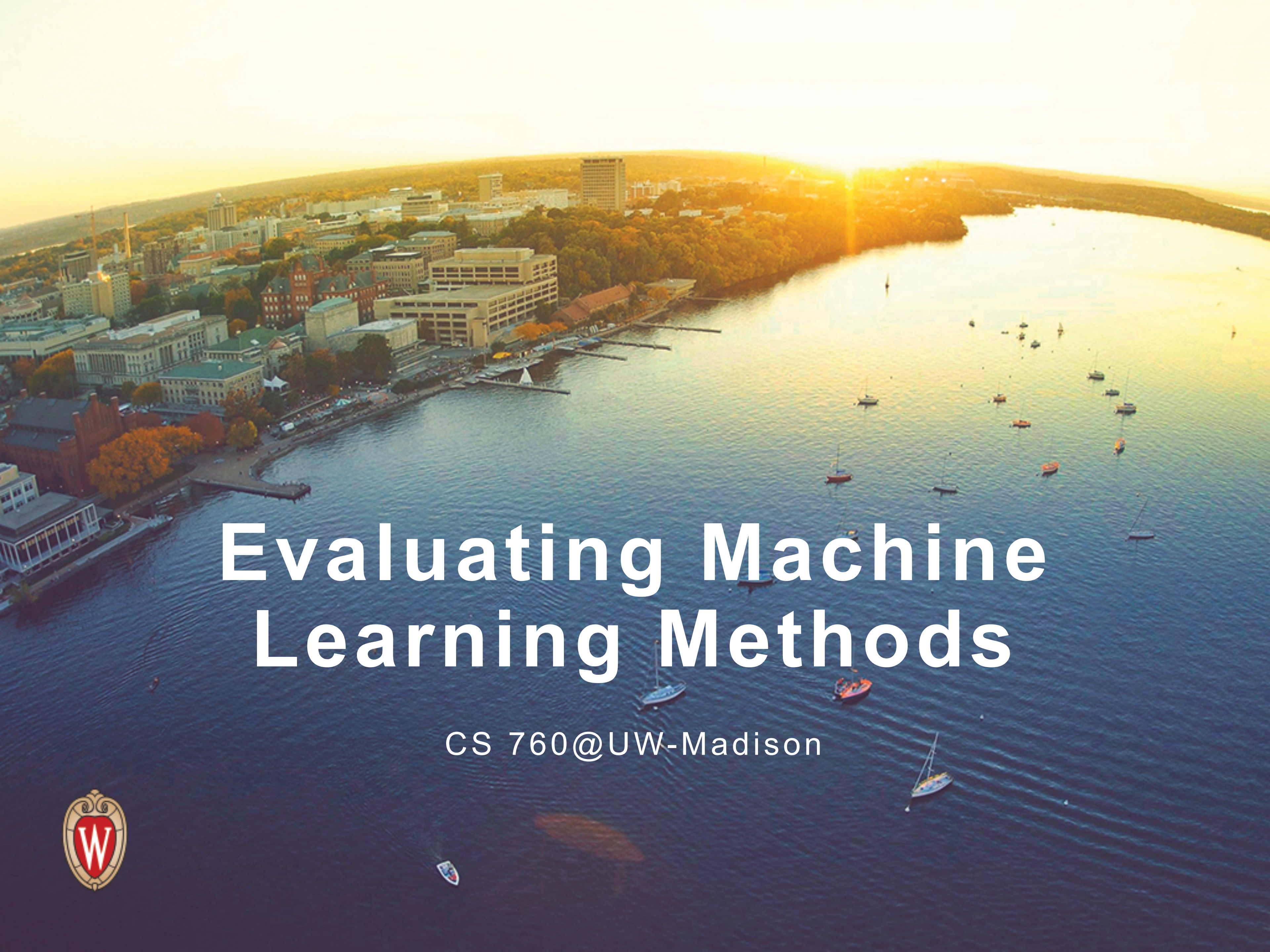


Evaluating Machine Learning Methods

CS 760@UW-Madison

Goals for the lecture

you should understand the following concepts

- bias of an estimator
- learning curves
- stratified sampling
- cross validation
- confusion matrices
- TP, FP, TN, FN
- ROC curves
- PR curves
- confidence intervals for error
- pairwise t-tests for comparing learning systems
- scatter plots for comparing learning systems
- lesion studies

Bias of an estimator

θ true value of parameter of interest (e.g. model accuracy)
$\hat{\theta}$ estimator of parameter of interest (e.g. test set accuracy)

$$\text{Bias}[\hat{\theta}] = \mathrm{E}[\hat{\theta}] - \theta$$

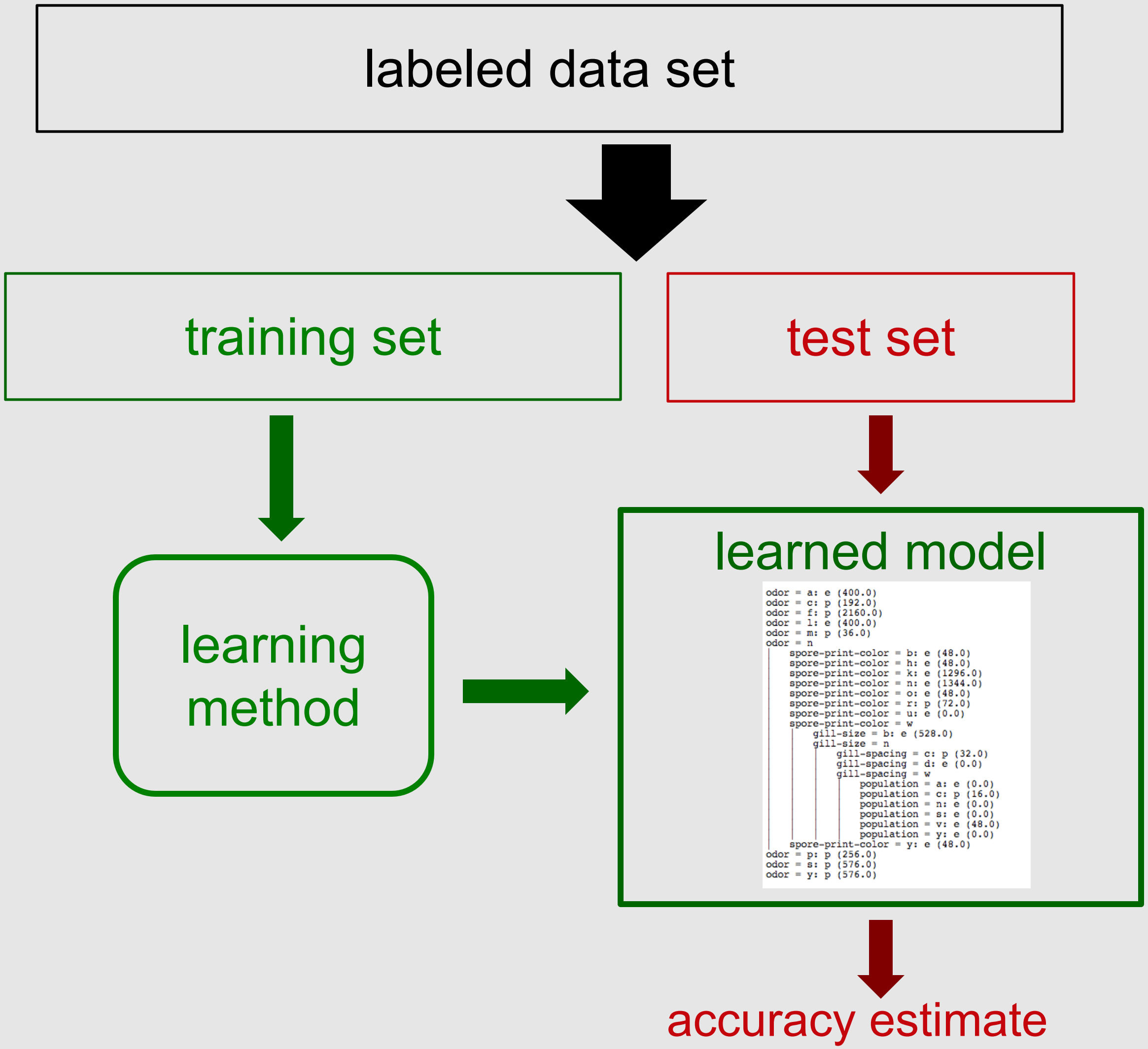
e.g. polling methodologies often have an inherent bias

FiveThirtyEight

POLLSTER	LIVE CALLER WITH CELLPHONES	INTERNET	NCPP/ AAPOR/ ROPER	POLLS ANALYZED	SIMPLE AVERAGE ERROR	RACES CALLED CORRECTLY	ADVANCED +/-	PREDICTIVE +/-	538 GRADE	BANNED BY 538	MEAN-REVERTED BIAS
SurveyUSA			●	763	4.6	90%	-1.0	**-0.8**	A		D+0.1
YouGov		●		707	6.7	93%	-0.3	**+0.1**	B		D+1.6
Rasmussen Reports/ Pulse Opinion Research				657	5.3	79%	+0.4	**+0.7**	C+		R+2.0
Zogby Interactive/JZ Analytics		●		465	5.6	78%	+0.8	**+1.2**	C-		R+0.8
Mason-Dixon Polling & Research, Inc.	●			415	5.2	86%	-0.4	**-0.2**	B+		R+1.0
Public Policy Polling				383	4.9	82%	-0.5	**-0.1**	B+		R+0.2
Research 2000				279	5.5	88%	+0.2	**+0.6**	F	✖	D+1.4

Test sets revisited

How can we get an unbiased estimate of the accuracy of a learned model?



