False Positives and False Negatives

In Machine Learning, analysts often treat every kind of error as equally bad

For example, suppose machine learning is being used to predict Pat's movie preferences. The classifier that is learned (e.g., a decision tree) predicts whether Pat will like a new movie or not. Over the next 100 movies, lets say there are 23 total errors

- 23 total errors / 100 predictions = 23% (overall) error rate or 77% (overall) accuracy

But lets break it down, because not all errors are equally bad. Call a "will like" prediction a "positive" prediction, and a "will not like" a "negative" prediction.

False Positives and False Negatives

Over the next 100 movies,

- the classifier predicts that Pat **will like** 37 of them (i.e., 37 positive predictions)

- the classifier predicts that Pat **will not like** 63 of them (i.e., 63 negative predictions)

False Positives and False Negatives

Over the next 100 movies,

- the classifier predicts that Pat **will like** 37 of them (i.e., 37 positive predictions)
  - 33 of these positive predictions are correct (i.e., true positives)
    - 33/37 = 0.89 is the proportion of true positives to positive predictions (*precision*)
  - 4 of these positive predictions are incorrect (i.e., false positives or Type 1 errors)
    - 4/37 = 0.11 is the proportion of false positives to predicted positives
- the classifier predicts that Pat **will not** like 63 of them (i.e., 63 negative predictions)
  - 51 of these negative predictions are correct (e.g., true negatives)
    - 51/63 = 0.81 is the proportion of true negatives to negative predictions
  - 19 of these negative predictions are incorrect (e.g., false negatives or Type 2 errors)
    - 19/63 = 0.19 is the proportion of false negatives to predicted negatives
- 23 total errors / 100 predictions = 23% (overall) error rate and 77% (overall) accuracy

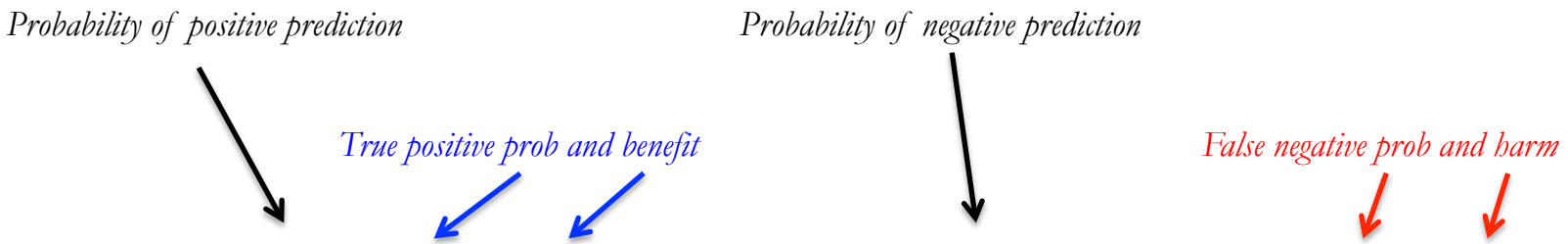
In movie prediction, what is the more harmful error – false positives or false negatives?

- What is the harm in a false positive (Prob = 0.11)? Perhaps it's a function of the following factors:
  - Pat will watch/start a movie that Pat does not like (e.g., with annoyance factor of -0.6)
- What is harm in a false negative (Prob = 0.19)? Perhaps it's a function of the following factors:
  - Pat will be unaware in the moment of a movie Pat would like
  - Pat has recommendations for plenty (?) of other movies that Pat will like
  - Pat may not be any the wiser about the overlooked movie (annoyance factor 0.0)
  - Pat will be annoyed if/when they later discover the movie (but probably won't associate with recommender system) (annoyance factor at recommender system -0.01)

  } quality of life factor

- What is the benefit of a true positive (Prob = 0.89)? Perhaps it's a function of the following factors:
  - Pat enjoys a movie (e.g., with QoL factor +0.92)
- What is the benefit of a true negative (Prob = 0.81)? Perhaps it's a function of the following factors:
  - Avoiding an annoyance (taken care of above)
  - Mitigating information overload (+0.02)

In movie prediction, how to judge classifiers based on mix of errors and correct predictions ?



