# Lecture 9: Evaluating Classifiers

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

# Testing a supervised learning model

- How do we validate software?
  - Test suite of carefully selected inputs
  - Compare output with expected answers
- What about classification models?
  - By definition, deploy on data where the outcome is unknown
  - If expected answer available, have a deterministic solution, model not needed!
- On what basis can we evaluate a supervised learning model?

# Creating a test set

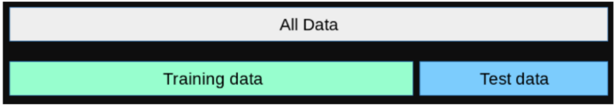
- Training data is labelled
  - No other source of inputs with expected answers
- Segregate some training data for testing
  - Terminology: training set and test set
  - Build model using training set, evaluate on test set
- Creating the test set
  - Need to choose a random sample
  - Can further use stratified sampling, preserve relative ratios (e.g., age wise distribution)
  - ML libraries can do this automatically

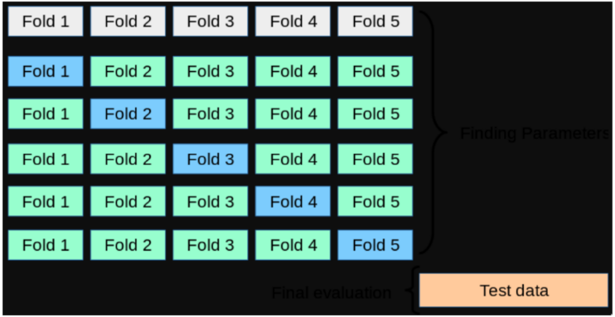
# Creating a test set

- How large should the test set be?
  - Typically 20-30% of labelled data
- Depends on labelled data available
  - Need enough training data to build the model

## Cross validation

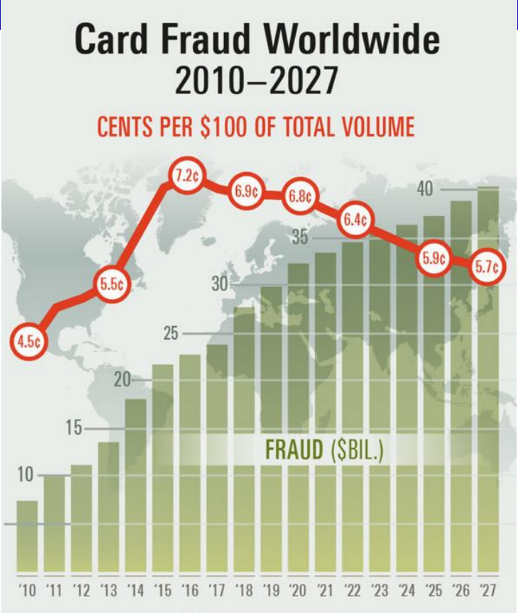
- Partition labelled data into $k$ chunks
- Hold out one chunk at a time
- Build $k$ models, using $k-1$ chunks for training, 1 for testing
- Useful if labelled data is scarce



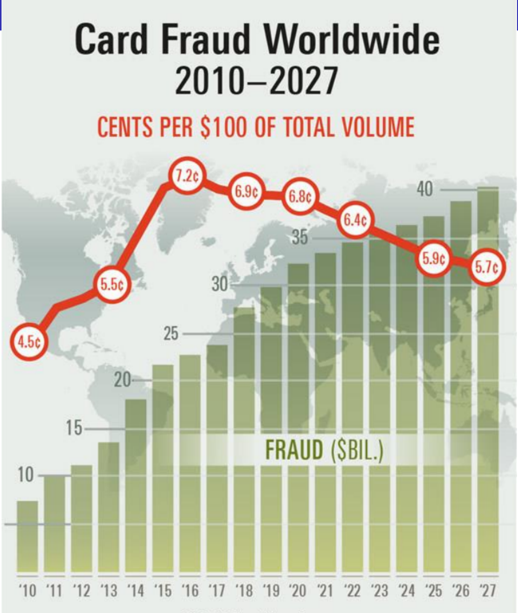
# What are we measuring?