Test sets revisited

How can we get an unbiased estimate of the accuracy of a learned model?

- when learning a model, you should pretend that you don’t have the test data yet (it is “in the mail”)

- if the test-set labels influence the learned model in any way, accuracy estimates will be biased

Learning curves

How does the accuracy of a learning method change as a function of the training-set size?

this can be assessed by plotting *learning curves*

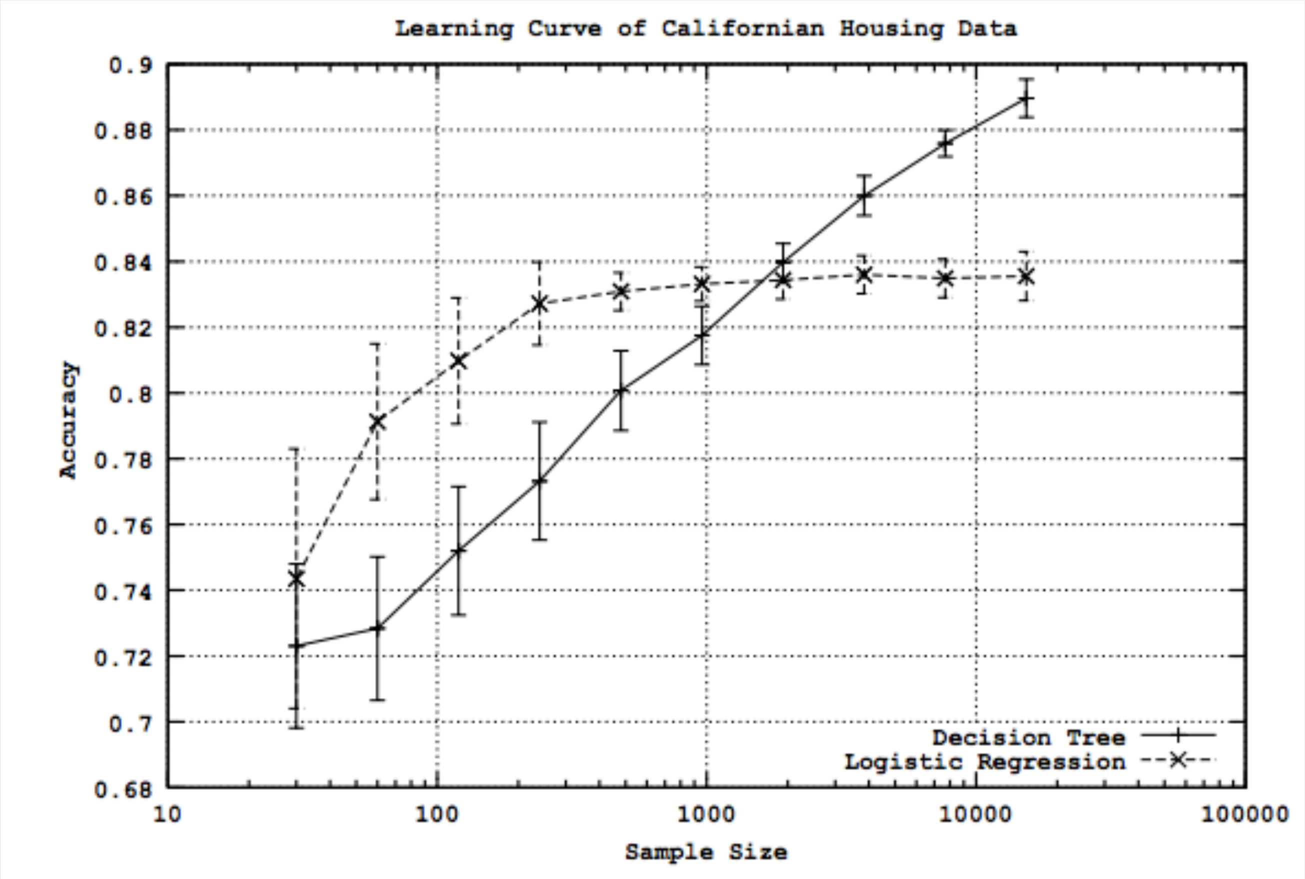
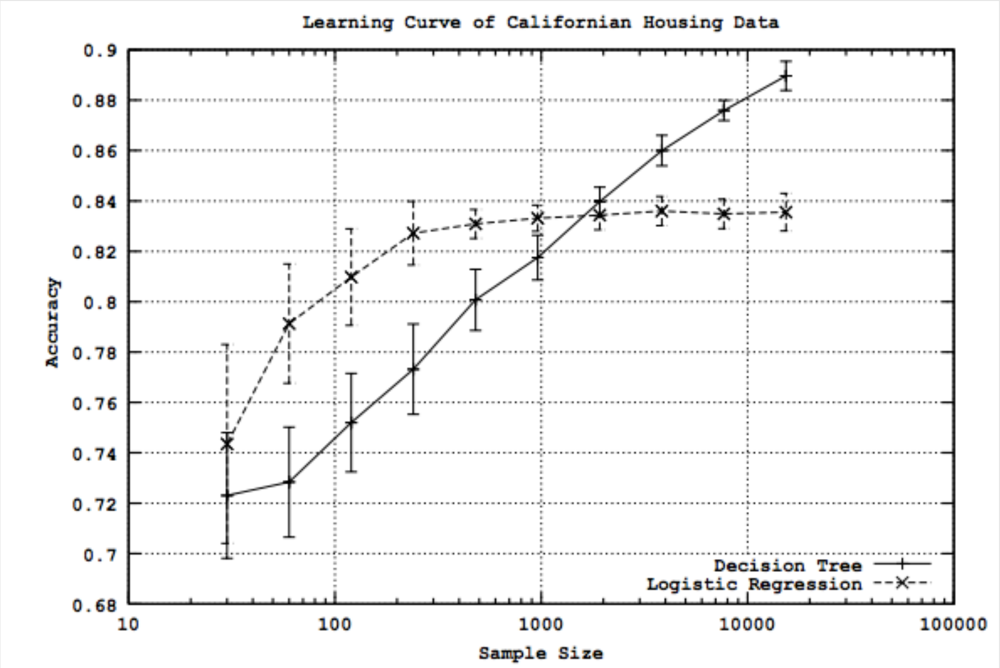


Figure from Perlich et al. *Journal of Machine Learning Research*, 2003

Learning curves

given training/test set partition

- for each sample size s on learning curve
  - (optionally) repeat n times
    - randomly select s instances from training set
    - learn model
    - evaluate model on test set to determine accuracy a
    - plot (s, a) or (s, avg. accuracy and error bars)



Limitations of a single training/test partition

- we may not have enough data to make sufficiently large training and test sets
  - a <u>larger test set</u> gives us more reliable estimate of accuracy (i.e. a lower variance estimate)
  - but… a <u>larger training set</u> will be more representative of how much data we actually have for learning process