= 0.37 * 0.75 + 0.63 * 0.02 = **0.29**

- A classifier with different proportions of positive and negative errors would vary in expected scores and thus in perceived quality
- Overall, in movie preference prediction it seems, false positives are more impactful than false negatives and true positives are more impactful of true negatives

- In predicting what brain tissue is normal (positive) and what is tumor (negative) in preparation for surgery, what errors (and correct prediction) types are more impactful?
- In predicting what past legal cases are relevant (positive) to a current case, what errors (and correct prediction) types more impactful?
    - For an AI that assists a legal team?
    - For an AI that is a lawyer?

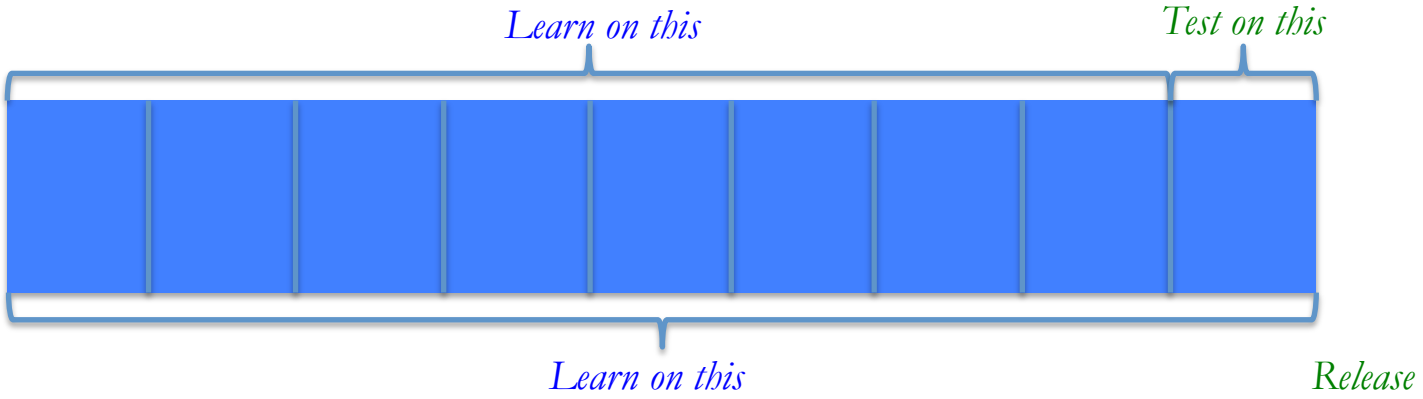
Evaluating an AI system

1. Utilities (benefits and harms) are determined by software developer team or by looking to see humans react in step 2

2. The various types of error (and accuracy) are measured on actual data (e.g., about movie preferences) and perhaps learned from reactions of humans on sample actual data too

3. How can we be sure of a product before its release?

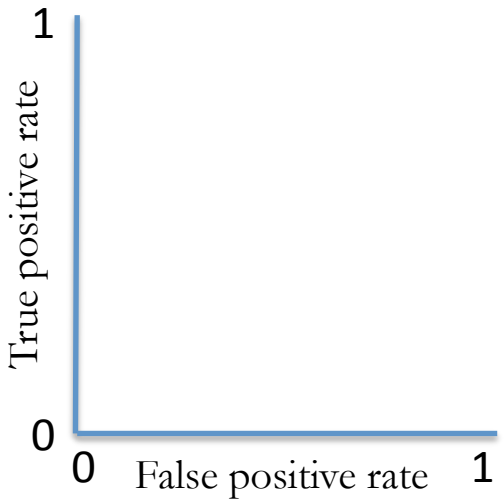
    a) If it is an expert system developed directly by software engineers and domain experts, then pretend" it was being used on a large collection of data, and compare the (new) AI system's pretend performance against the current (AI and/or human) system's performance
    b) If AI developed through machine learning, then use *cross validation*



More Lingo

- 33 of these positive predictions are correct (i.e., true positives)
  - 33/37 = 0.89 is the proportion of true positives to positive predictions (*precision*)
  - #tp / (#tp + #fp) is the *precision*
  - #tp / (#tp + #fn) is the *recall* (also *true positive rate*)
  - #fp / (#fp + #tn) is the *false positive rate*

ROC Space



Precision/Recall Space