Computer Science 477

Estimating Classifier Accuracy

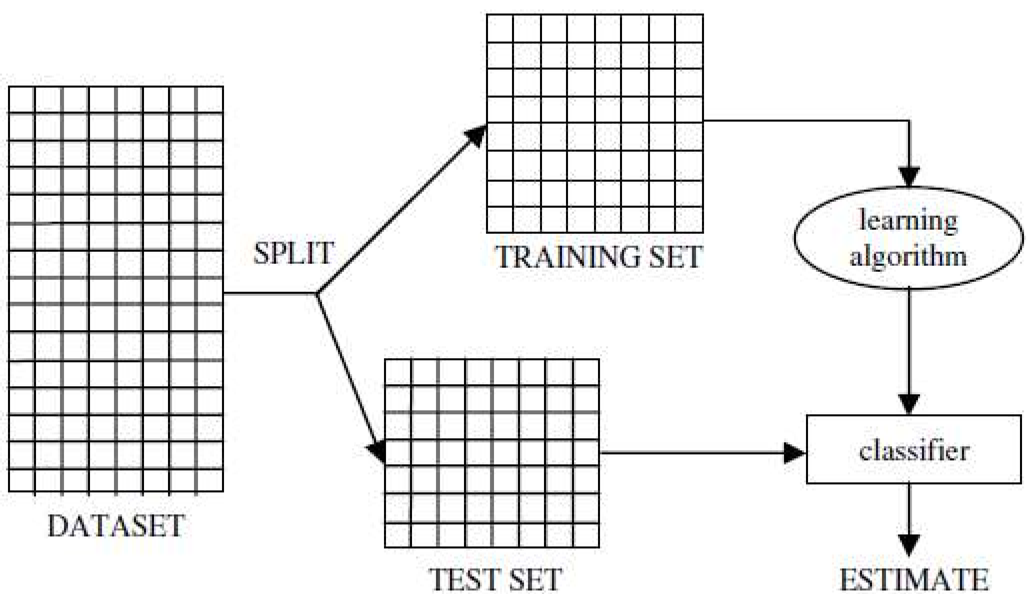
Lecture 8

Predictive Classifier Accuracy

- See how well it works in practice
- Applies to any classifier
  - Illustrated here with tree classifiers.
- *Estimate* the predictive accuracy by measuring its accuracy for a sample of data not used when generated.
- Three methods:
  - Divide into training set and test set
  - k-fold cross-validation
  - N-fold (or leave-one-out) cross-validation.

Separate Training and Test Sets

- Use training to construct a classifier (decision tree, neural net etc.).
- Classifier is used to predict the classification for the instances in the test set.
- Test set contains *N* instances of which *C* are correctly classified the *predictive accuracy* $p = C/N$.



Some datasets (UCI) come with predefined test sets.

Common ratios of data to test: 1:1, 2:1, 70:30, 60:40

Standard Error

- Predictive accuracy is an *estimate* of performance of classifier.
- Find range of values within which the true value of predictive accuracy
- Use *standard error* associated with an estimated value *p*
- If *p* is calculated using a test set of *N* instances, standard error is $\sqrt{p(1-p)N}$
- Standard error enables to assert with a specified probability *p* that the true predictive accuracy is "so-many" standard errors below of above the estimated value of *p*.
- The more certain we wish to be, the greater the number of standard errors.
- The Probability is *confidence level, denoted CL written* Z_{CL}

Confidence Level

- Typical relation between CL and Z_{CL}

Confidence Level (CL)	0.9	0.95	0.99
Z_{CL}	1.64	1.96	2.58

- If the predictive accuracy of a test set is *p*
  - With standard error *S* then
  - With confidence level CL,
  - True predictive accuracy lies in the interval $p \pm Z_{CL} \times S$

Confidence Level Example

- Let 80 of 100 instances be predicted correctly
- Predictive accuracy, $p = 0.8$
- Standard error: $\sqrt{0.8 \times 0.2/100} = \sqrt{0.0016} = 0.04$
- With probability 0.95 the true and predicted accuracy lies in the interval $0.8 \pm 1.96 \times 0.04$, between 0.7216 and 0.8784
- Predictive accuracy also known as error rate
  - If p = 0.9, error rate is 10%

Repeated Train and Test

- Here: classifier used on *k* test sets (not just one)
- If all test sets are the same size, predictive average simply averaged
- Total number of instances *kN*, standard error of the estimate is $\sqrt{p(1-p)/kN}$
- Test sets not the same size, calculation more complicated.