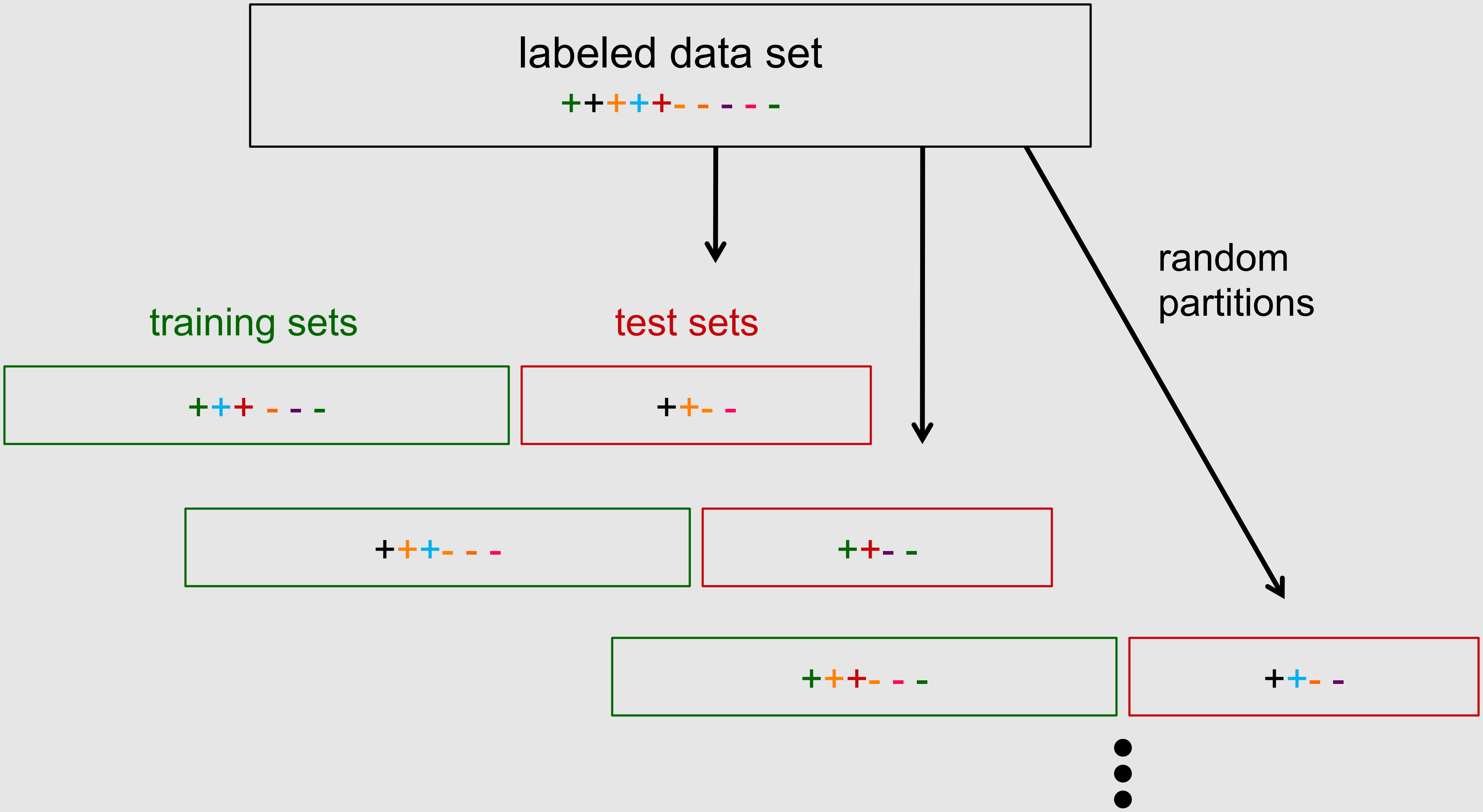
- a single training set doesn’t tell us how sensitive accuracy is to a particular training sample

Using multiple training/test partitions

- two general approaches for doing this
  - random resampling
  - cross validation

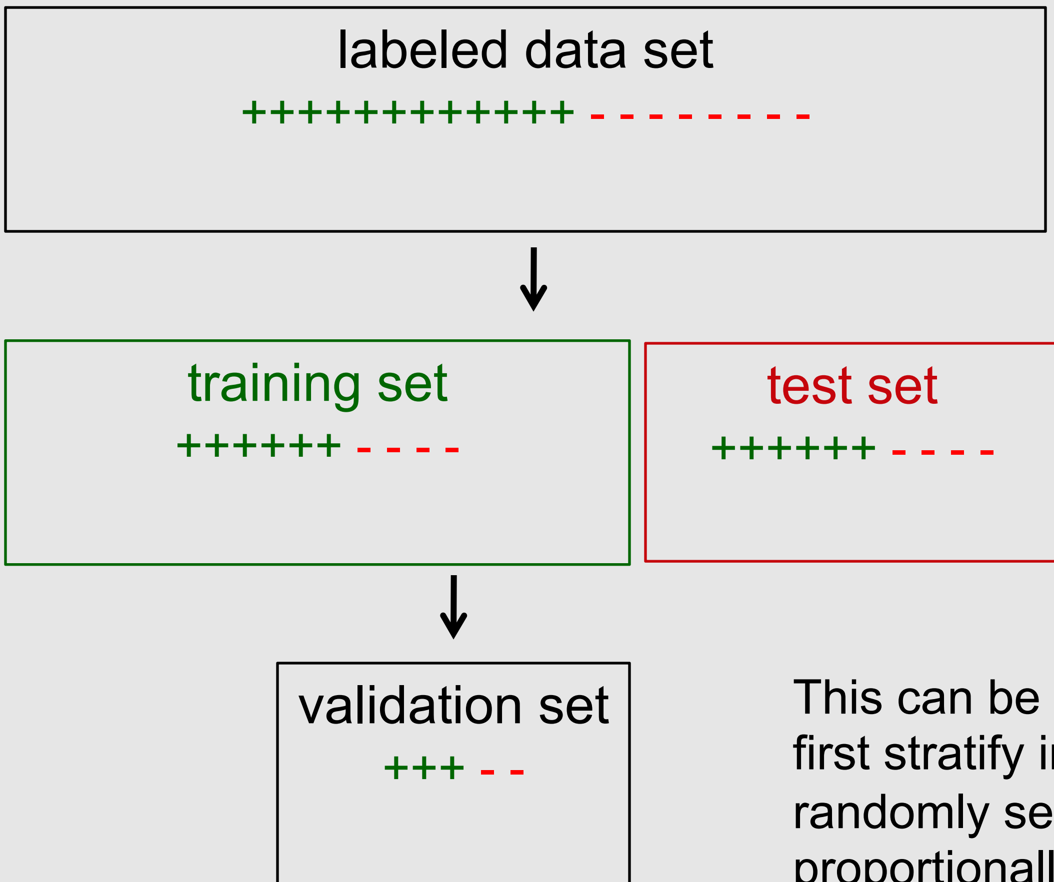
Random resampling

We can address the second issue by repeatedly randomly partitioning the available data into training and test sets.



Stratified sampling

When randomly selecting training or validation sets, we may want to ensure that class proportions are maintained in each selected set



This can be done via stratified sampling: first stratify instances by class, then randomly select instances from each class proportionally.

Cross validation

partition data
into *n* subsamples

labeled data set

s_1	s_2	s_3	s_4	s_5

iteratively leave one subsample out for the test set, train on the rest

iteration	**train on**	**test on**
1	s_2 s_3 s_4 s_5	s_1
2	s_1 s_3 s_4 s_5	s_2
3	s_1 s_2 s_4 s_5	s_3
4	s_1 s_2 s_3 s_5	s_4
5	s_1 s_2 s_3 s_4	s_5

Cross validation example

Suppose we have 100 instances, and we want to estimate accuracy with cross validation

iteration	train on	test on	correct
1	s_2 s_3 s_4 s_5	s_1	11 / 20
2	s_1 s_3 s_4 s_5	s_2	17 / 20
3	s_1 s_2 s_4 s_5	s_3	16 / 20
4	s_1 s_2 s_3 s_5	s_4	13 / 20
5	s_1 s_2 s_3 s_4	s_5	16 / 20

accuracy = 73/100 = 73%

Cross validation

- 10-fold cross validation is common, but smaller values of *n* are often used when learning takes a lot of time
- in *leave-one-out* cross validation, *n* = # instances
- in *stratified* cross validation, stratified sampling is used when partitioning the data
- CV makes efficient use of the available data for testing
- note that whenever we use multiple training sets, as in CV and random resampling, we are evaluating a <u>learning method</u> as opposed to an <u>individual learned hypothesis</u>

Confusion matrices

How can we understand what types of mistakes a learned model makes?

task: activity recognition from video

actual class

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	89	0	0	0	11	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	89	0	11	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	0	0	100	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	67	33
wave2	0	0	0	0	0	0	0	0	0	100

predicted class

figure from vision.jhu.edu

Confusion matrix for 2-class problems

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

$$\text{error} = 1 - \text{accuracy} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Is accuracy an adequate measure of predictive performance?

accuracy may not be useful measure in cases where

- there is a large class skew
  - Is 98% accuracy good when 97% of the instances are negative?
- there are differential misclassification costs – say, getting a positive wrong costs more than getting a negative wrong
  - Consider a medical domain in which a false positive results in an extraneous test but a false negative results in a failure to treat a disease
- we are most interested in a subset of high-confidence predictions

