MODEL OVERFITTING

SYRACUSE UNIVERSITY
School of Information Studies

MODEL EVALUATION

Topics:

Model overfitting

Model evaluation methods and metrics

Model comparison and selection

Reproducible research

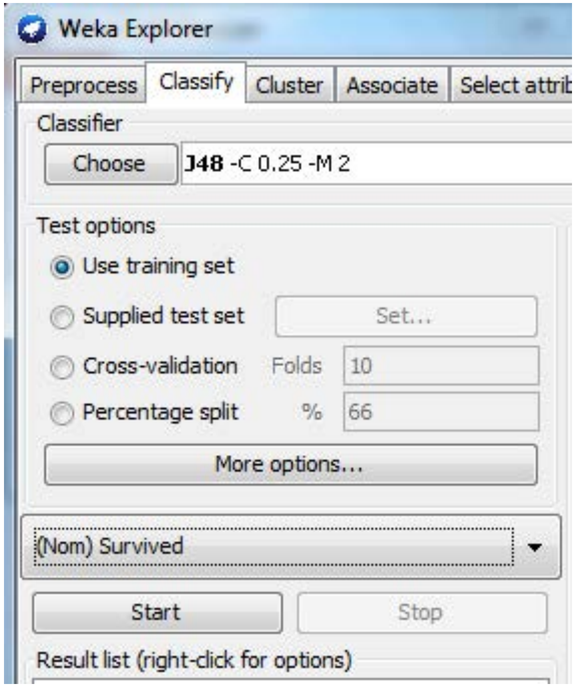
MODEL OVERFITTING

Two fundamental concepts

Training error: Train a model (e.g., a decision tree) on a training set, then test the model on the same training set. The error rate is called “training error,” which measures how well the model fits the training data.

Test error: Test the model on a test set that is different from the training set. The error rate is called “test error,” which measures how well the model generalizes to new, unseen data.

TRAINING ERROR VS. TEST ERROR



Classifier

J48 -C 0.25 -M 2

Test options

● Use training set

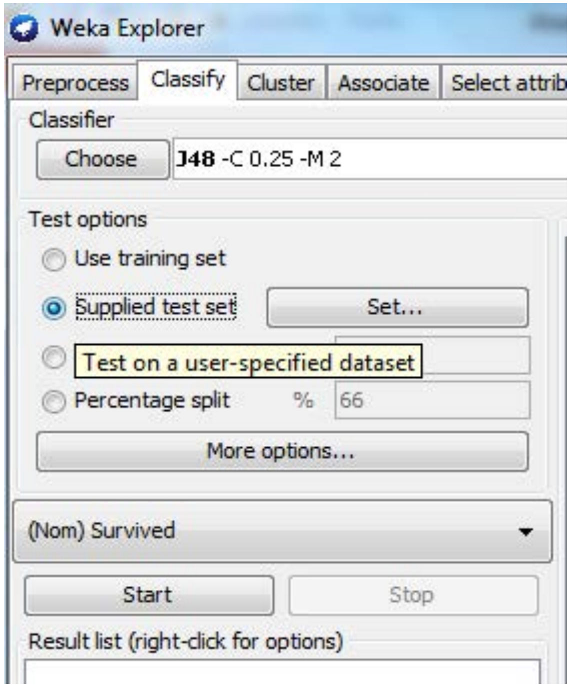
○ Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Nom) Survived

Result list (right-click for options)



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

● Supplied test set

○

○ Percentage split % 66

(Nom) Survived

Result list (right-click for options)

Weka: The evaluation option to obtain training error

Weka: The evaluation option to obtain test error

MODEL OVERFITTING (CONT.)

Overfitting means a model fits the training data very well but generalizes to unseen data poorly.

Therefore, if the test error is much higher than training error, the model is more likely to be overfitting.

MODEL OVERFITTING

SYRACUSE UNIVERSITY
School of Information Studies

MODEL COMPLEXITY AND OVERFITTING

Complex models are more likely to overfit than simple models.

