

Evaluating Machine Learning Methods

ECE 5470 Lecture 24

Evaluating Machine Learning Methods

With edits and additions by
APR

C. David Page Jr.

[Department of Biostatistics and Medical Informatics](#)
and [Department of Computer Sciences](#)
School of Medicine and Public Health
University of Wisconsin-Madison

www.cs.wisc.edu/~dpage/cs760/

1

Goals for the lecture

you should understand the following concepts

- test sets
- learning curves
- validation (tuning) sets
- stratified sampling
- cross validation
- internal cross validation
- confusion matrices
- TP, FP, TN, FN
- ROC curves
- confidence intervals for error
- pairwise t -tests for comparing learning systems
- scatter plots for comparing learning systems
- lesion studies

2

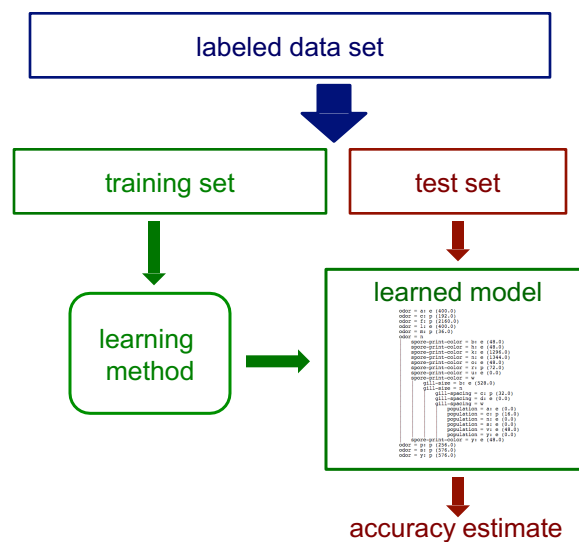
Goals for the lecture (continued)

- recall/sensitivity/true positive rate (TPR)
- precision/positive predictive value (PPV)
- specificity and false positive rate (FPR or 1-specificity)
- precision-recall (PR) curves

3

Test sets revisited

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



4

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

* In some applications it is reasonable to assume that you have access to the feature vector (i.e. x) but not the y part of each test instance.

5

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*

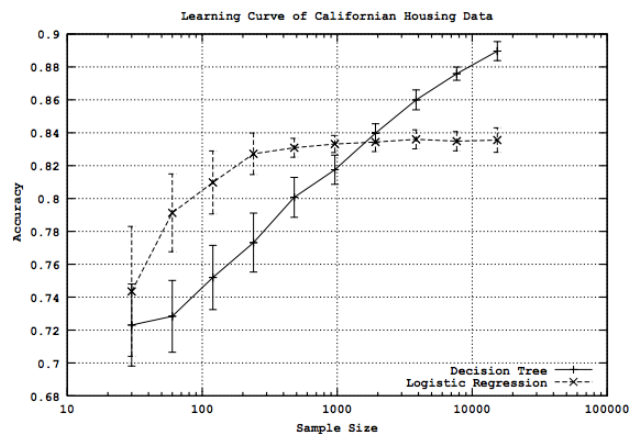


Figure from Pertich et al. *Journal of Machine Learning Research*, 2003

6



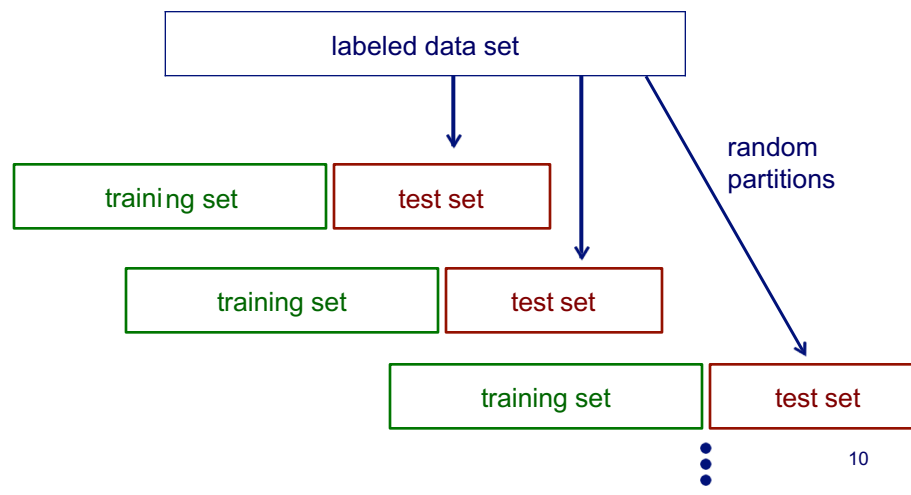
Limitations of using a single training/test partition