Other accuracy metrics

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

Other accuracy metrics

		actual class: positive	actual class: negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{true positive rate (recall)} = \frac{\text{TP}}{\text{actual pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Other accuracy metrics

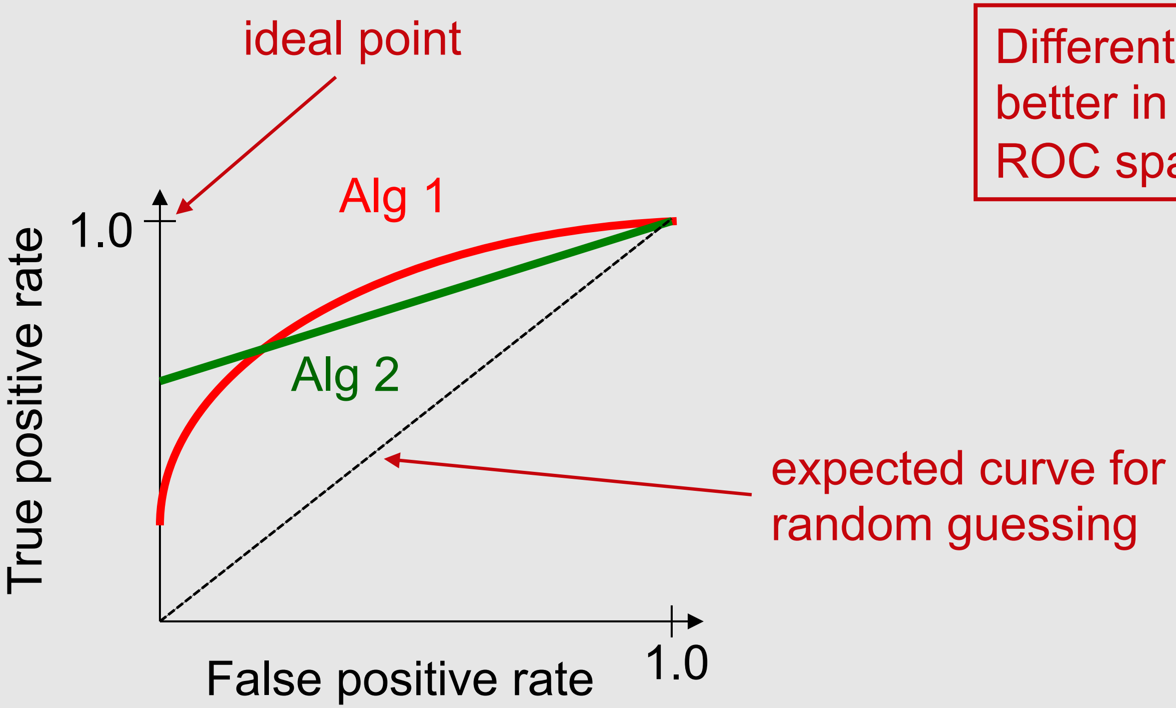
		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{true positive rate (recall)} = \frac{TP}{\text{actual pos}} = \frac{TP}{TP + FN}$$

$$\text{false positive rate} = \frac{FP}{\text{actual neg}} = \frac{FP}{TN + FP}$$

ROC curves

A *Receiver Operating Characteristic* (*ROC*) curve plots the TP-rate vs. the FP-rate as a threshold on the confidence of an instance being positive is varied



Different methods can work better in different parts of ROC space.

Algorithm for creating an ROC curve

let $\left(\left(y^{(1)}, c^{(1)}\right) \ldots \left(y^{(m)}, c^{(m)}\right)\right)$ be the test-set instances sorted according to predicted confidence $c^{(i)}$ that each instance is positive

let num_neg, num_pos be the number of negative/positive instances in the test set

$TP = 0, \; FP = 0$

$last\_TP = 0$

for $i = 1$ to m

  // find thresholds where there is a pos instance on high side, neg instance on low side

  if (i > 1) and ($c^{(i)} \neq c^{(i-1)}$) and ($y^{(i)} ==$ neg) and ($TP > last\_TP$)

    $FPR = FP \,/\, num\_neg, \quad TPR = TP \,/\, num\_pos$

    output (FPR, TPR) coordinate

    $last\_TP = TP$

  if $y^{(i)} ==$ pos

    $++TP$

  else

    $++FP$

$FPR = FP \,/\, num\_neg, \; TPR = TP \,/\, num\_pos$

output (FPR, TPR) coordinate

Plotting an ROC curve

instance	confidence positive		correct class
Ex 9	.99		+
Ex 7	.98	TPR= 2/5, FPR= 0/5	+
Ex 1	.72		-
Ex 2	.70		+
Ex 6	.65	TPR= 4/5, FPR= 1/5	+
Ex 10	.51		-
Ex 3	.39		-
Ex 5	.24	TPR= 5/5, FPR= 3/5	+
Ex 4	.11		-
Ex 8	.01	TPR= 5/5, FPR= 5/5	-

