CS 456 Data Mining

Central Washington University

Today …

- Confusion Matrix
- Cross-Validation

Metrics

- We train on our training data Train = $\{x_i, y_i\}_{1,m}$
- We test on Test data.
- We often set aside part of the training data as a development set, especially when the algorithms require tuning.
  - In the Project we asked you to present results also on the Training; why?
- When we deal with binary classification we often measure performance simply using Accuracy:

$$\text{accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{test instances}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\#\ \text{incorrect predictions}}{\#\ \text{test instances}}$$

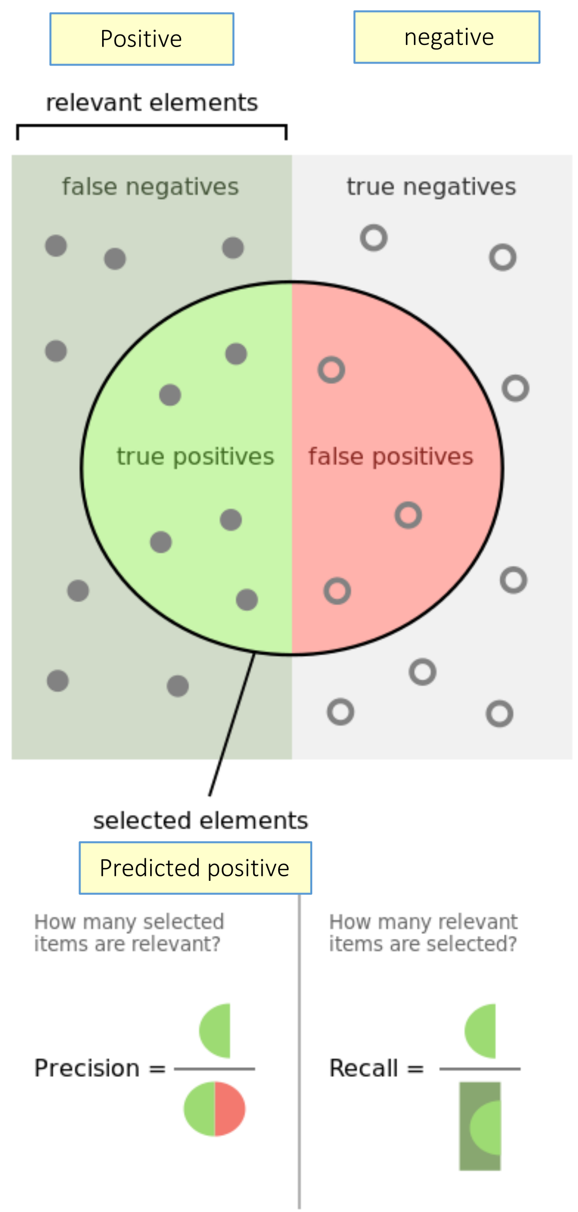
- Any possible problems with it?

Alternative Metrics

- If the Binary classification problem is biased
  - In many problems most examples are negative
- Or, in multiclass classification
  - The distribution over labels is often non-uniform
- Simple accuracy is not a useful metric.
  - Often we resort to task specific metrics
- However one important example that is being used often involves Recall and Precision

- **Recall:** $\frac{\textbf{\# (positive identified = true positives)}}{\textbf{\# (all positive)}}$

- **Precision:** $\frac{\textbf{\# (positive identified = true positives)}}{\textbf{\# (predicted positive)}}$