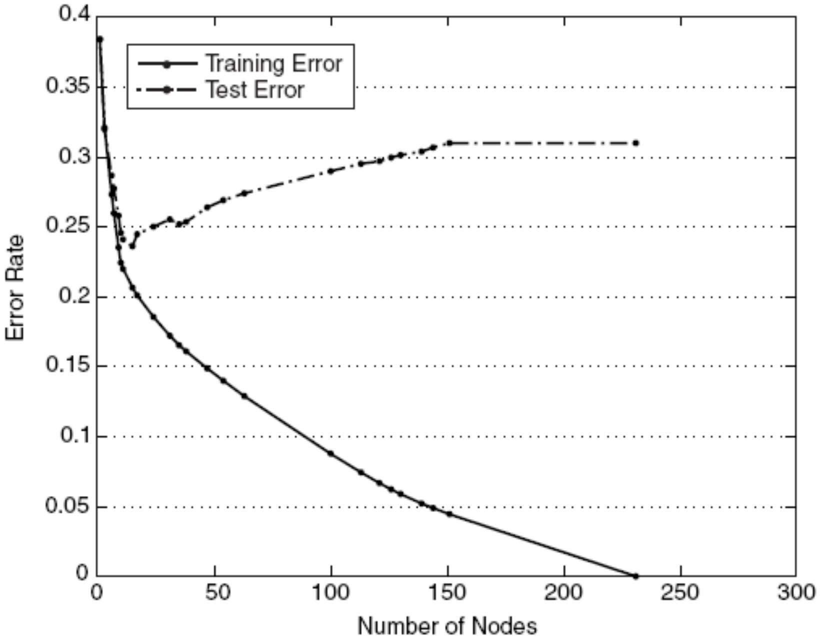
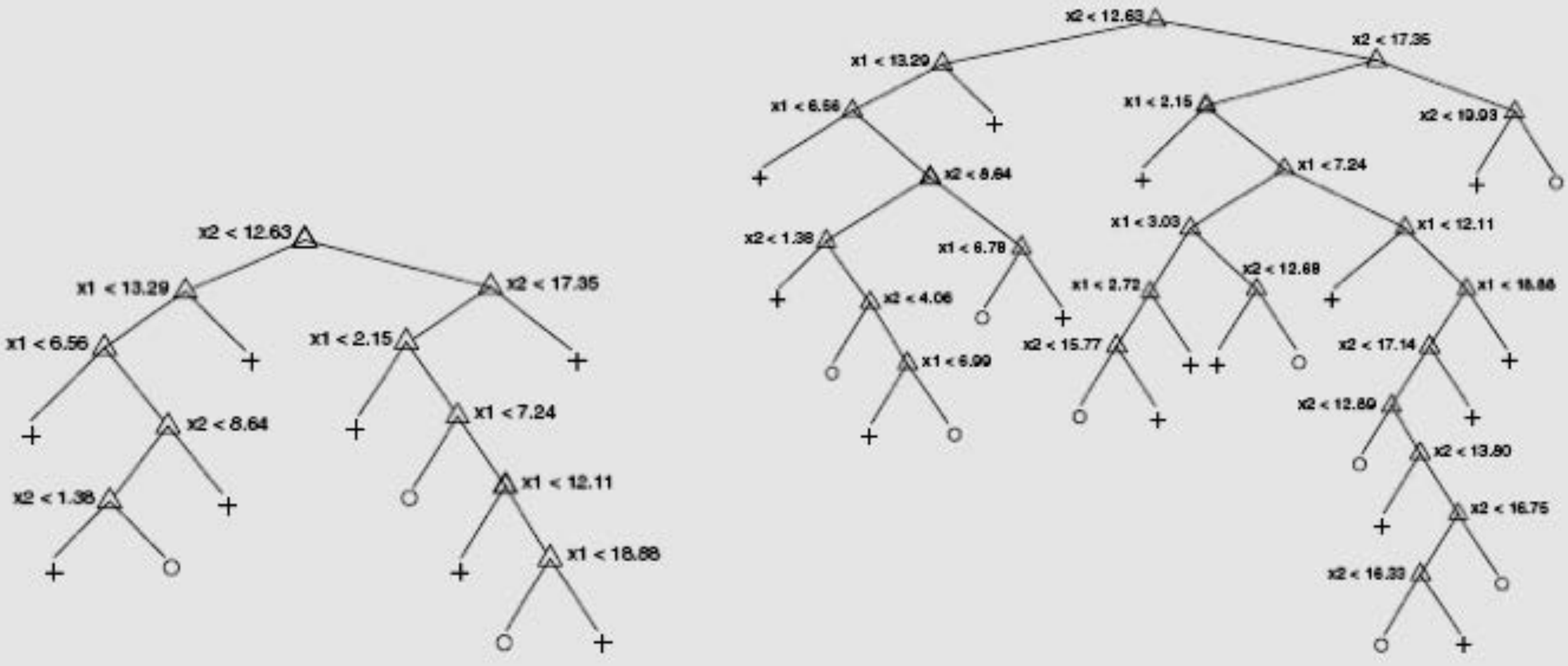


Figure 4.23. Training and test error rates.

For decision tree, number of nodes indicates model complexity.

Higher number of nodes -> higher model complexity -> lower training error and higher test error



(a) Decision tree with 11 leaf nodes.

(b) Decision tree with 24 leaf nodes.