ROC curve example

task: recognizing genomic units called operons

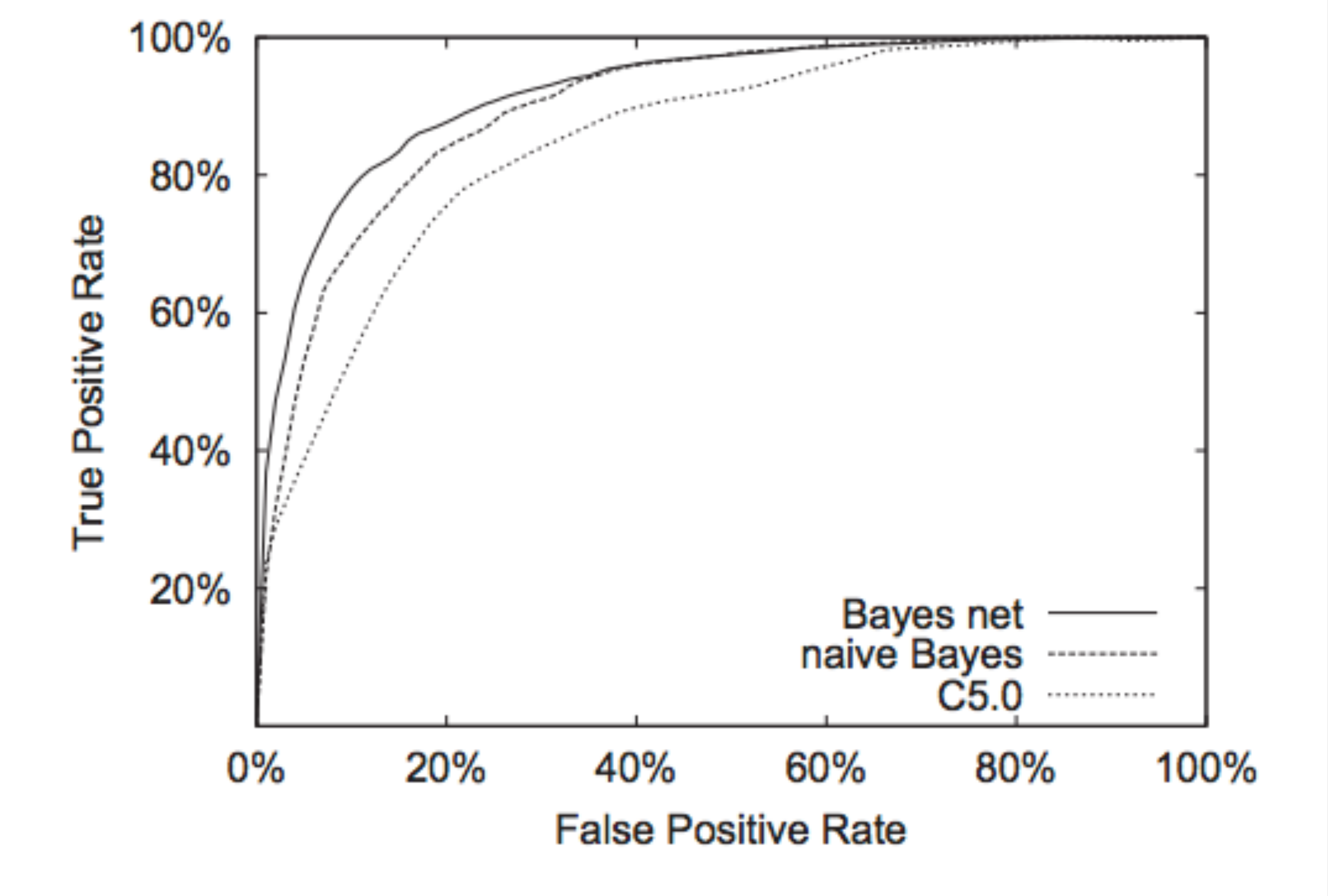
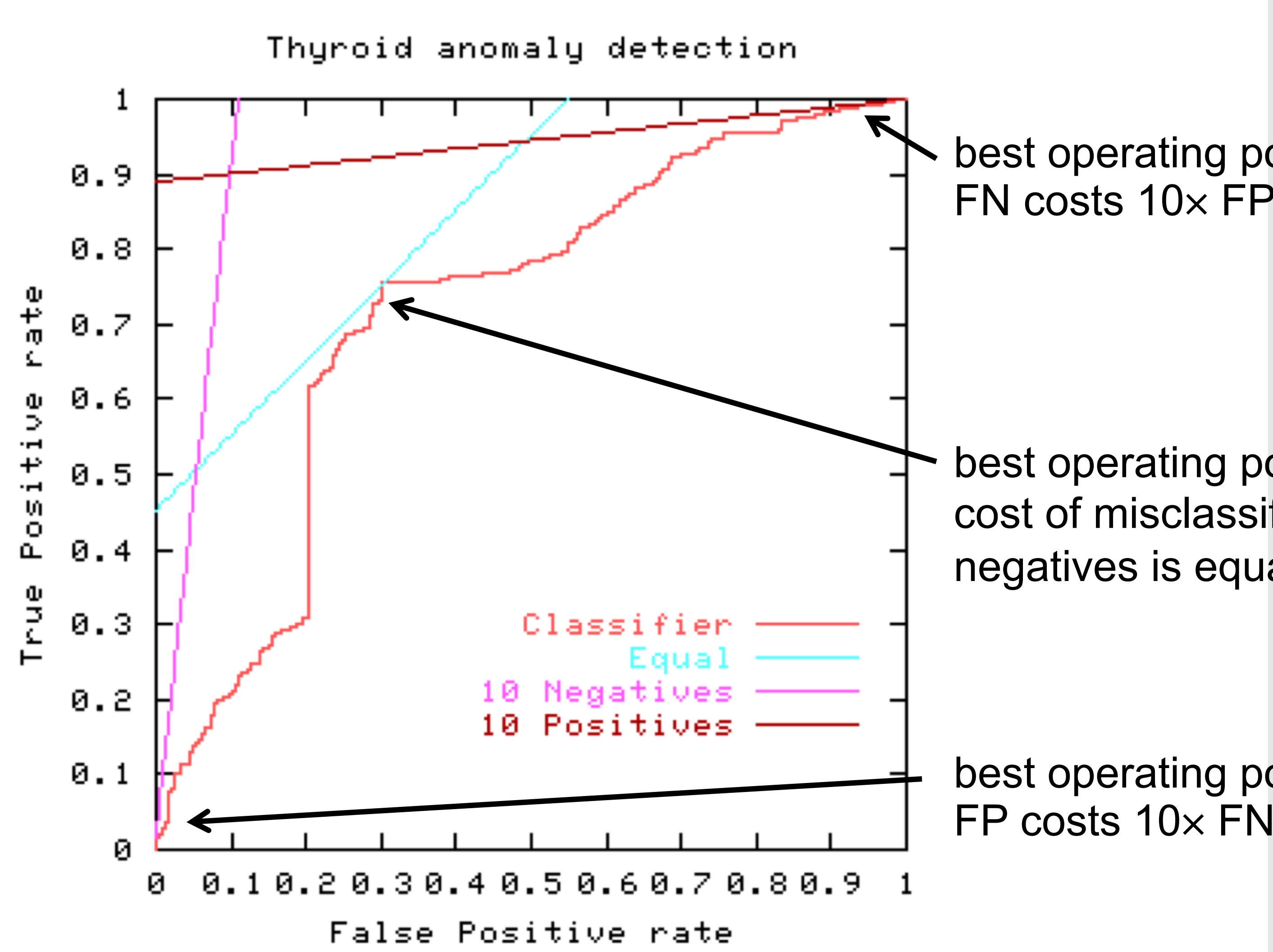


figure from Bockhorst et al., *Bioinformatics* 2003

ROC curves and misclassification costs

The best operating point depends on the relative costs of FN and FP misclassifications



best operating point when FN costs 10× FP

best operating point when cost of misclassifying positives and negatives is equal

best operating point when FP costs 10× FN

ROC curves

Does a low false-positive rate indicate that most positive predictions (i.e. predictions with confidence > some threshold) are correct?

suppose our TPR is 0.9, and FPR is 0.01

fraction of instances that are positive	fraction of positive predictions that are correct
0.5	0.989
0.1	0.909
0.01	0.476
0.001	0.083

Other accuracy metrics

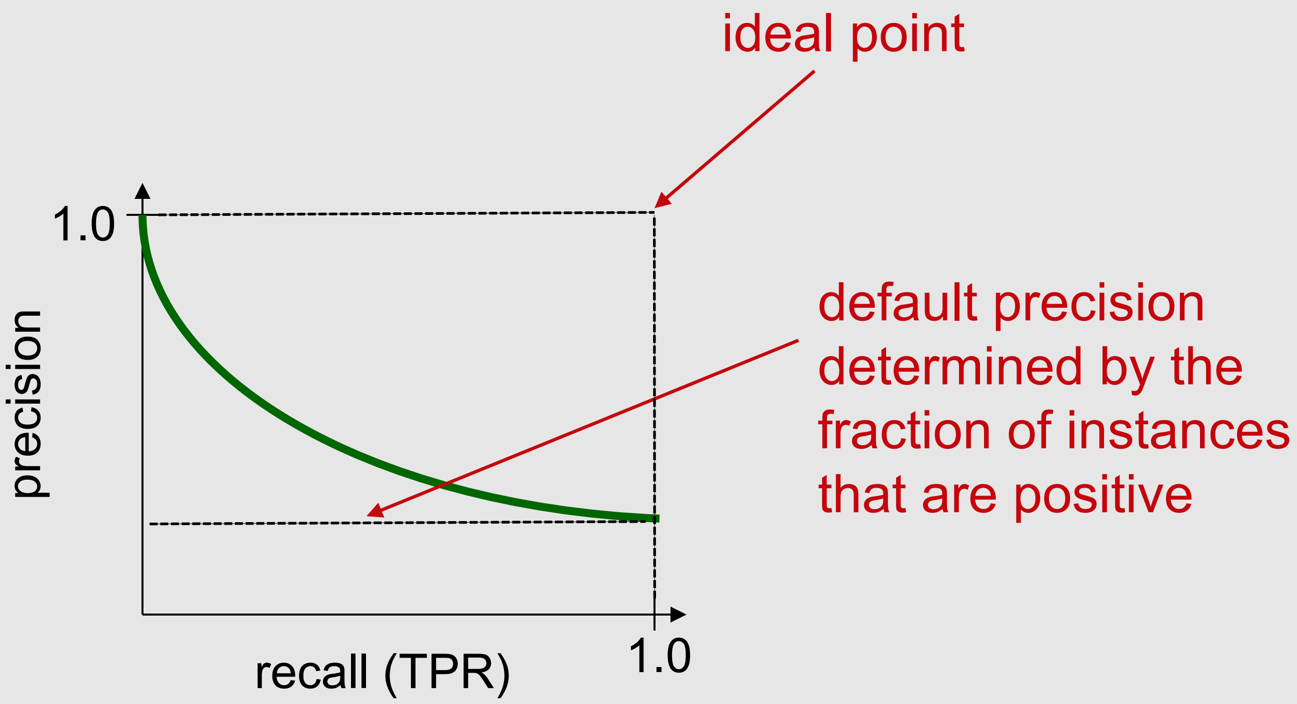
predicted class		actual class: positive	actual class: negative
	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{recall (TP rate)} = \frac{\text{TP}}{\text{actual pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{precision (positive predictive value)} = \frac{\text{TP}}{\text{predicted pos}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision/recall curves

A *precision/recall curve* plots the precision vs. recall (TP-rate) as a threshold on the confidence of an instance being positive is varied