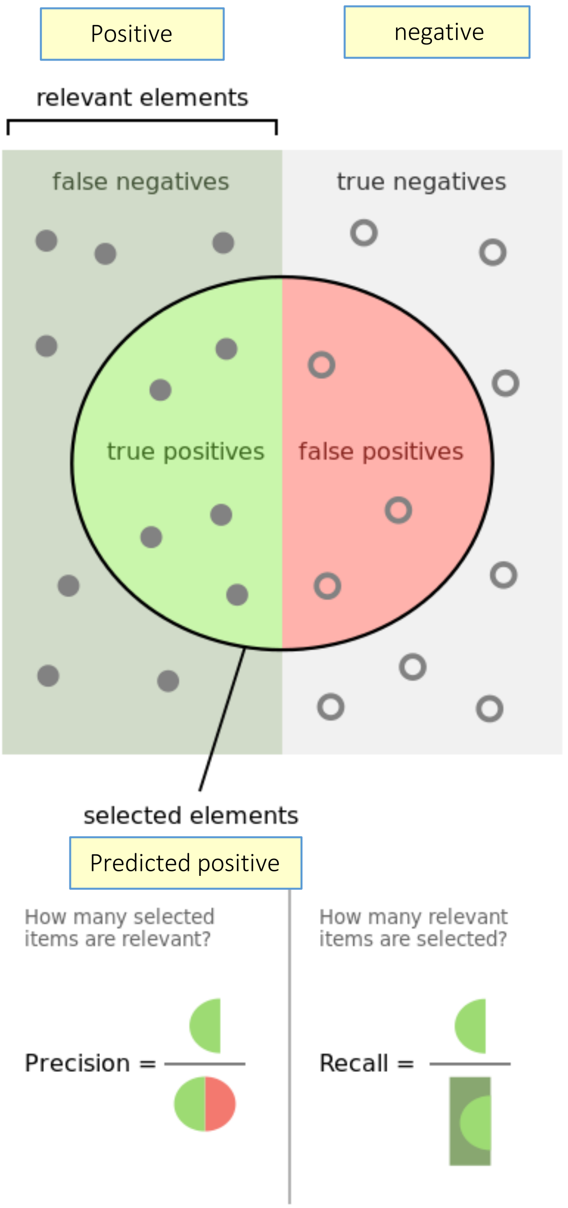
Example

- 100 examples, 5% are positive.

- Just say NO: your accuracy is 95%
  - Recall = precision = 0
- Predict 4+, 96-; 2 of the +s are indeed positive
  - Recall:2/5; Precision: 2/4

- **Recall:** $\dfrac{\textbf{\# (positive identified = true positives)}}{\textbf{\# (all positive)}}$

- **Precision:** $\dfrac{\textbf{\# (positive identified = true positives)}}{\textbf{\# (predicted positive)}}$



Confusion Matrix

- Given a dataset of P positive instances and N negative instances:

Actual Class \ Predicted Class	Yes	No
Yes	TP	FN
No	FP	TN

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

- Imagine using classifier to identify positive cases (i.e., for information retrieval)

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

Probability that a randomly selected positive prediction is indeed positive

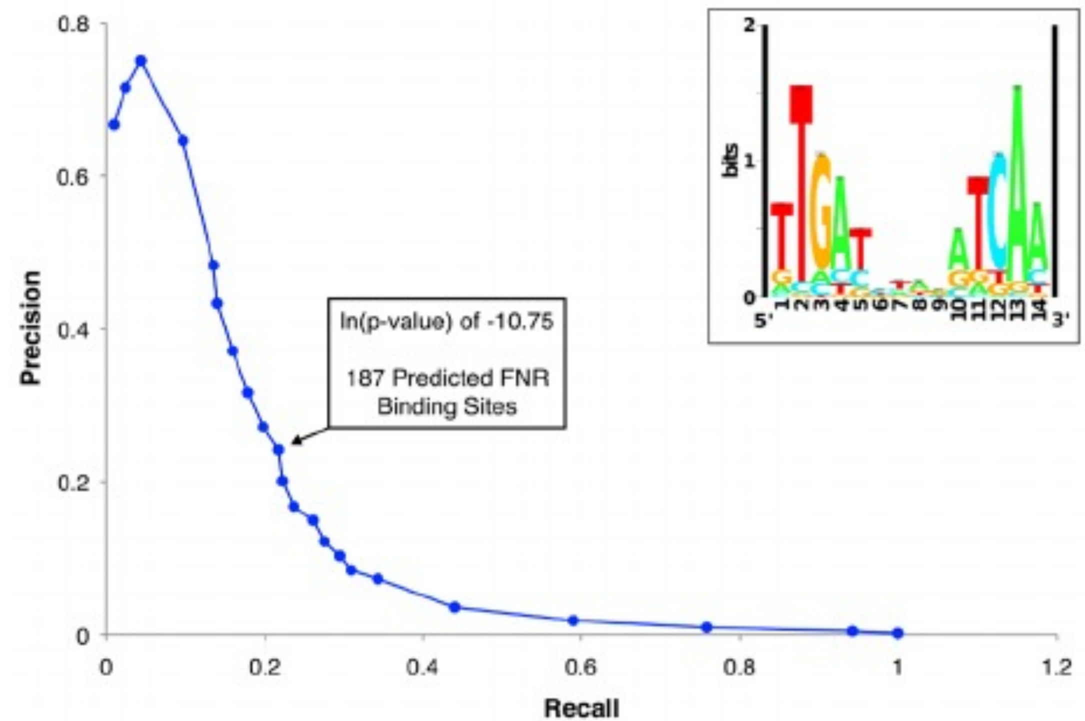
$$\text{recall} = \frac{TP}{TP + FN}$$

Probability that a randomly selected positive is identified

The notion of a confusion matrix can be usefully extended to the multiclass case
(i,j) cell indicate how many of the i-labeled examples were predicted to be j