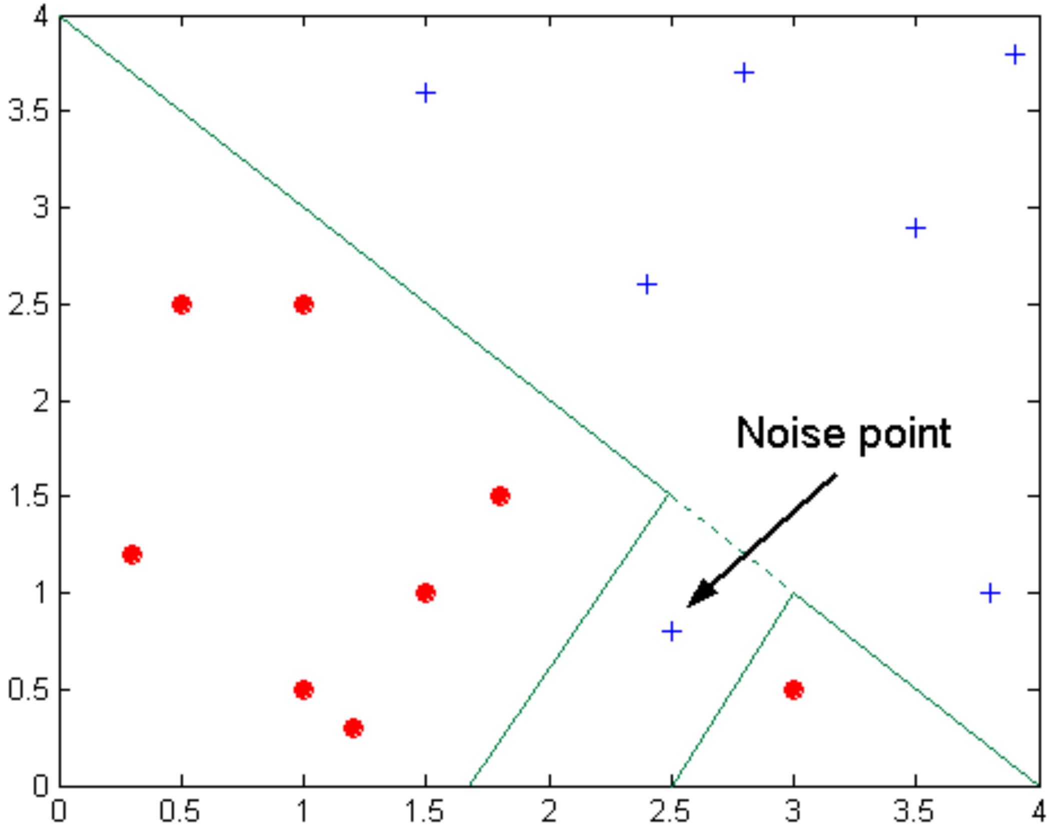
Figure 4.24. Decision trees with different model complexities.

MAIN REASONS FOR MODEL OVERFITTING

Overfitting due to noise

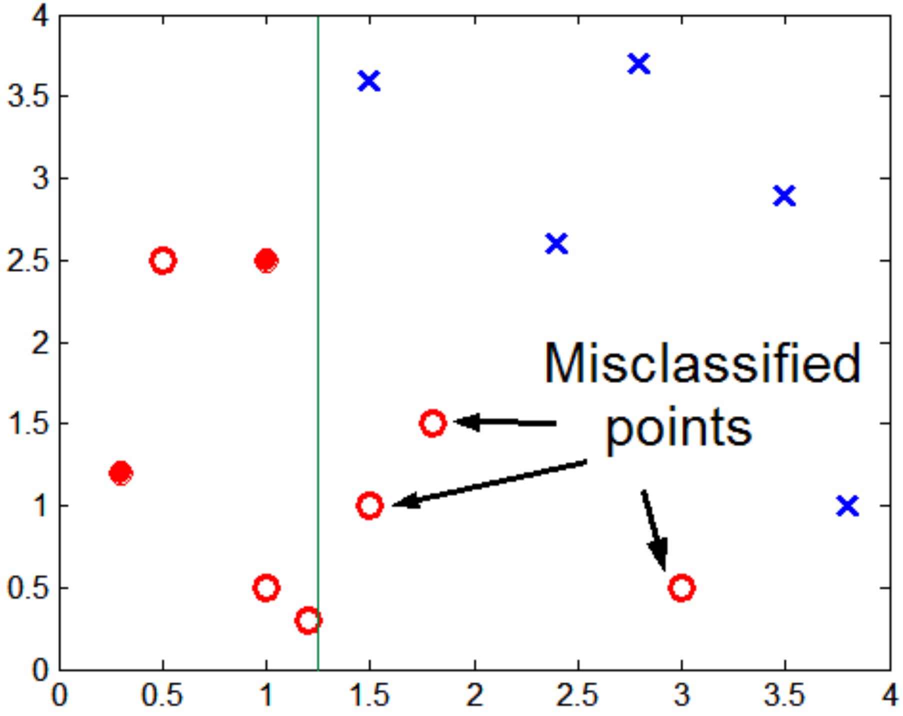
Overfitting due to insufficient samples

OVERFITTING DUE TO NOISE



The decision boundary (supposedly a straight line) is distorted by the noise point. The overfitted decision boundary is indicated by the solid blue lines.

OVERFITTING DUE TO INSUFFICIENT EXAMPLES



Blue crosses and solid red dots are training data.

Red circles are test data.

The green vertical line is the decision boundary created by a simple decision tree (if x > 1.25, label = blue; otherwise, label = red).

Lack of data points in the lower half of the diagram makes it difficult to predict correctly the class labels in that region.

OCCAM'S RAZOR

Given two models of similar generalization errors, the simpler model is preferred over the more complex model.

For a complex model, there is a greater chance that it was overfitted accidentally by errors in data or data imbalance.

Therefore, model complexity should be considered when evaluating a model.

MODEL EVALUATION METHODS

SYRACUSE UNIVERSITY
School of Information Studies

MODEL EVALUATION METHODS

What methods can measure model fitness before using it in real predictions?

Some evaluation methods have been designed to test the model on training data while controlling model overfitting.