Generalizing

- Given N_i instances in the *i*th test set $(1 \leq i \leq k)$ and the predicted accuracy for the *i*th test set is p_i the overall predictive accuracy p is

$$\sum_{i=1}^{i=k} p_i N_i / T$$

where

$$T = \sum_{i=1}^{i=k} N_i$$

- (weighted average of p_i values)
- Standard error is $\sqrt{p(1-p) \times T}$

k-fold Cross-validation

- Divide dataset of N instances into k equal subset
  - k typically a small number such as 5 or 10.
- (If N is not exactly divisible by k, the final part will have fewer instances than the other $k - 1$ parts.)
- Series of k runs is now carried out.
- Each of the k parts in turn is used as a test set
- Other $k - 1$ parts are used as a training set.
- The total number of instances correctly classified (in all k runs combined)
- Divided by the total number of instances N to give an overall level of predictive accuracy p, with standard error $\sqrt{p(1 - p)/N}$.

Repeated Train & Test

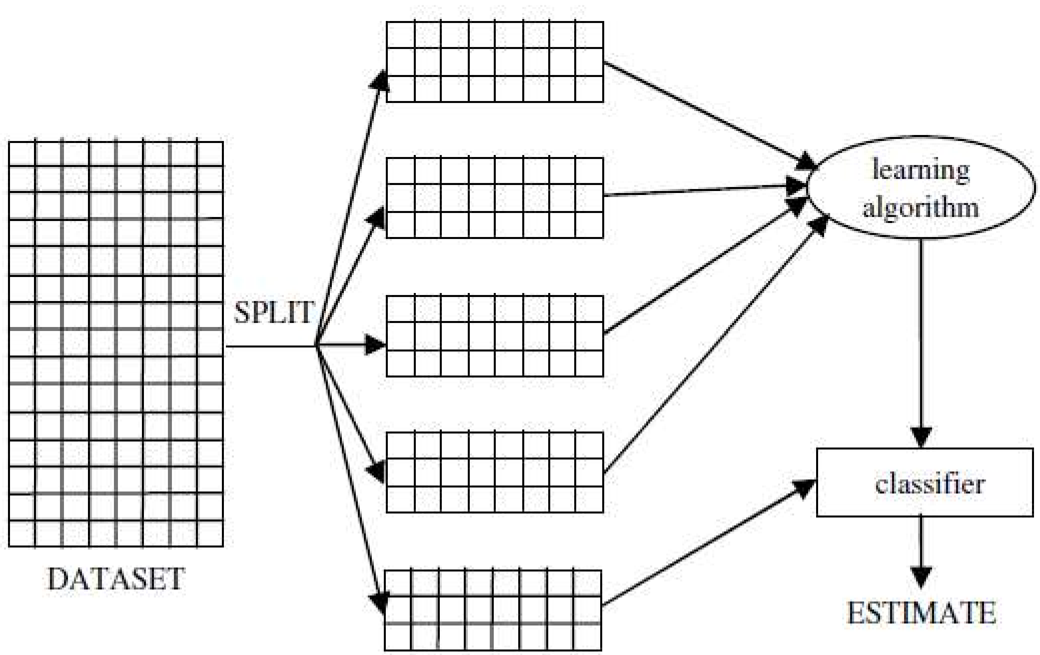
- Classifying k test sets (not just one).
- Average to produce overall estimate of p.
- If total number in each test set is N, standard error p is $\sqrt{p(1-p)/kN}$
- If test sets differ in size:
  - N_i instances in the ith test set $(1 \leq i \leq k)$
  - Predictive accuracy for ith test set is p_i, overall predictive accuracy is

$$\sum_{i=1}^{i=k} \frac{p_i N_i}{T}$$

where $T = \sum_{i=1}^{i=k} N_i$

- Standard error is $\sqrt{p(1-p)/T}$

k-fold Cross-validation



N-fold Cross-validation

- *N*-fold cross-validation is an extreme case of *k*-fold cross-validation, often known as ‘leave-one-out’ cross-validation or jack-knifing
- Dataset is divided into as many parts as there are instances,
  - Each instance effectively forming a test set of size one.
- *N* classifiers are generated, each from $N - 1$ instances, and each is used to classify a single test instance.
- Predictive accuracy *p* is the total number correctly classified divided by the total number of instances.
- Standard error is $\sqrt{p(1 - p)/N}$.