- we may not have enough data to make sufficiently large training and test sets
 - a larger test set gives us more reliable estimate of accuracy (i.e. a lower variance estimate)
 - but... a larger training set 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

9

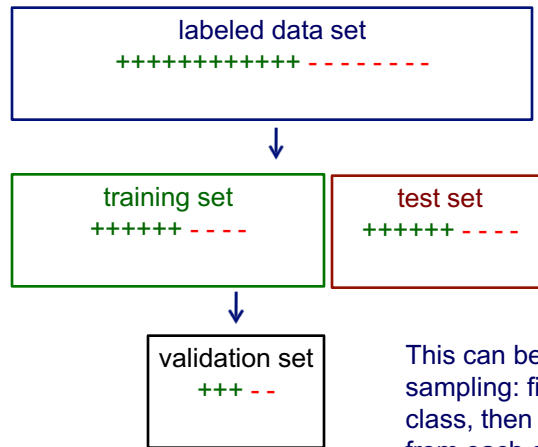
Random resampling

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

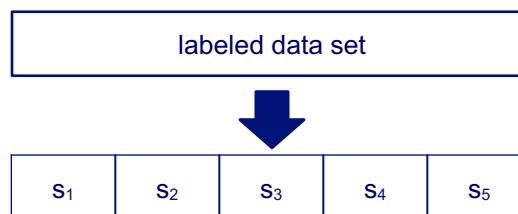


This can be done via stratified sampling: first stratify instances by class, then randomly select instances from each class proportionally.

11

Cross validation

partition data into n subsamples



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

12

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

$$\text{accuracy} = 73/100 = 73\%$$

13

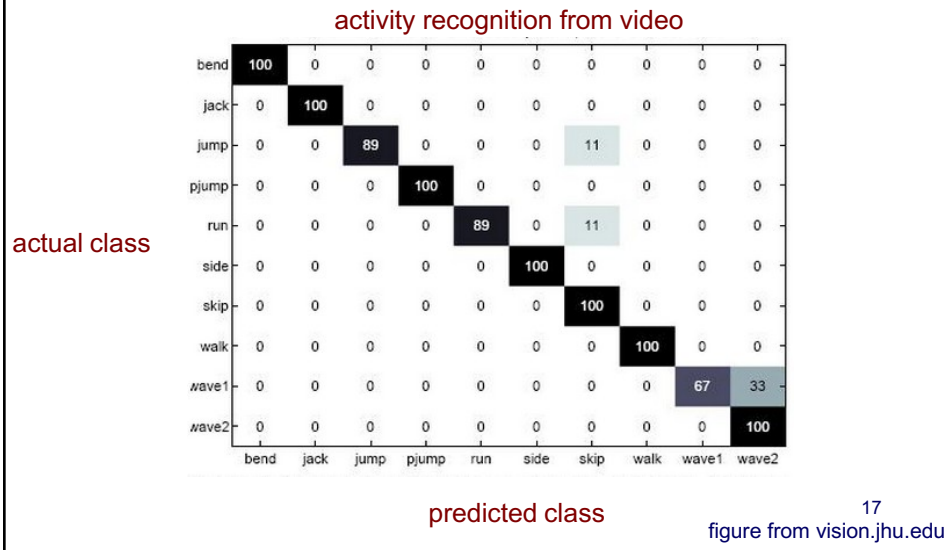
Cross validation

- 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 learning method as opposed to an individual learned model

14

Confusion matrices

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



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)

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

Is accuracy an adequate measure of predictive performance?