- Hold-out test
- Cross-validation

HOLD-OUT TEST

Split the training data to two subsets, using one subset for training and the other for testing.

- The splitting ratio is determined by the training set size in that both subsets cannot be too small.

50/50 or 2:1 are common splitting ratios.

Advantage: Fast

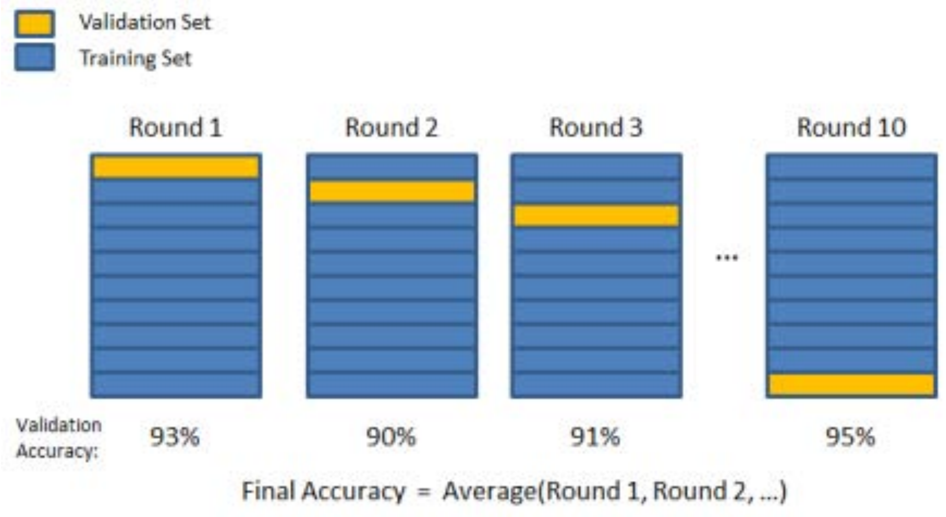
Shortcoming: When the split changes, the test result changes too

- High variability in the test result

CROSS-VALIDATION (CV)

N is determined by the training set size. The larger the N, the longer it takes to run the experiment.

Five and 10 are common choices for N.



http://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

LEAVE ONE OUT

An extreme case of cross-validation

- N equals the training set size S

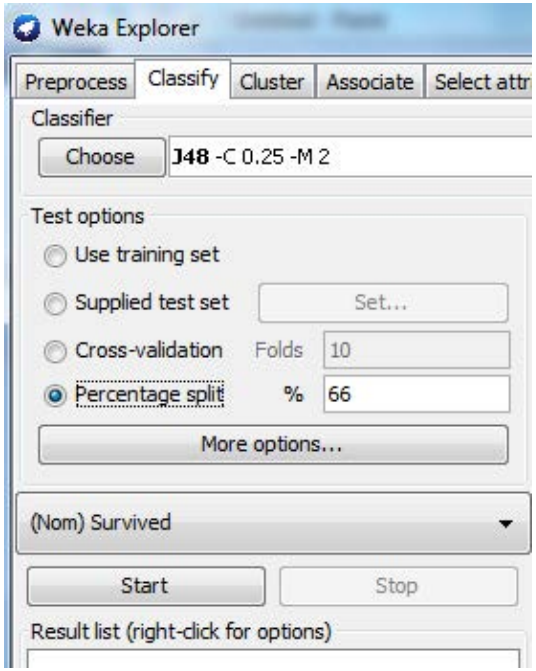
Advantage

- No variability in the test result (always get the same result)

Problems

- The most time-consuming method
- Usually used on very small data sets

HOLD-OUT TEST VS. CROSS-VALIDATION



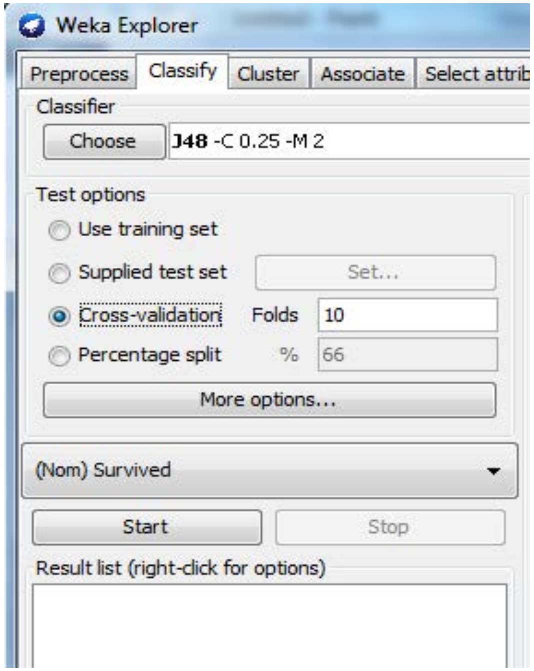
Classifier
J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) Survived

Result list (right-click for options)

Weka test option for hold-out test



Classifier
J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Survived

Result list (right-click for options)

Weka test option for cross-validation

HOLD-OUT TEST VS. CROSS-VALIDATION

Hold-out test

Pro: Fast

Con: High variability in the result, depending on the split

Cross-validation

Pro: Less variability and thus more reliable error estimation

Con: Takes longer time

WHICH MODEL EVALUATION METHODS TO CHOOSE?