Precision/recall curve example

predicting patient risk for VTE

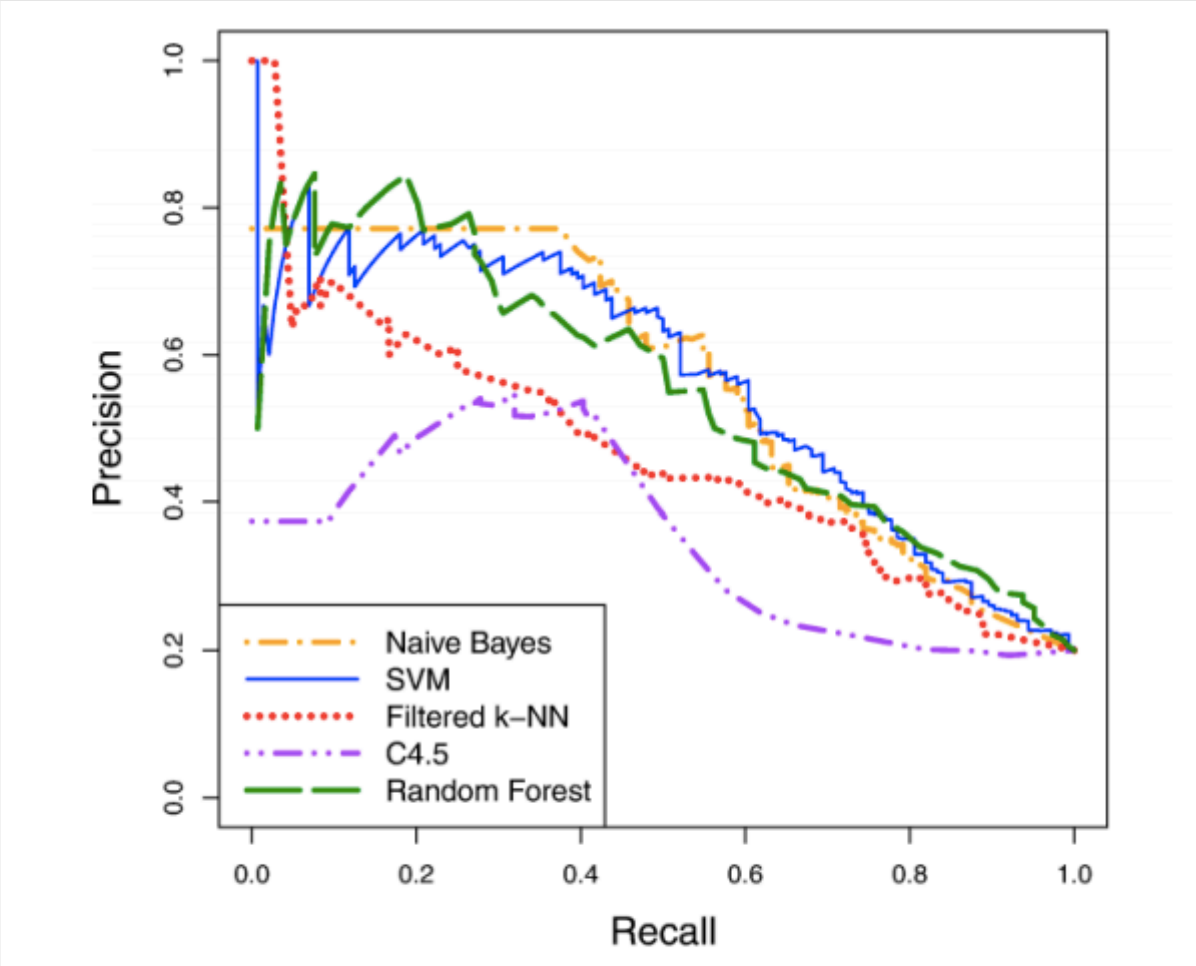


figure from Kawaler et al., *Proc. of AMIA Annual Symosium,* 2012

How do we get one ROC/PR curve when we do cross validation?

Approach 1

- make assumption that confidence values are comparable across folds
- pool predictions from all test sets
- plot the curve from the pooled predictions

Approach 2 (for ROC curves)

- plot individual curves for all test sets
- view each curve as a function
- plot the average curve for this set of functions

Comments on ROC and PR curves

both

- allow predictive performance to be assessed at various levels of confidence
- assume binary classification tasks
- sometimes summarized by calculating *area under the curve*

ROC curves

- insensitive to changes in class distribution (ROC curve does not change if the proportion of positive and negative instances in the test set are varied)
- can identify optimal classification thresholds for tasks with differential misclassification costs

precision/recall curves

- show <u>the fraction of predictions</u> that are false positives
- well suited for tasks with lots of negative instances

Confidence intervals on error

Given the observed error (accuracy) of a model over a limited sample of data, how well does this error characterize its accuracy over additional instances?

Suppose we have

- a learned model h
- a test set S containing n instances drawn independently of one another and independent of h
- $n \geq 30$
- h makes r errors over the n instances

our best estimate of the error of h is

$$error_S(h) = \frac{r}{n}$$

Confidence intervals on error

With approximately C% probability, the true error lies in the interval

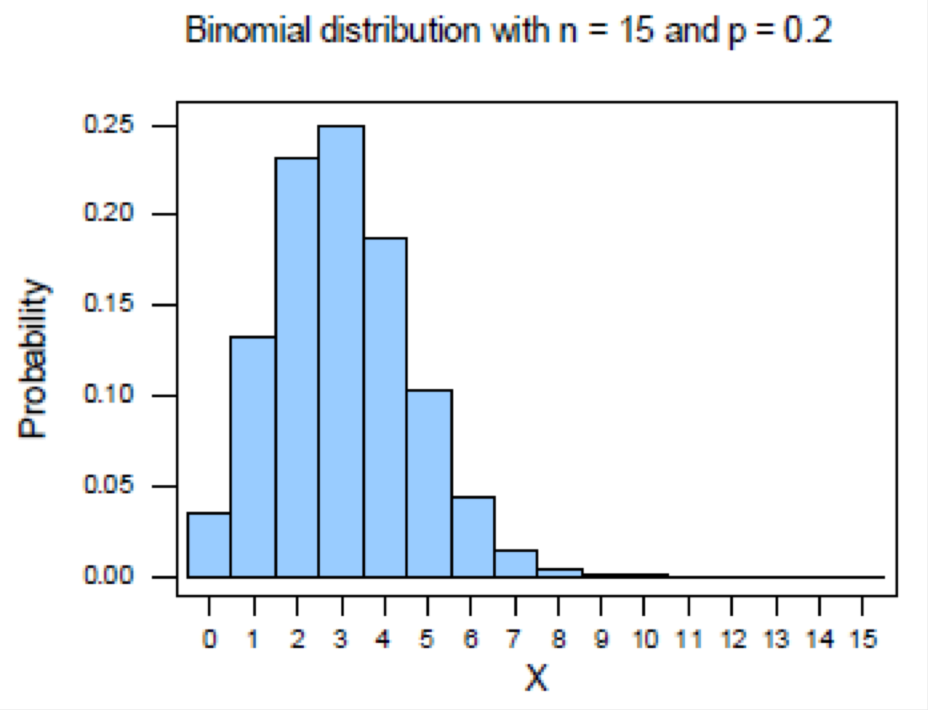
$$error_S(h) \pm z_C \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

where z_C is a constant that depends on C (e.g. for 95% confidence, z_C =1.96)

Confidence intervals on error

How did we get this?

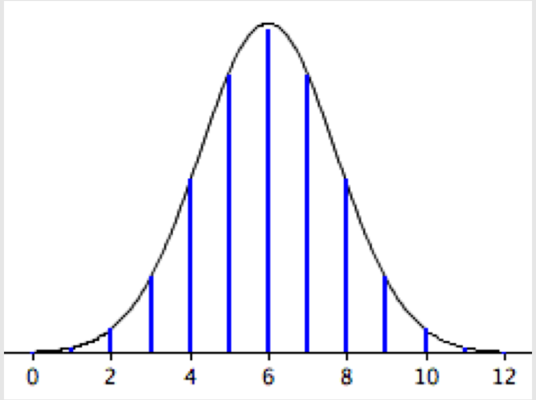
1. Our estimate of the error follows a binomial distribution given by n and p (the true error rate over the data distribution)



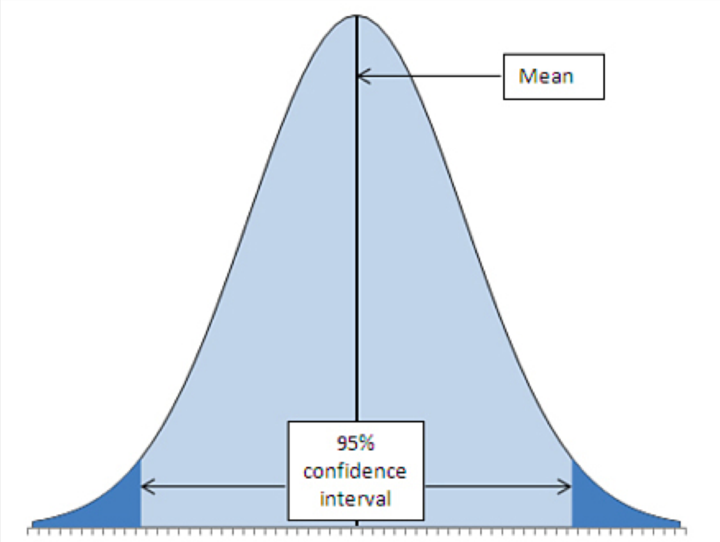
2. Most common way to determine a binomial confidence interval is to use the *normal approximation* (although can calculate exact intervals if n is not too large)

Confidence intervals on error

2. When $n \geq 30$, and p is not too extreme, the normal distribution is a good approximation to the binomial



3. We can determine the C% confidence interval by determining what bounds contain C% of the probability mass under the normal



Comparing learning systems

How can we determine if one learning system provides better performance than another

- for a particular task?
- across a set of tasks / data sets?