Accuracy may not be useful measure in cases where:

- A. there is a large class skew
 - Is 98% accuracy good if 97% of the instances are negative?
- B. 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
- C. we are most interested in a subset of high-confidence predictions

19

Other accuracy metrics

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

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

(Sensitivity)

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

(1 - Specificity)

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

20

True and False Positives and Negatives

		True condition			
		Condition positive	Condition negative	Prevalence $= \frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision $= \frac{\sum \text{True positive}}{\sum \text{Test outcome positive}}$	False discovery rate (FDR) $= \frac{\sum \text{False positive}}{\sum \text{Test outcome positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) $= \frac{\sum \text{False negative}}{\sum \text{Test outcome negative}}$	Negative predictive value (NPV) $= \frac{\sum \text{True negative}}{\sum \text{Test outcome negative}}$
Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$		True positive rate (TPR), Sensitivity, Recall $= \frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR-}}$
		False negative rate (FNR), Miss rate $= \frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) $= \frac{\text{FNR}}{\text{TNR}}$	

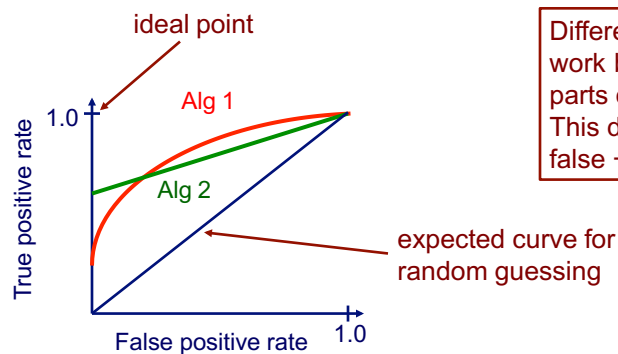
Wikipedia



Cornell University
Vision and Image Analysis

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. This depends on cost of false + vs. false -

ROC curve example

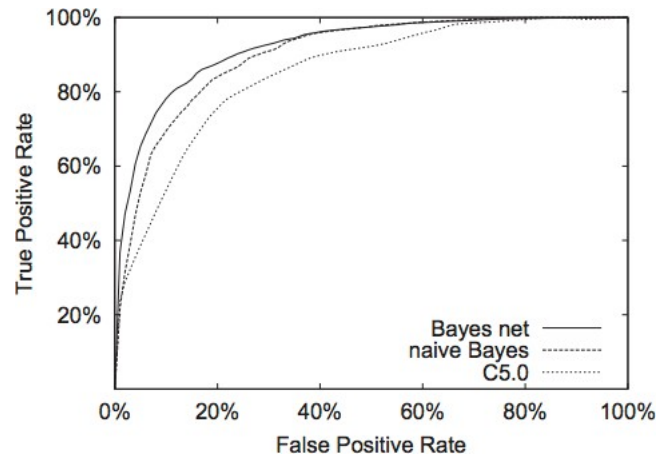
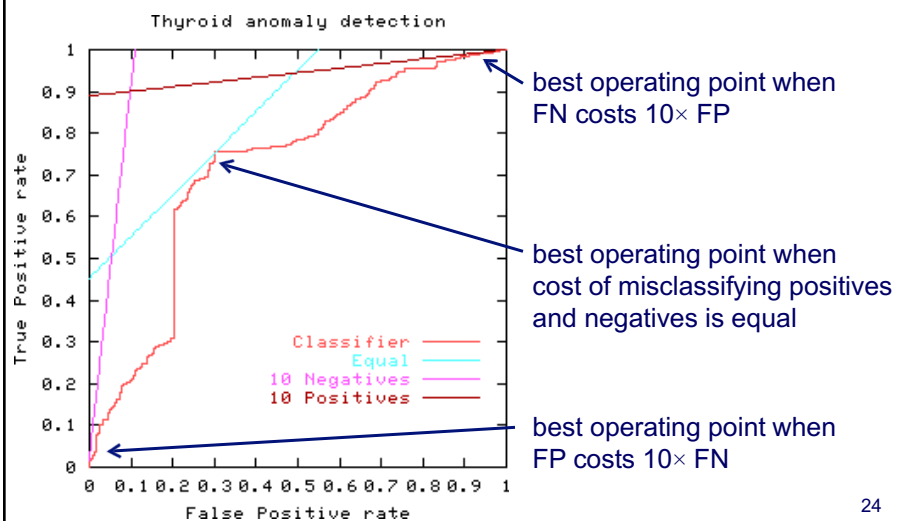