CV is the standard method.

When data set is huge, hold-out test can save time.

When data set is small, leave one out can be considered.

MODEL EVALUATION METRICS

SYRACUSE UNIVERSITY
School of Information Studies

METRICS FOR MODEL PERFORMANCE

Accuracy is the most common measure, but it has limitations, especially on skewed data set.

Data set with similar number of examples in each category is "balanced," otherwise "unbalanced" or "skewed."

Titanic training data set is skewed, with more negative examples than positive ones.

549 "0": Did not survive

342 "1": Survived

PROBLEM WITH ACCURACY MEASURE

We need to learn some fundamental concepts first:

Confusion matrix for two classes (can be extended to multiple classes)

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

ACCURACY DEFINITION BASED ON CONFUSION MATRIX

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

LIMITATION OF ACCURACY

Consider a two-class problem:

- Number of Class 0 examples = 9,990
- Number of Class 1 examples = 10

If a model predicts every test example as “0,” the model’s accuracy is 9,990/10,000 = 99.9 %.

- Accuracy is misleading because the trivial model does not detect any Class 1 example.

TWO TYPES OF ERROR

Market analysis: To predict if a student is going to buy new computer or not.

Prediction result in a confusion matrix:

	Predictions		
Classes	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6,000	1,000	7,000
buy_computer = no	500	2,500	3,000
Total	6,500	3,500	10,000

False negative: Missed customers

False positive: Wrong targets

WHICH TYPE OF ERROR MATTERS MORE?

For a company, one type of error might be more costly than the other.

- E.g., one would rather send out more coupons than miss a potential buyer.
- E.g., one would rather tolerate some junk mail in inbox than risk misclassify a regular mail to junk.

The accuracy measure does not differentiate these two types of errors, but precision and recall would do.

PRECISION AND RECALL

Concepts borrowed from the information retrieval field

Define precision and recall on each category

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

PRECISION

$$\text{Precision}_{\text{class=yes}} = \frac{a}{a+c} = \frac{TP}{TP+FP}$$

Meaning: Among all positive predictions, how many are correct?

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

RECALL

$$\text{Recall}_{\text{class=yes}} = \frac{a}{a+b} = \frac{TP}{TP+FN}$$

Meaning: Among all positive examples, how many are correctly predicted?

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

EXAMPLE: CALCULATE PRECISION AND RECALL

	Predictions			
Classes	buy_computer = yes	buy_computer = no	Total	Recall(%)
buy_computer = yes	6,000	1,000	7,000	6,000/7,000
buy_computer = no	500	2,500	3,000	2,500/3,500
Total	6,500	3,500	10,000	
Precision (%)	6,000/6,500	2,500/3,500		

F-MEASURE

An ideal model would achieve high precision and recall on all categories.

But in reality, precision and recall are like the two sides of a seesaw: If one goes up, the other might go down.

F-measure is a weighted average of precision and recall.

$$\mathrm{F}_{\text{class=yes}} = \frac{2 \cdot \ precision \cdot \ recall}{precision + recall}$$

MODEL COMPARISON

SYRACUSE UNIVERSITY
School of Information Studies

BASELINES FOR MODEL EVALUATION

If your classification model reached 80% accuracy, is it "good enough"?

Two common baselines for comparison

- **Random guess**: If there are two categories, a model based on random guess would result in 50% accuracy.
- **Majority vote**: If the data set is skewed, a trivial model would assign all test data to the larger category.
  - In the Titanic training data set, the majority vote model would result in 549/891 = 62% accuracy.

Your model is expected to outperform the common baselines.

MAJORITY VOTE BASELINE

	Predictions			
Classes	buy_computer = yes	buy_computer = no	Total	Recall(%)
buy_computer = yes	7,000	0	7,000	1
buy_computer = no	3,000	0	3,000	0
Total	10,000	0	10,000	
Precision (%)	.70	n/a		

FAIR COMPARISON

When comparing the performance of two models, e.g., an unpruned tree vs. a pruned tree, make sure the comparison is fair, meaning, the test data should be exactly the same.

Common mistakes:

Run hold-out test on one model but cross-validation on another model.

Set up different numbers of folds for the two models when using cross-validation.

Set up different split ratio for the two models when using hold-out test.

OTHER ASPECTS OF EVALUATION

When comparing two classification models, predictive capability (as measured by accuracy, precision, recall, etc.) is only one aspect to examine.

Other aspects:

Speed

Robustness

Scalability

Model interpretability

OTHER ASPECTS OF EVALUATION

Speed

Time to construct model (training time)

Time to use the model (classification/prediction time)

Robustness

Handling noise and missing values

Scalability

The data set size keeps increasing

Interpretability

Understanding the insight provided by the model

IS THE MODEL GOOD ENOUGH?

There is always room for improvement for nontrivial prediction tasks.