Relevant Metrics

- It makes sense to consider Recall and Precision together or combine them into a single metric.
- Recall-Precision Curve:
- F-Measure:
  - A measure that combines precision and recall is the harmonic mean of precision and recall.
  - F1 is the most commonly used metric.



$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

Comparing Classifiers

Say we have two classifiers, *C1* and *C2*, and want to choose the best one to use for future predictions

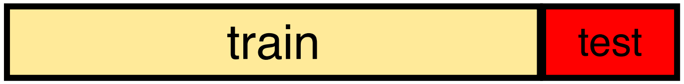
Can we use training accuracy to choose between them?

- No!

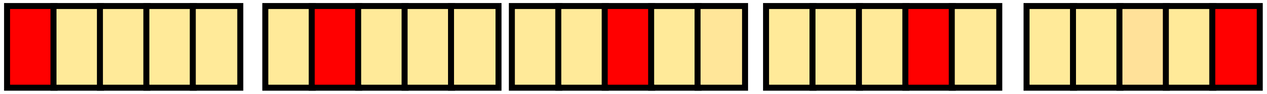
- What about accuracy on test data?
- Yes, but…
  - We basically want to look at more than a single number; gather some statistical evidence.

N-fold cross validation

- Instead of a single test-training split:

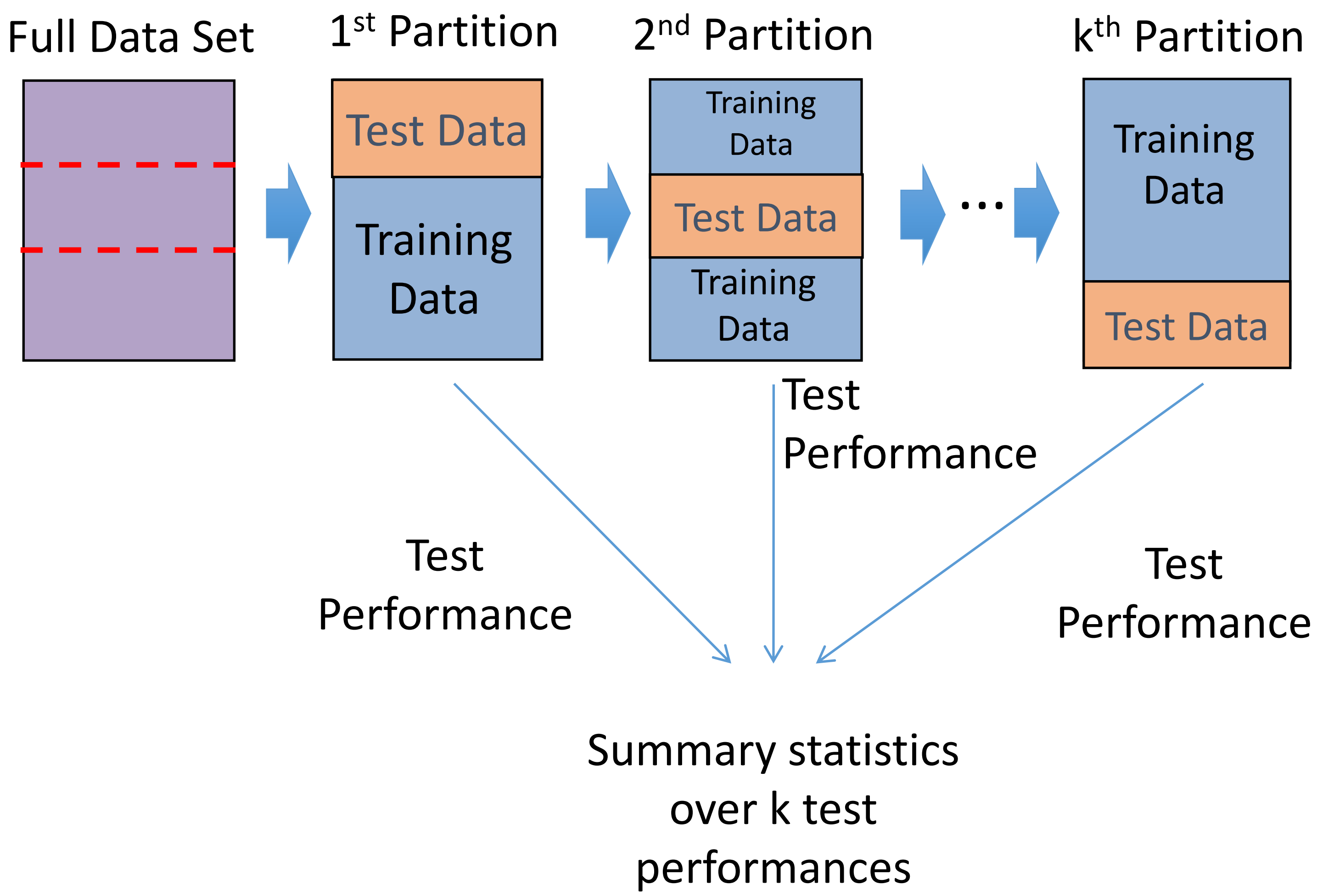


- Split data into N equal-sized parts



- Train and test N different classifiers
- Report average accuracy and standard deviation of the accuracy