- Accuracy is an obvious measure
  - Fraction of inputs where classification is correct
- Classifiers are often used in asymmetric situations
  - Less than 1% of credit card transactions are fraud
- "Is this transaction a fraud?"
  - Trivial classifier — always answer "No"
  - More than 99% accurate, but useless!



# Catching the minority case

- The minority case is the useful case
  - Assume question is phrased so that minority answer is "Yes"
  - Want to flag as many "Yes" cases as possible
- Aggressive classifier
  - Marks borderline "No" as "Yes"
  - False positives
- Cautious classifier
  - Marks borderline "Yes" as "No"
  - False negatives



# Confusion matrix

- Four possible combinations
  - Actual answer: Yes / No
  - Prediction: Yes / No
- Record all four possibilities in confusion matrix
  - Correct answers
    - True positives, true negatives
  - Wrong answers
    - False positives, false negatives

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | True Positive (TP) | False Negative (FN) |
| Actual negative | False Positive (FP) | True Negative (TN) |

# Performance measures

## Precision

- What percentage of positive predictions are correct?

$$\frac{\text{TP}}{\text{TP} + \text{FP}}$$

## Recall

- What percentage of actual positive cases are discovered?

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | True Positive (TP) | False Negative (FN) |
| Actual negative | False Positive (FP) | True Negative (TN) |

# Performance measures

- Precision 1, Recall 0.01

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | 1 | 99 |
| Actual negative | 0 | 900 |

# Performance measures

- Precision 1, Recall 0.01
- Recall up to 0.4, but precision down to 0.29

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | 40 | 60 |
| Actual negative | 100 | 800 |

# Performance measures

- Precision 1, Recall 0.01
- Recall up to 0.4, but precision down to 0.29
- Recall up to 0.99, but precision down to 0.165
- Precision-recall tradeoff
  - Strict classifiers : fewer false positives (high precision), miss more actual positives (low recall)
  - Permissive classifiers : catch more actual positives (high recall) but more false positives (low precision)

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | 99 | 1 |
| Actual negative | 500 | 400 |

# Performance measures

- Which measure is more useful?
  - Depends on situation
- Hiring
  - Screening test: high recall
  - Interview: high precision
- Medical diagnosis
  - Immunization: high recall
  - Critical illness diagnosis: high precision

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | True Positive (TP) | False Negative (FN) |
| Actual negative | False Positive (FP) | True Negative (TN) |

# Performance measures

## Other measures, terminology

- Recall is also called sensitivity
- Accuracy: (TP+TN)/(TP+TN+FP+FN)
- Specificity: TN/(TN+FP)
- Threat score: TP/(TP+FP+FN)
  - TN usually majority, ignore, not useful

| | Classified positive | Classified negative |
|---|---|---|
| Actual positive | True Positive (TP) | False Negative (FN) |
| Actual negative | False Positive (FP) | True Negative (TN) |

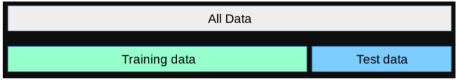
## F Score

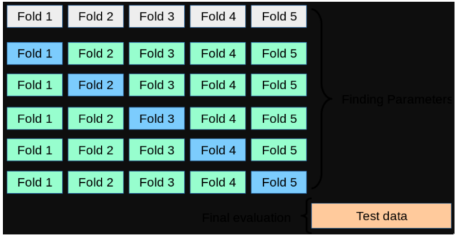
- A single combined score
- Harmonic mean of precision, recall

$$\frac{2pr}{p+r}$$

# Summary

- Need to carve out a test set to evaluate a classifier
- Can use cross-validation if labelled data is scarce
- Accuracy is not a very useful metric — categories are asymmetric
- Confusion matrix captures different types of correct and wrong answers
- Precision and recall are most commonly used measures
- Tradeoff one for the other based on the situation




| | |
|---|---|
| TP | FN |
| FP | TN |