figure from Bockhorst et al., *Bioinformatics* 2003

22

ROC curves and misclassification costs



24

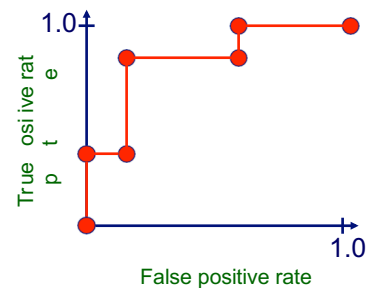
Algorithm for creating an ROC curve

1. sort test-set predictions according to confidence that each instance is positive
2. step through sorted list from high to low confidence
 - i. locate a *threshold* between instances with opposite classes (keeping instances with the same confidence value on the same side of threshold)
 - ii. compute TPR, FPR for instances above threshold
 - iii. output (FPR, TPR) coordinate

25

Plotting an ROC curve

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



26

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

27

Other accuracy metrics

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

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

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

28

Document Retrieval

precision (also called positive predictive value) is the fraction of retrieved instances that are relevant

recall (also known as sensitivity) is the fraction of relevant instances that are retrieved

Wikipedia

		True condition		
		Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$
Predicted condition	Predicted condition positive	True positive	False positive (Type I error)	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{True positive}}{\Sigma \text{Test outcome positive}}$
	Predicted condition negative	False negative (Type II error)	True negative	False omission rate (FOR) $= \frac{\Sigma \text{False negative}}{\Sigma \text{Test outcome negative}}$
Accuracy (ACC) =		True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$

F-measure, harmonic mean of precision and recall

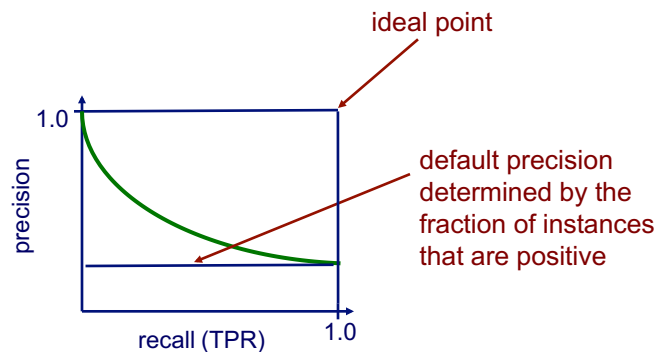
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$



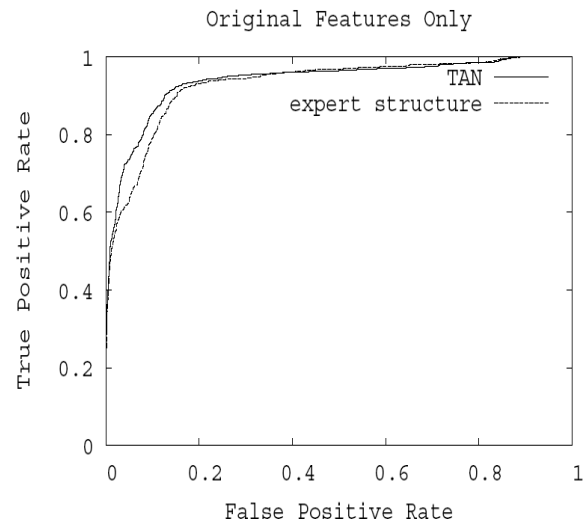
Cornell University
Vision and Image Analysis