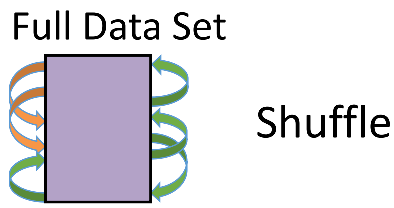
Example 3-Fold CV



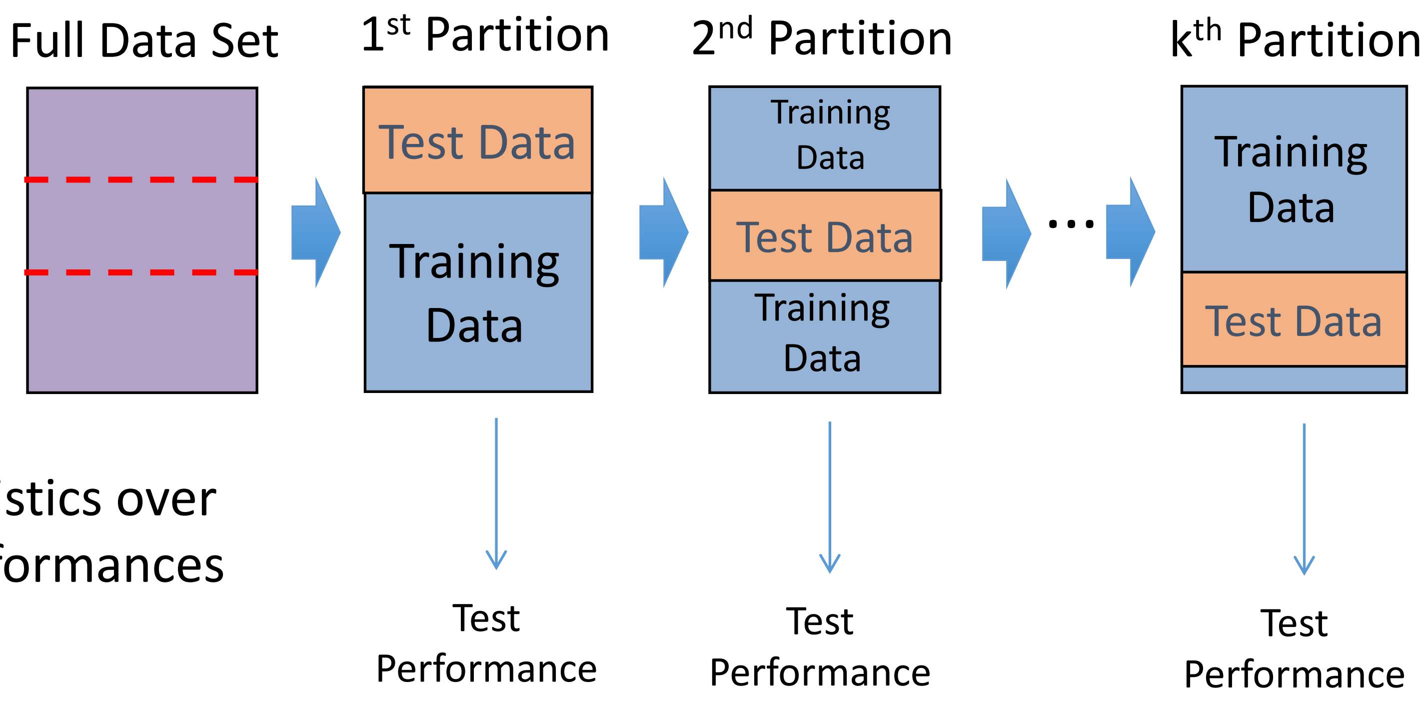
Multiple Trials of k-Fold CV

1.) Loop for t trials:

a.) Randomize Data Set



b.) Perform k-fold CV

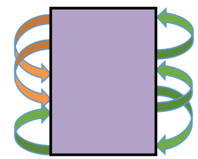


2.) Compute statistics over $t \times k$ test performances