N-fold cross-validation

- Unsuitable for use with large datasets.
- Utility questionable
- Most likely to be of benefit with very small datasets where as much data as possible needs to be used to train the classifier

Datasets with Missing Values

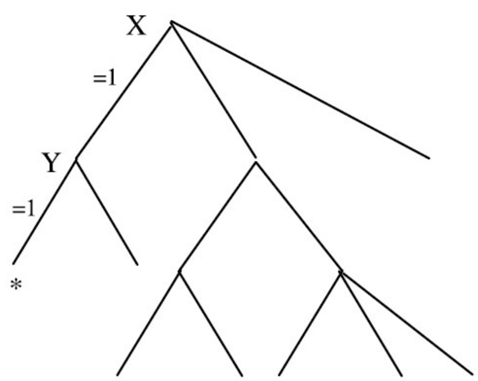
- Discard Instances
- Advantage:
  - Avoid introducing any data errors.
- Disadvantages
  - Discarding data may damage the reliability of the resulting classifier.
  - Cannot be used when a high proportion of the instances in the training set have missing values
  - Not possible with this strategy to classify any instances in the test set that have missing values.
- Replace by Most Frequent or Average Value
  - Works better in practice

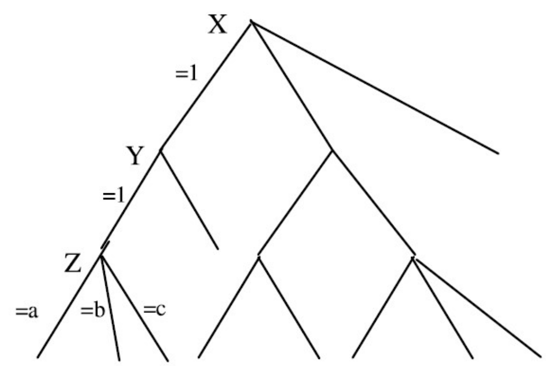
Missing Classifications

- More problematic
- Replace missing class with most common
  - Unsatisfactory in practice
- Generally must disregard such instances.

Note on Missing Branches

- Missing branches can occur at any stage of decision tree generation
- Likely to occur lower down where the number of instances under consideration is smaller
- Suppose that tree construction has reached the following stage:





- Suppose that at $*$ it is decided to split on categorical attribute Z, which has four possible values a, b, c and d, but no instance has value d.
- Cannot classify any new instance that has d for attribute Z

Experimental Results - I

- TDIDT classification of four data sets
- Information gain for attribute selection
- (See appendix B of the text).

Dataset	Description	classes	attributes$^+$		instances	
			categ	cts	training set	test set
vote	Voting in US Congress in 1984	2	16		300	135
pima-indians	Prevalence of Diabetes in Pima Indian Women	2		8	768	
chess	Chess Endgame	2	7		647	
glass	Glass Identification	7		9*	214	

- Three of four datasets from UCI repository.

Experimental Results - I

- Vote datasets has separate training and test sets
- Other three: every third instance reserved for a test set

Dataset	Test set (instances)	Correctly classified	Incorrectly classified	Unclassified
vote	135	126 (93% $\pm$ 2%)	7	2
pima-indians	256	191 (75% $\pm$ 3%)	65	
chess	215	214 (99.5% $\pm$ 0.5%)	1	
glass	71	50 (70% $\pm$ 5%)	21	

- Unclassified instances assigned a default classification (largest class)
- Unclassified instances rare, various rival policies inconsequential

10-fold, N-fold Cross-Validation

- 10-fold Cross-Validation:

Dataset	Instances	Correctly classified	Incorrectly classified
vote	300	275 (92% $\pm$ 2%)	25
pima-indians	768	536 (70% $\pm$ 2%)	232
chess	647	645 (99.7% $\pm$ 0.2%)	2
glass	214	149 (70% $\pm$ 3%)	65

- N-fold Cross Validation:

Dataset	Instances	Correctly classified	Incorrectly classified
vote	300	278 (93% $\pm$ 2%)	22
pima-indians	768	517 (67% $\pm$ 2%)	251
chess	647	646 (99.8% $\pm$ 0.2%)	1
glass	214	144 (67% $\pm$ 3%)	70

Experimental Results – Missing Values - II

- TDIDT with information gain:

categ = categorical
cts - continuous

( ) – at least one missing value

Dataset	Description	classes	attributes+		instances	
			categ	cts	training set	test set
crx	Credit Card Applications	2	9	6	690 (37)	200 (12)
hypo	Hypothyroid Disorders	5	22	7	2514 (2514)	1258 (371)
labor-ne	Labor Negotiations	2	8	8	40 (39)	17 (17)

- Two strategies for missing values
  - Discard instances
  - Replacement
    - Most frequent
    - Average

Strategy 1 – Discard Instances

- Advantage: don't introduce data distortions
- Disadvantage: lose information
- Large proportion of missing attribute values – can't use
  - Labor negotiations
  - Hyperthyroid disorders
- Applied to crx dataset:

Dataset	MV strategy	Rules	Test set	
			Correct	Incorrect
crx	Discard Instances	118	188	0

- Correctly classifies all 188 complete test set

Strategy 2: Replace by most frequent/average

- Categorical – replace by most common attribute value
- Continuous – replace with average value

Dataset	MV strategy	Rules	Test set	
			Correct	Incorrect
crx	Discard Instances	118	188	0
crx	Most Frequent/Average Value	139	200	0

- All 200 of the test set correctly classified.