Evaluation from system perspective

Evaluation from user perspective

TRAINING SET SIZE

SYRACUSE UNIVERSITY
School of Information Studies

TRAINING DATA SIZE AFFECTS ACCURACY

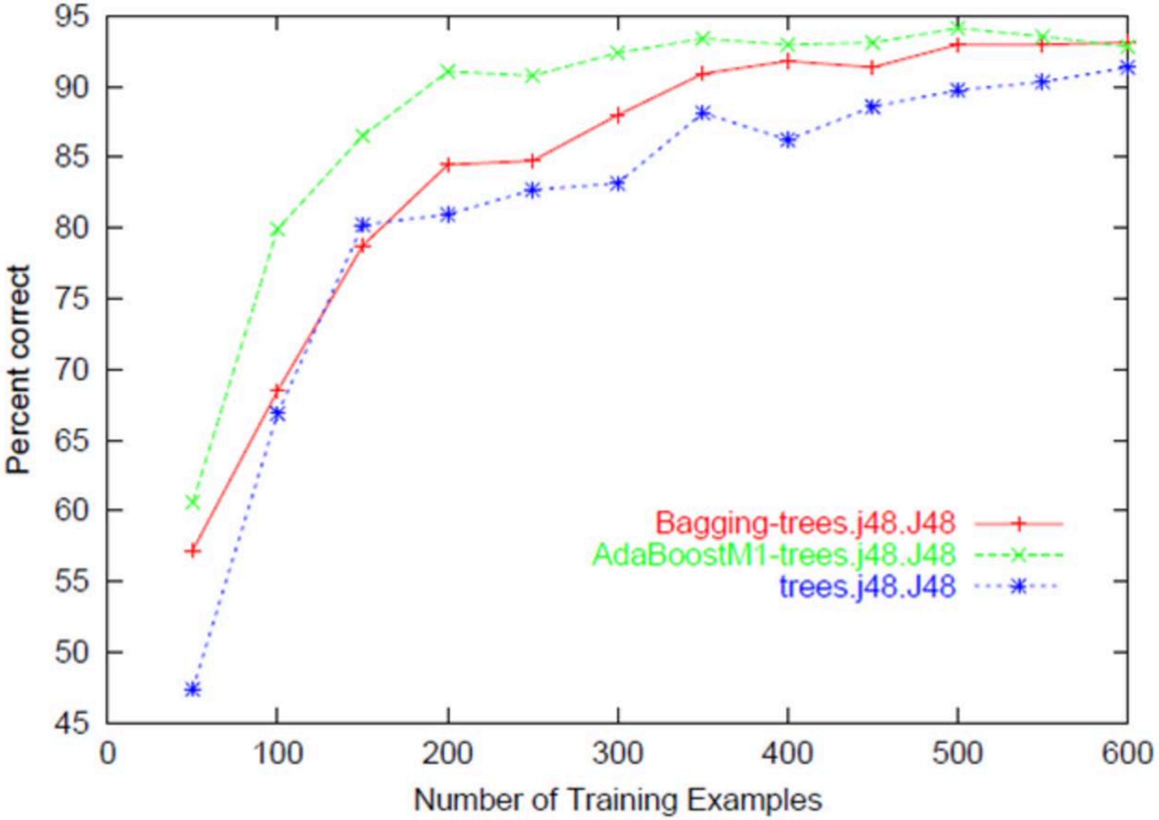
Larger training data set usually helps improve the model, but not always.

- Data saturation
- Noise in data

How many is "enough"?

- Depends on many factors, e.g., data availability, cost to obtain data, data quality

LEARNING CURVE



http://stackoverflow.com/questions/4617365/what-is-a-learning-curve-in-machine-learning

TRAINING DATA ACQUISITION

SYRACUSE UNIVERSITY
School of Information Studies

NOT ENOUGH DATA?

Semi-supervised learning

Active learning

Crowdsourcing

SEMI-SUPERVISED LEARNING

Utilize the strength of current model.

Assume the most confident predictions are highly accurate.

Process:

Build model on current training data.

Apply model to test data.

Rank test data by prediction confidence.

Add the most confident ones into training data.