Multiple Trials of k-Fold CV

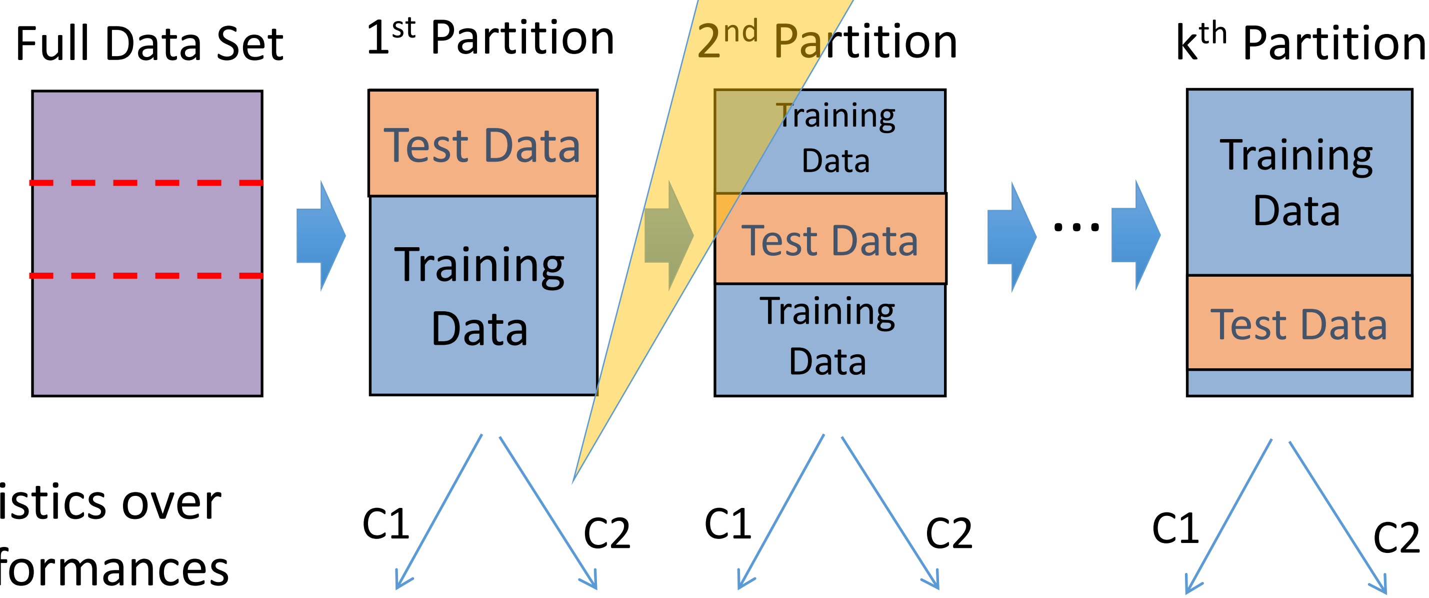
1.) Loop for t trials:

a.) Randomize Data Set



Test each candidate learner on same training/testing splits

b.) Perform k-fold CV



2.) Compute statistics over $t \times k$ test performances

Allows us to do paired summary statistics (e.g., paired t-test)