Precision/recall curves

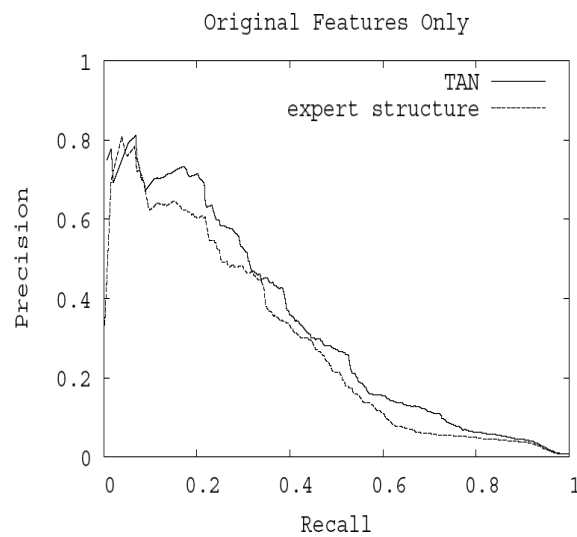
A *precision/recall curve* plots the precision vs. recall (TP-rate) as a threshold on the confidence of an instance being positive is varied

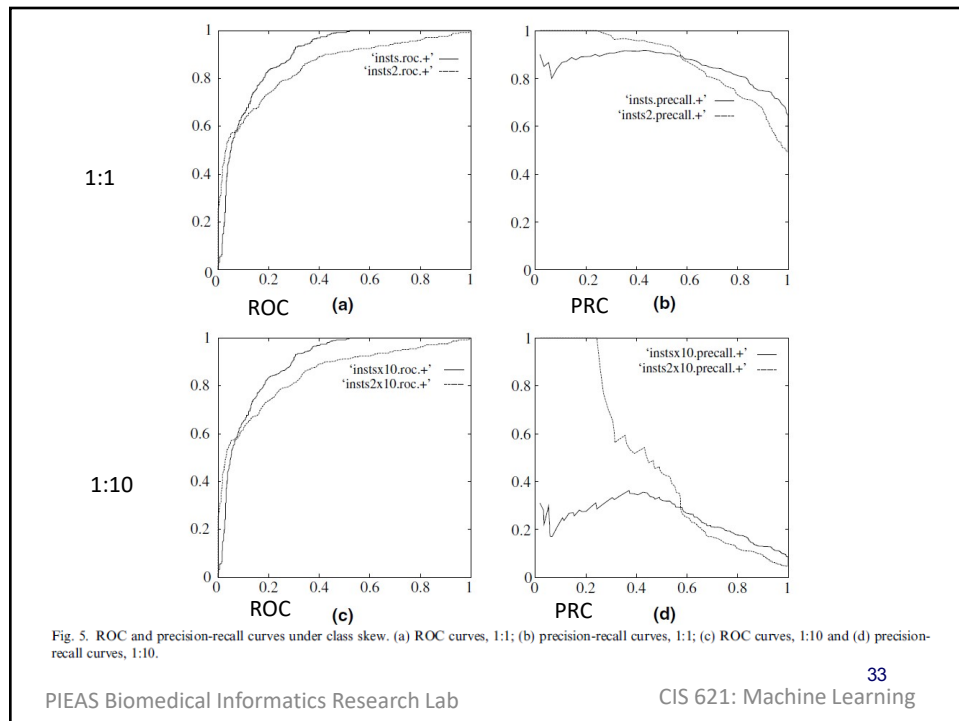


Mammography Example: ROC



Mammography Example: PR

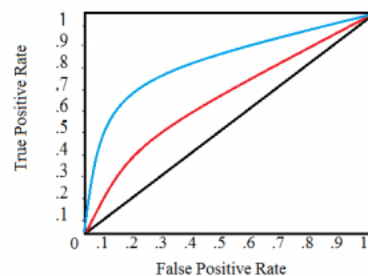




Area Under the Curve (AUC) Logistic regression C-statistic

Often used to compare overall performance between models

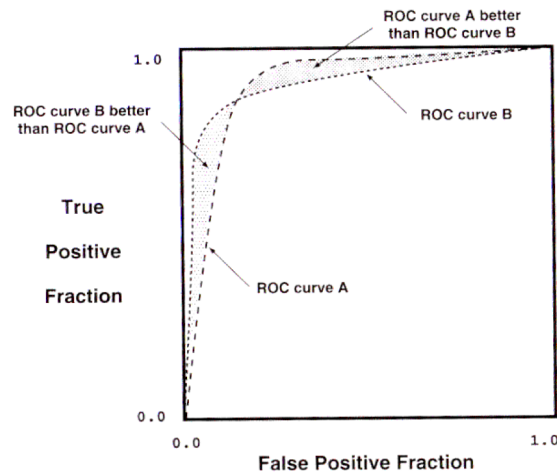
- A value below 0.5 indicates a very poor model.
- A value of 0.5 means that the model is no better than predicting an outcome than random chance.
- Values over 0.7 indicate a good model.
- Values over 0.8 indicate a strong model.
- A value of 1 means that the model perfectly predicts those group members who will experience a certain outcome and those who will not.