Motivating example

<u>Accuracies on test sets</u>

System A:	80%	50	75	…	99
System B:	79	49	74	…	98
δ :	+1	+1	+1	…	+1

- Mean accuracy for System A is better, but the standard deviations for the two clearly overlap
- Notice that System A is always better than System B

Comparing systems using a paired t test

- consider δ's as observed values of a set of i.i.d. random variables

- *null hypothesis*: the 2 learning systems have the same accuracy
- *alternative hypothesis*: one of the systems is more accurate than the other

- hypothesis test:
  - use paired t-test to determine probability p that mean of δ's would arise from null hypothesis
  - if p is sufficiently small (typically < 0.05) then reject the null hypothesis

Comparing systems using a paired *t* test

1. calculate the sample mean

$$\bar{\delta} = \frac{1}{n} \sum_{i=1}^{n} \delta_i$$

2. calculate the *t* statistic

$$t = \frac{\bar{\delta}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (\delta_i - \bar{\delta})^2}}$$

3. determine the corresponding *p*-value, by looking up *t* in a table of values for the Student's *t*-distribution with *n-1* degrees of freedom

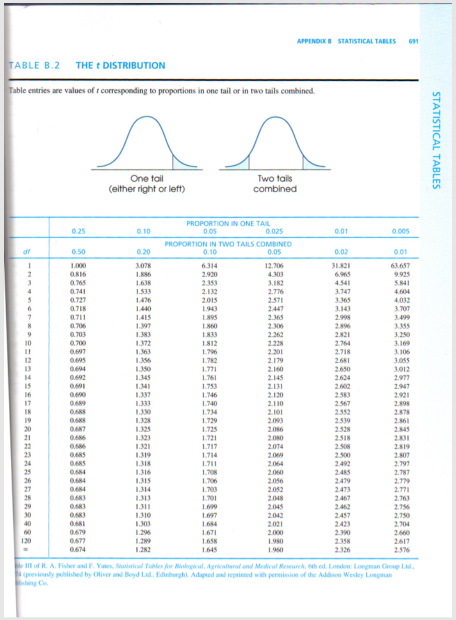


TABLE B.2 THE *t* DISTRIBUTION

Table entries are values of *t* corresponding to proportions in one tail or in two tails combined.

One tail
(either right or left)

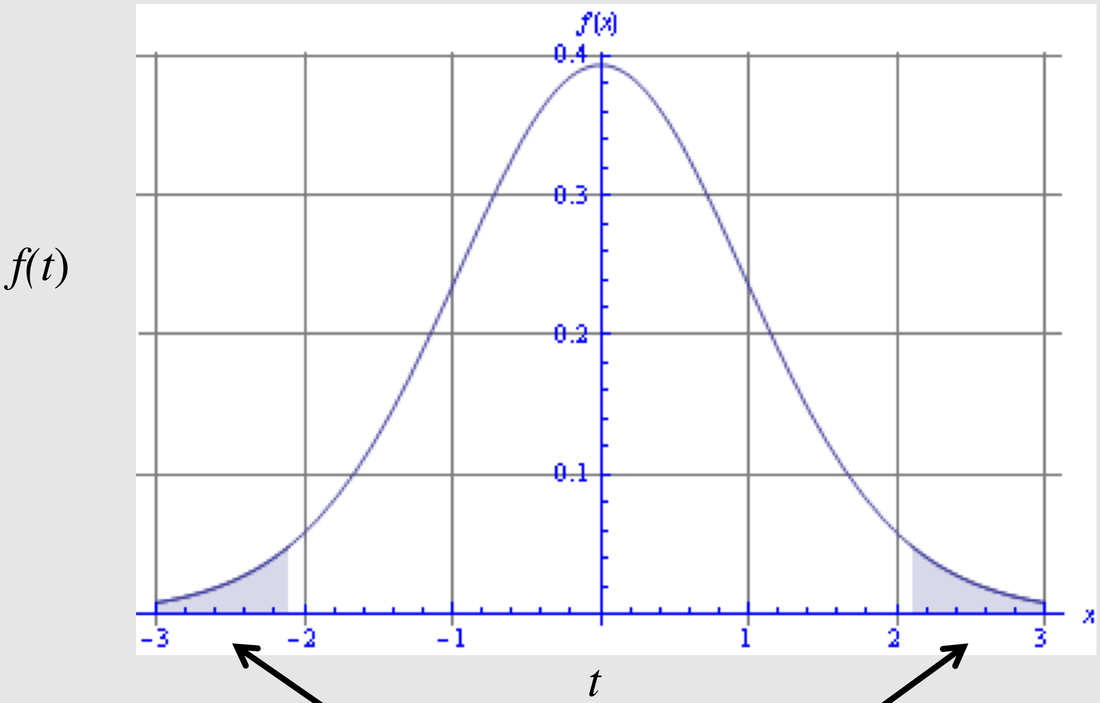
Two tails
combined

	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

ble III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman Group Ltd., 74 (previously published by Oliver and Boyd Ltd., Edinburgh). Adapted and reprinted with permission of the Addison Wesley Longman blishing Co.

STATISTICAL TABLES

Comparing systems using a paired t test



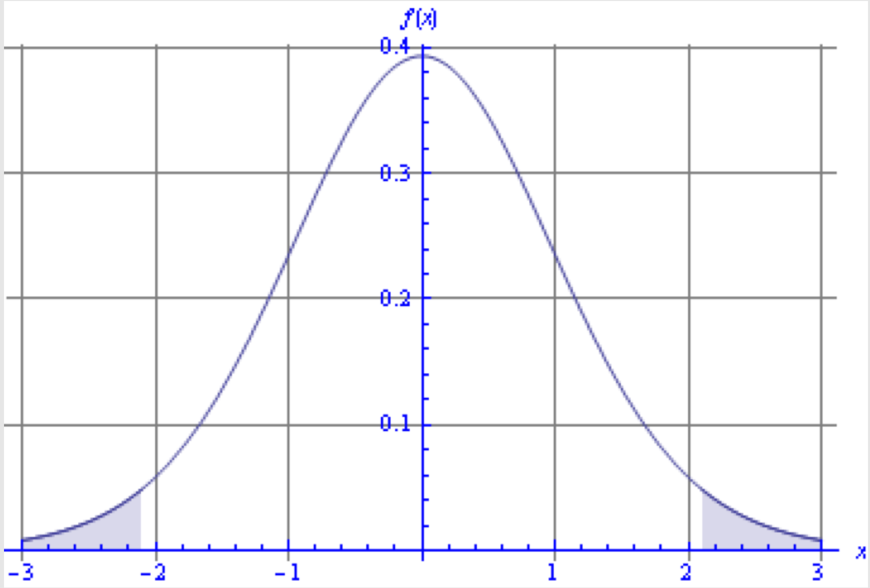
The null distribution of our t statistic looks like this

The p-value indicates how far out in a tail our t statistic is

If the p-value is sufficiently small, we reject the <u>null hypothesis,</u> since it is unlikely we'd get such a t by chance

for a two-tailed test, the p-value represents the probability mass in these two regions

Why do we use a two-tailed test?