ACTIVE LEARNING

Goal: Adding data to reduce current model's weakness

Also rank test data by prediction confidence

Choose the least confident ones

Confirm these predictions with human experts

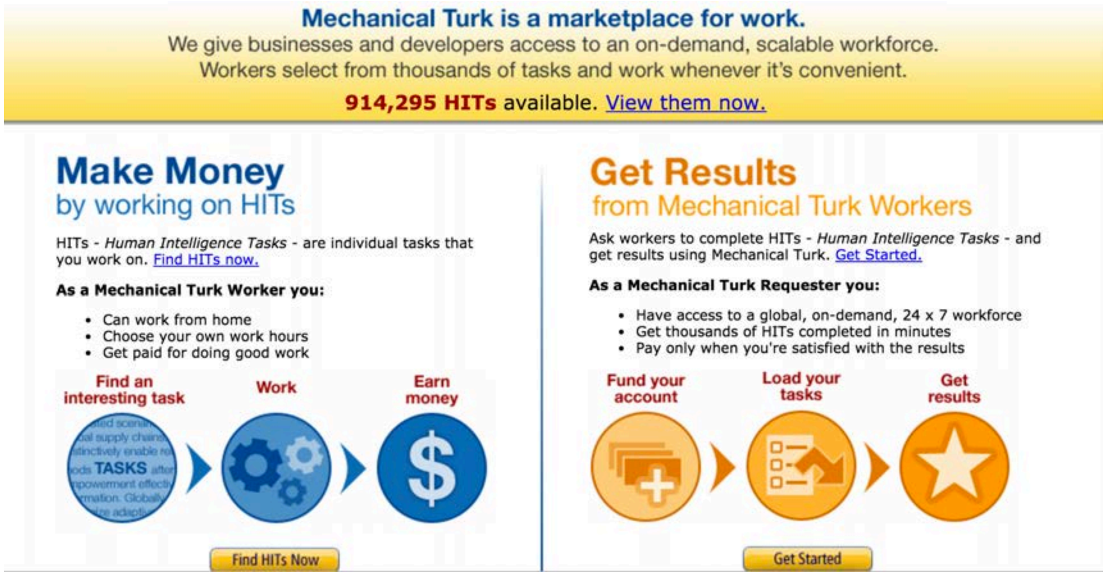
Add them to training data

CROWDSOURCING

Divide and conquer

Ask many people to each label a few examples for you.

Amazon Mechanical Turk



Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

914,295 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

HOW TRUSTWORTHY IS HUMAN ANNOTATION?

- Reliability test
  - If asking two or more people to mark the sentiment of a collection of tweets, to what extent will they agree with each other?

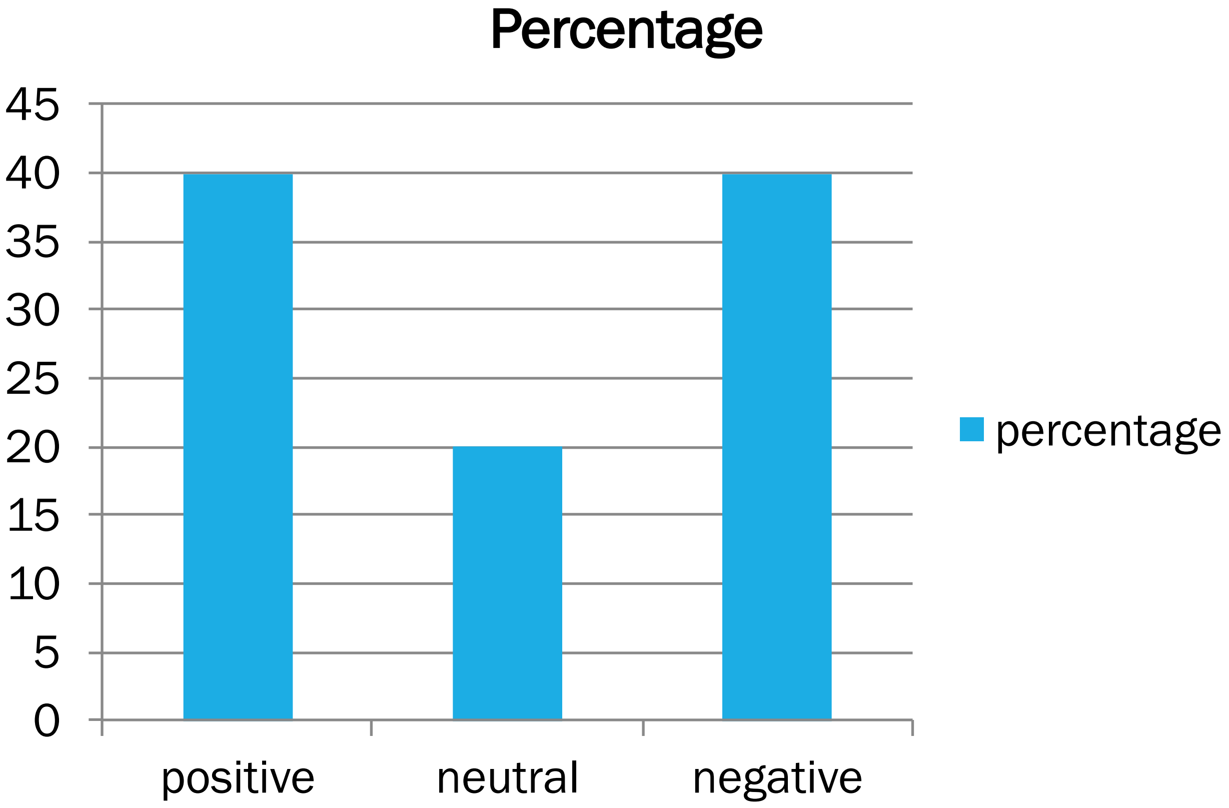
SUBJECTIVITY IN CLASSIFICATION

Some classification tasks involve a certain level of subjectivity in decision.