32

Area Under the Curve (AUC) Logistic regression C-statistic

AUC may not capture all the important information for a comparison



32

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

Approach 1

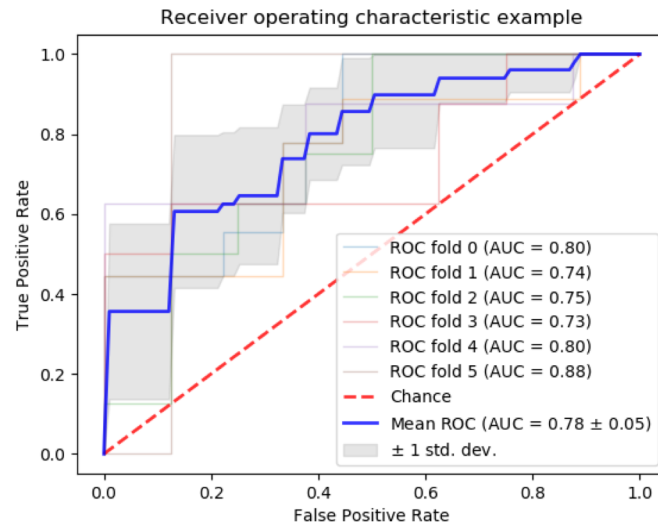
- make assumption that confidence values are comparable across folds
- pool predictions from all test sets
- plot the curve from the pooled predictions

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

32

ROC curve for CV with confidence from scikit-learn



http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html#sphx-glr-auto-examples-model-selection-plot-roc-crossval-py

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 the fraction of predictions that are false positives
- well suited for tasks with lots of negative instances

To Avoid Cross-Validation 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

39

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

40

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

41

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

$$\text{error}_S(h) = \frac{r}{n}$$

42

Confidence intervals on error

With approximately $N\%$ probability, the true error lies in the interval

$$error_s(h) \pm z_N \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}}$$

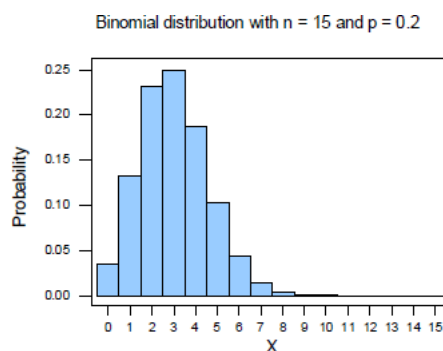
where z_N is a constant that depends on N (e.g. for 95% confidence, $z_N = 1.96$)

43

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)

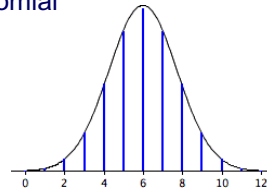


2. Simplest (and most common) way to determine a binomial confidence interval is to use the *normal approximation*

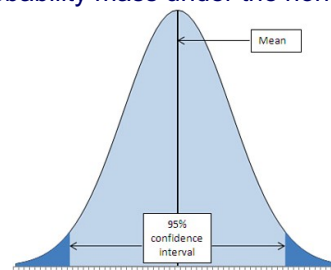
39

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 $N\%$ confidence interval by determining what bounds contain $N\%$ of the probability mass under the normal



45

Empirical Confidence Bounds

- Bootstrapping: Given n examples in data set, randomly, uniformly, independently (with replacement) draw n examples – bootstrap sample
- Repeat 1000 (or 10,000) times:
 - Draw bootstrap sample
 - Repeat entire cross-validation process
- Lower (upper) bound is result such that 2.5% of runs yield lower (higher)

46

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?