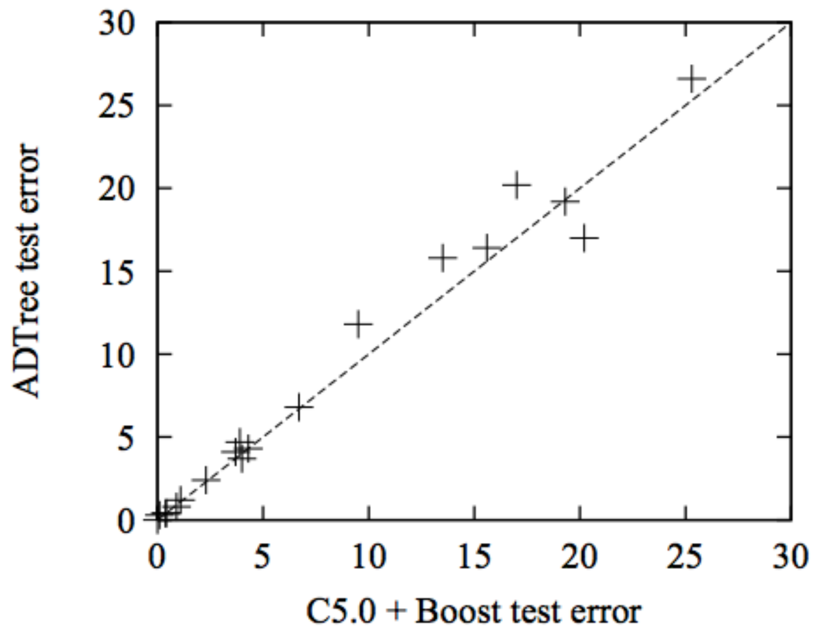
- a two-tailed test asks the question: is the accuracy of the two systems different
- a one-tailed test asks the question: is system A better than system B
- a priori, we don't know which learning system will be more accurate (if there is a difference) – we want to allow that either one might be

Comments on hypothesis testing to compare learning systems

- the paired t-test can be used to compare two learning systems
- other tests (e.g. McNemar's χ^2 test) can be used to compare two learned models
- a statistically significant difference is not necessarily a large-magnitude difference

Scatter plots for pairwise method comparison

We can compare the performance of two methods *A* and *B* by plotting (*A performance*, *B performance*) across <u>numerous data sets</u>



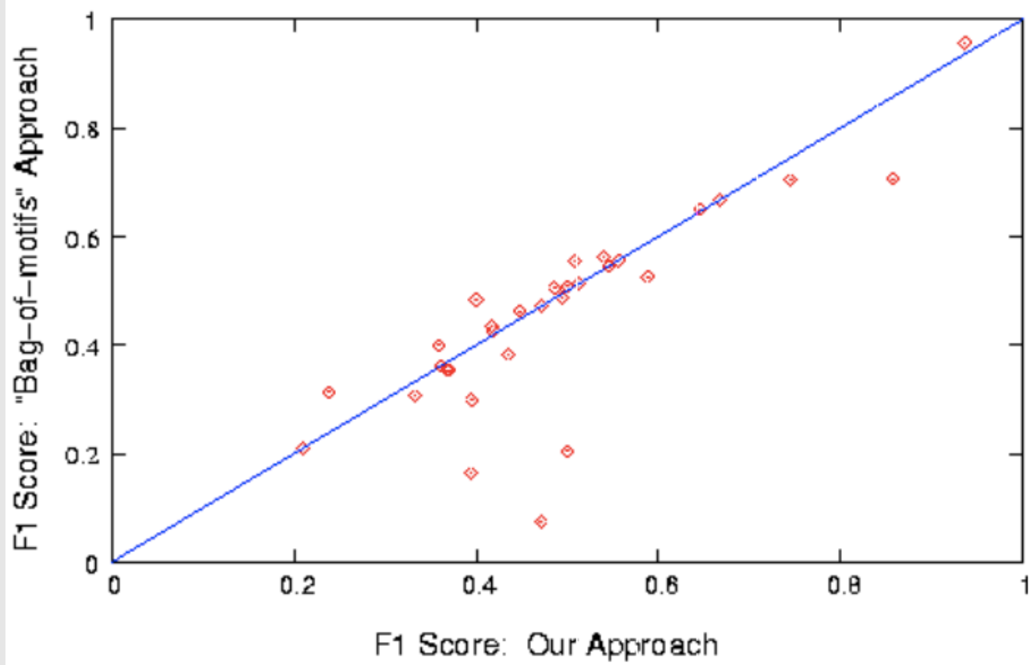


figure from Freund & Mason, *ICML* 1999

figure from Noto & Craven, *BMC Bioinformatics* 2006

Lesion (ablation) studies

We can gain insight into what contributes to a learning system's performance by removing (lesioning) components of it

The ROC curves here show how performance is affected when various feature types are removed from the learning representation

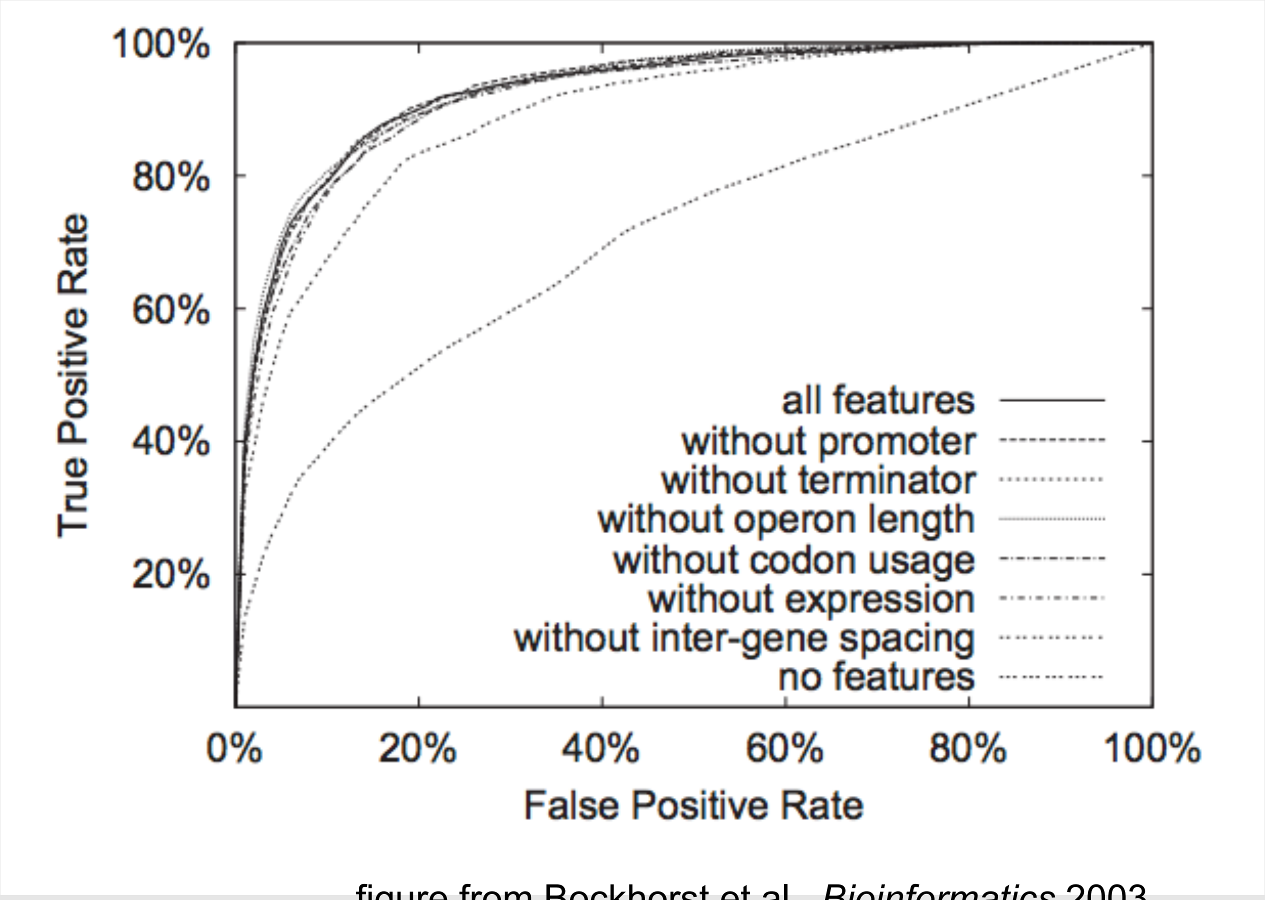


figure from Bockhorst et al., *Bioinformatics* 2003

To avoid pitfalls, ask

1. Is my held-aside test data really representative of going out to collect new data?

   - Even if your methodology is fine, someone may have collected features for positive examples differently than for negatives – should be randomized

   - Example: samples from cancer processed by different people or on different days than samples for normal controls

To avoid pitfalls, ask

2. Did I repeat my entire data processing procedure on every fold of cross-validation, using only the training data for that fold?

   - On each fold of cross-validation, did I ever access in any way the label of a test instance?

   - Any preprocessing done over entire data set (feature selection, parameter tuning, threshold selection) must not use labels

To avoid pitfalls, ask

3. Have I modified my algorithm so many times, or tried so many approaches, on this same data set that I (the human) am overfitting it?

- Have I continually modified my preprocessing or learning algorithm until I got some improvement on this data set?
- If so, I really need to get some additional data now to at least test on

THANK YOU

Some of the slides in these lectures have been adapted/borrowed from materials developed by Yingyu Liang, Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, and Pedro Domingos.