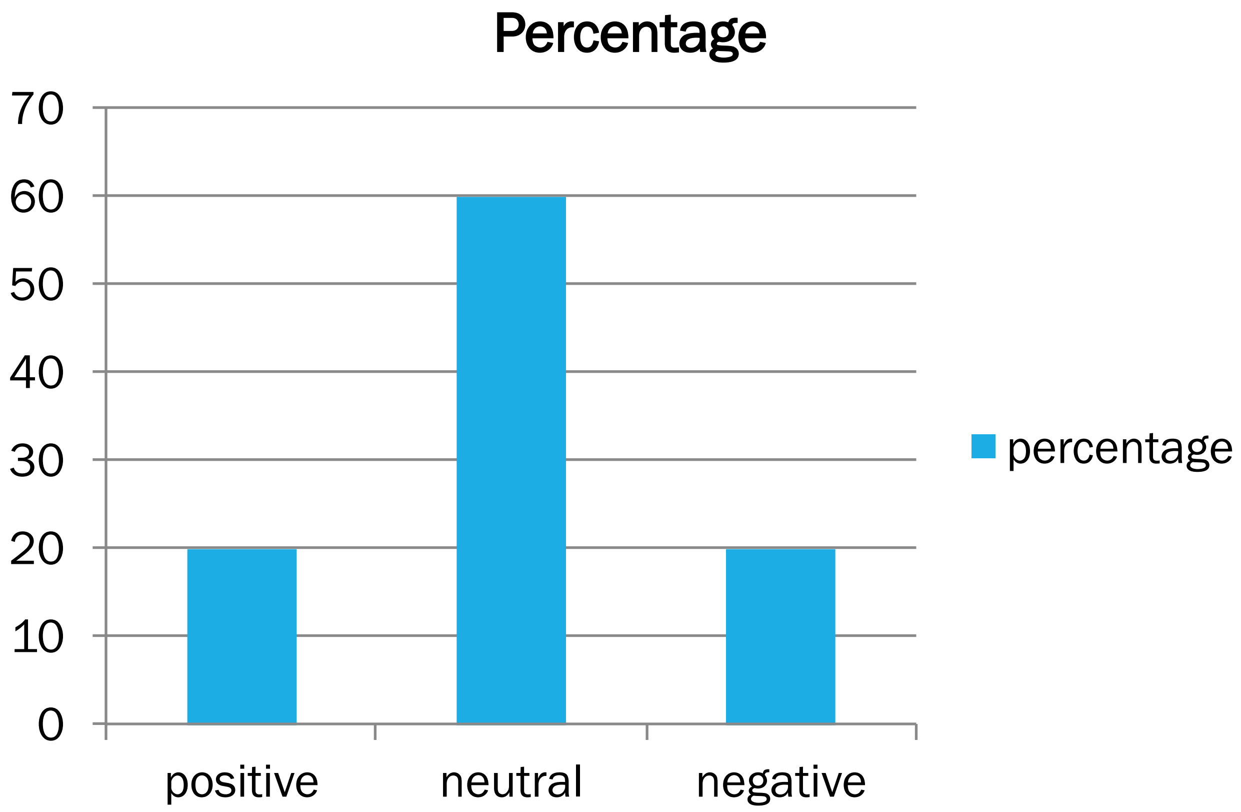
Whether a tweet is positive or neutral can be a subjective decision.

Different people may annotate the same tweet with different labels, e.g., “positive,” “neutral.”

A “POLARIZED” CODER



A "NEUTRAL" CODER



INTERCODER AGREEMENT

Measures to evaluate the reliability of human annotation

Percentage of agreement

Cohen’s kappa

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Po: Observed agreement
Pe: Chance of agreement

INTERCODER AGREEMENT

Raw agreement:

a = count(agreed_items)/total_items

Problem with raw agreement

Skewed categories: 90% raw agreement in both tables

	Coder A		
Coder B		positive	negative
	positive	45	5
	negative	5	45

	Coder A		
Coder B		positive	negative
	positive	90	10
	negative	0	0

COHEN'S KAPPA

a = raw_agreement

c = chance_agreement

$K = (a - c)/(1 - c)$

	Coder A		
Coder B		positive	negative
	positive	45	5
	negative	5	45

	Coder A		
Coder B		positive	negative
	positive	90	10
	negative	0	0

COHEN’S KAPPA

a = raw_agreement

c = chance_agreement

$K = (a - c)/(1 - c)$

	Coder A		
Coder B		positive	negative
	positive	45	5
	negative	5	45

	Coder A		
Coder B		positive	negative
	positive	90	10
	negative	0	0

HOW TO CALCULATE KAPPA

Given a confusion matrix of two coders:

	Coder A		
Coder B		positive	negative
	positive	45	5
	negative	5	45

HOW TO CALCULATE KAPPA

Calculate marginal distribution:

	Coder A			
Coder B		positive	negative	
	positive	45	5	50%
	negative	5	45	50%
		50%	50%	

HOW TO CALCULATE KAPPA

Calculate raw agreement (a = 0.9)

Calculate:

P(both A and B gives “positive” label) = 0.25

P(both A and B gives “negative” label) = 0.25

Chance_agreement: c = 0.25 + 0.25 = 0.5

Kappa = (a – c)/(1 – c) = (0.9 – 0.5)/(1 – 0.5) = 0.4/0.5 = 0.8

TOOLS TO CALCULATE KAPPA

Online tool:

http://vassarstats.net/kappa.html

EXERCISE: CALCULATE KAPPA AGREEMENT

	Coder A		
Coder B		positive	negative
	positive	89	9
	negative	1	1

REPRODUCIBLE RESEARCH

SYRACUSE UNIVERSITY
School of Information Studies

REPRODUCIBLE RESEARCH

Reproducible research is a cornerstone of scientific research.

Report your data mining approach and results in a reproducible way.

Use tools like RMD to document the process.

If possible, open data access.