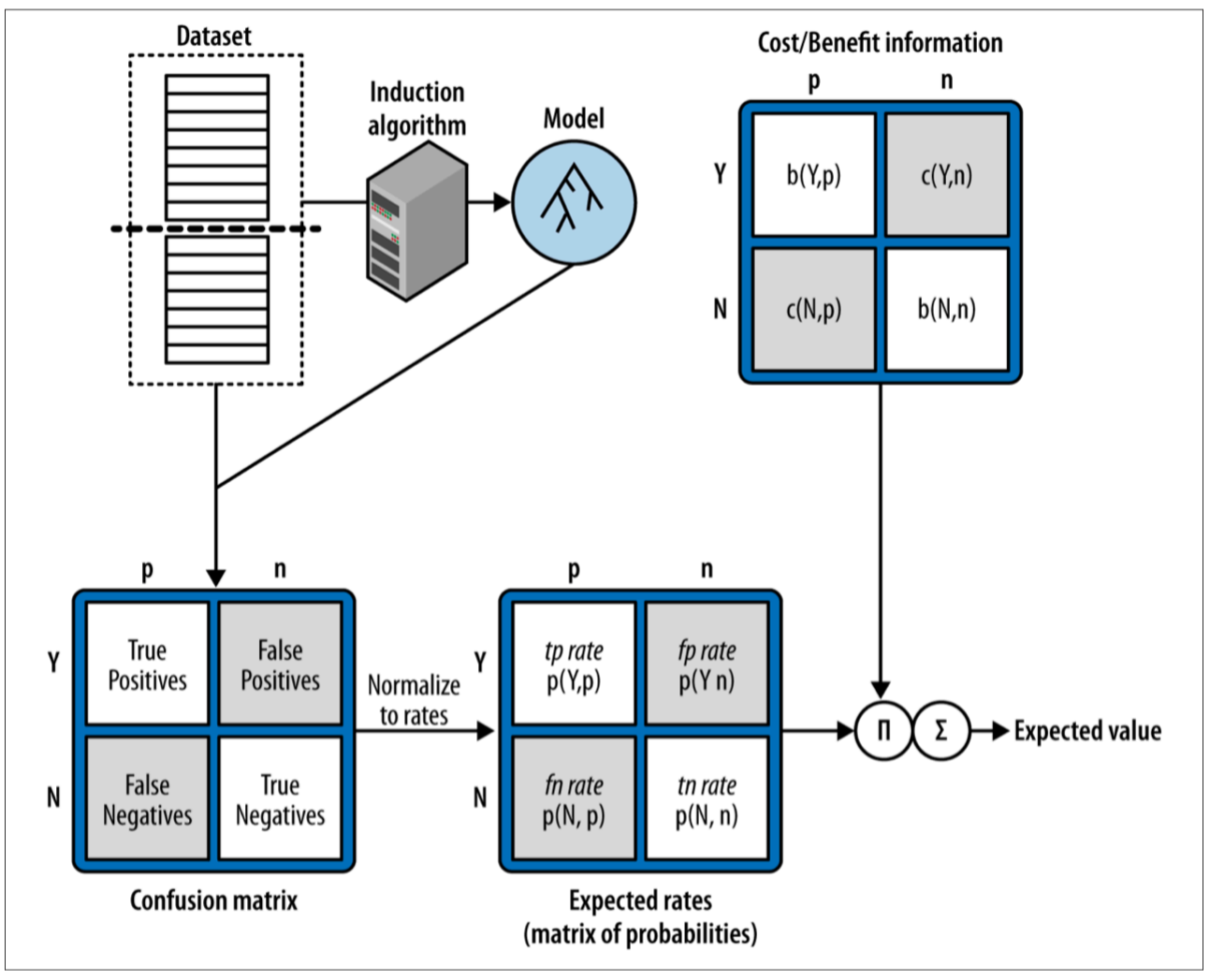
A key analytical framework: Expected Value

1. *To frame classifier us*

$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

- Each term: Probability of responding, value of responding, probability of no response and value of no response
- Usually value of no response would be minus
- We can use this formula to calculate threshold for p_r, by setting expected benefits = 0
- Then for all customers with probability of response higher than the threshold, it's worth reaching out.

2. *To frame classifier evaluation*



- Use confusion matrix to get expected rates (by normalizing)
- Create cost benefit matrix from external information (domain knowledge)
- $$Expected\ profit = p(\mathbf{Y}, \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N}, \mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + p(\mathbf{N}, \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y}, \mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n})$$
- Then decompose by Law of Conditional Probability:

$$\text{Expected profit} = p(\mathbf{p}) \cdot [p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + p(\mathbf{n}) \cdot [p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot c(\mathbf{Y}, \mathbf{n})]$$

- The conditional terms are TPR, FPR, etc.
- This form allows us to compare two models even with different base rate, just need to replace the priors of probability

3. *Side notes for formulating cost benefits matrix*

a. Make sure the signs are consistent (cost & profit)
b. An easy mistake is "double count" by putting benefit in one cell and negative cost in another. A good way to check is to calculate "benefit from improvement" for a certain instance

Baseline Performance

Some common ideas for choosing baseline models

1. Random model (random classification)
2. Majority classifier (for classification problem)
   Always choose the majority class in training set
3. Average predictor (for regression problem)
   Always predict the mean of training set
4. Models that consider only a very small amount of features
   E.g. decision stump (decision tree with only one internal node)
5. The idea can be extended to comparison of worth of data sources
6. Sometimes adopting domain knowledge alone can produce good models too
   E.g. increasing of usage of credit card to predict defraud