47

Motivating example

	<u>Accuracies on test sets</u>				
System 1:	80%	50	75	...	99
System 2:	79	49	74	...	98
δ :	+1	+1	+1	...	+1

- Mean accuracy for System 1 is better, but the standard deviations for the two clearly overlap
- Notice that System 1 is always better than System 2

48

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

44

Comparing systems using a paired t test

1. calculate the sample mean

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

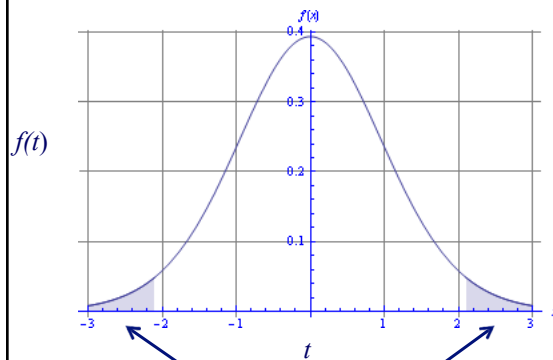
2. calculate the t statistic

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

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

45

Comparing systems using a paired t test



The null distribution of our t statistic looks like this

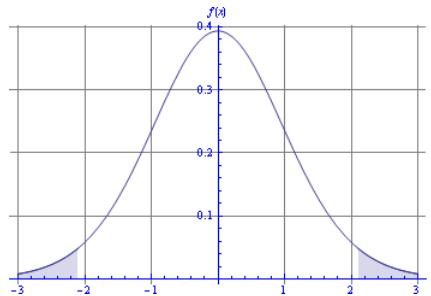
The p -value indicates how far out in a tail our t statistic is

If the p -value is sufficiently small, we reject the null hypothesis, 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

46

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

47

Sign Test

- If less than 300 examples, we won't have 30 test examples per fold
- Prefer leave-one-out cross-validation
- Count “wins” for Algorithm A and B over the N test examples on which they disagree
- Let M be the larger of these counts
- What is probability under $b(N, 0.5)$ that either A or B would win *at least* M times

53

Scatter plots for pairwise method comparison

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

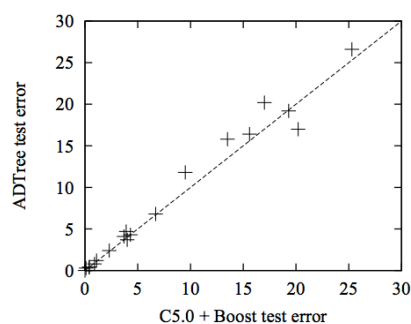


figure from Freund & Mason, *ICML* 1999

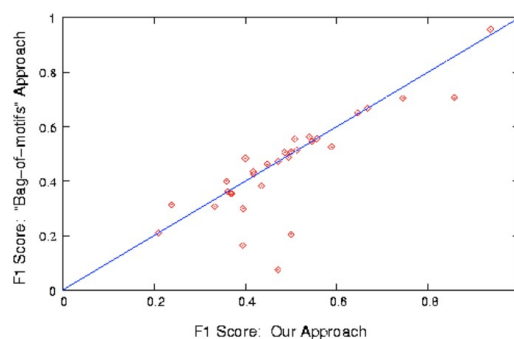


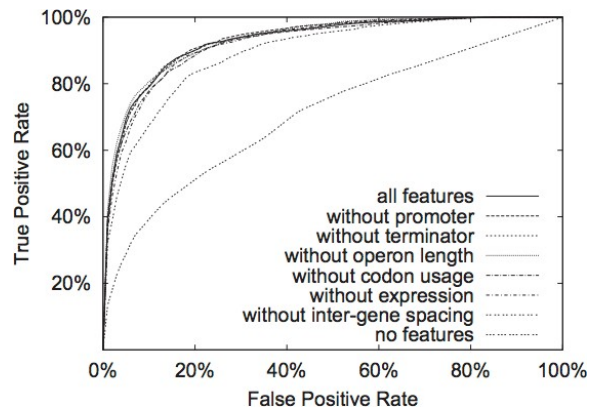
figure from Noto & Craven, *BMC Bioinformatics* 2006

54

Lesion 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



50

figure from Bockhorst et al., *Bioinformatics* 2003