Replacement with Hyperthyroid & Labor Negotiations

- Hyperthyroid disorders:

Dataset	MV strategy	Rules	Test set	
			Correct	Incorrect
hypo	Most Frequent/Average Value	15	1251	7

- Classifies correctly 1251 of 1258 (99%)
- Impressive, since every single instance has missing attribute values
- Labor Negotiations:

Dataset	MV strategy	Rules	Test set	
			Correct	Incorrect
labor-ne	Most Frequent/Average Value	5	14	3

- Correctly classifies 14 of 17 instances in training set

Confusion Matrix

- Displays how frequently instances of class *X* were correctly classified as class *X* or misclassified as some other class.

Correct classification	Classified as democrat	Classified as republican
democrat	81 (97.6%)	2 (2.4%)
republican	6 (11.5%)	46 (88.5%)

- Confusion Matrix for a Binary Classification
- 81 correctly classified as Democrat, 2 Democrats incorrectly classified as Republican
- 6 Republicans incorrectly classified as Democrat, 46 correctly classified as Republican

Confusion Matrix with non-binary classifications

- Six classifications:

Correct classification	Classified as 1	2	3	5	6	7
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
5	0	2	0	10	0	1
6	0	1	0	0	7	1
7	1	3	0	1	0	24

- 52 1s correctly classified, 10 1s incorrectly classified as 2, 7 as 3, 1 as 7.
- 24 7s correctly classified, 1 incorrectly classified as 1, 3 as 2, 1 as 5.

Confusion Matrix

Correct classification	Classified as	
	+	−
+	true positives	false negatives
−	false positives	true negatives

- Confusion matrix interpretation
- When two classes: one regarded as positive
  - Class of especial interest.

Value of TP, FN, FP, TN

- TP, FN, FP, TN Rate not depend on the relative sizes of *P* and *N*.
  - Similarly: any combination of two ‘rate’ values calculated from *different* rows of the confusion matrix

- Predictive Accuracy and other measures from values in *both* rows of the table are affected by the relative sizes of *P* and *N*
  - Can be a serious weakness.

Example – Driving Test

- Positive class corresponds to those who pass a driving test at the first attempt
  - Negative class corresponds to those who fail.
- Relative proportions in the real world are 9 to 10
  - Test set correctly reflects this.
- Implied confusion matrix:

		Predicted class		Total instances
		+	−	
Actual class	+	8,000	1,000	9,000
	−	2,000	8,000	10,000

- True positive rate of 0.89 and a false positive rate of 0.2
  - Assume: a satisfactory result

Example – Driving Test

- Suppose that the number of successes grows
  - Because of improved training,
  - Higher proportion of passes.
- Possible confusion matrix:

		Predicted class		Total instances
		+	−	
Actual class	+	8,000	1,000	9,000
	−	2,000	8,000	10,000

		Predicted class		Total instances
		+	−	
Actual class	+	80,000	10,000	90,000
	−	2,000	8,000	10,000

- Both confusion matrices the values of TP Rate and FP Rate are the same
  - (0.89 and 0.2 respectively).
- Values of the Predictive Accuracy measure are different.
- For the original confusion matrix, Predictive Accuracy is 16,000/19,000 =
- 0.842. For the second one, Predictive Accuracy is 88,000/100,000 = 0.88.

Driving Test – Alternative Scenario

- A large increase in the relative proportion of failures
  - Because of an increase in the number of younger people being tested.
- Possible confusion matrix:

		Predicted class		Total instances
		+	−	
Actual class	+	8,000	1,000	9,000
	−	20,000	80,000	100,000

- Predictive Accuracy is now 88,000/109,000 = 0.807.
- TP, FP Rate invariant
- FP Rate values would be the same.
- Three Predictive Accuracy values vary from 81% to 88%,
  - Reflecting changes in the relative numbers of positive and negative values in the test set, rather than any change in the quality of the classifier.

Non-binary Confusion Matrix

Correct classification	Classified as 1	2	3	5	6	7
1	52	10	7	0	0	1
2	15	50	6	2	1	2
3	5	6	6	0	0	0
5	0	2	0	10	0	1
6	0	1	0	0	7	1
7	1	3	0	1	0	24

- Confusion matrix for a classification with seven possible values