CS522 Advanced Database Systems
Classification: Evaluation and Comparison of Classifiers

Chengyu Sun
California State University, Los Angeles

Difficulties in Classifier Evaluation and Comparison

- Training error is not a good indicator of testing error
- Data with known class labels are often in short supply
- *Costs* of errors need to be taken into account
- Evaluation results must be evaluated themselves

Accuracy and Error Rate

- Accuracy `Acc(M)` of a classifier on a given test set is the percentage of the test records that are correctly classified
- Error rate: `1-Acc(M)`

Confusion Matrix

Actual Class \ Predicted Class	cancer	not cancer
cancer	20	5
not cancer	10	1000

Confusion Matrix for Binary Classification

Actual Class \ Predicted Class	C1	C2
C1	true positive (`t_pos`)	false negative (`f_neg`)
C2	false positive (`f_pos`)	true_negative (`t_neg`)

C1 is the main class of interests

Costs of Misclassification

- The costs (or risks) of a false negative tend to be far greater than that of a false positive
  - E.g. cancer vs. not cancer

Precision and Recall

$$precision = \frac{t\_pos}{t\_pos + f\_pos}$$

$$recall = \frac{t\_pos}{t\_pos + f\_neg}$$

Sensitivity and Specificity

- Sensitivity = Recall
- Specificity = `t_neg/(t_neg+f_pos)`

Accuracy Measure Examples

		Predicted Class	
		cancer	not cancer
Actual Class	cancer	20	5
	not cancer	10	1000

- Accuracy and error rate??
- Precision and Recall??
- Sensitivity and specificity??

Utilizing Records with Known Class Labels

- For both training and testing
  - More training records → better classifier
  - More testing records → better accuracy estimate

The Holdout Method

- Randomly partition the given records into two non-overlapping subsets: a training set and a testing set
  - Typically 2/3 for training and 1/3 for testing

Problems of the Holdout Method

- More records for training means less for testing, and vice versa
- Distribution of the data in the training/testing set may be different from the original dataset
- Some classifiers are sensitive to random fluctuations in the training data

Random Subsampling

- Repeat the holdout method k times
- Take the average accuracy over the k iterations
- Random subsampling methods
  - Bootstrap Method
  - Cross-validation

Bootstrap Method

- Each iteration uses a sample to train the classifier, and the remaining records for testing
- Uniform sampling with replacement – *bootstrapping*
  - The sample record may be selected more than once

.632 Bootstrap …

- Select d samples out of a dataset of size d and use them as the training set, and the rest are used for testing
- On average, 63.2% of the records will be selected into the training set
- The probability of not being selected:

$$(1-1/d)^d \xrightarrow[d\to\infty]{} e^{-1} = 2.718^{-1} = 0.368$$

… .632 Bootstrap

- Overall accuracy over k iterations

$$\frac{1}{k}\sum_{i=1}^{k}(0.632\times Acc(M_i)_{test\_set} + 0.368\times Acc(M_i)_{all\_records})$$

K-fold Cross Validation

- Randomly divide the data into k non-overlapping subsets of roughly equal size called *folds*
- Each iteration uses (k-1) subsets for training, and the remaining subset for testing

Variants of K-fold Cross Validation

- Stratified folds: the class distribution in each fold is roughly the same as in the original dataset
- Leave-one-out
- 10-fold Cross Validation

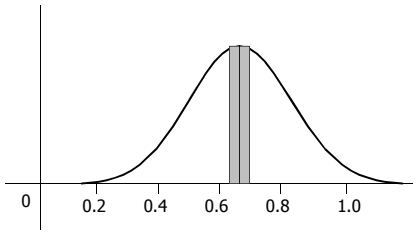
Accuracy Using K-fold Cross Validation

$$\frac{\text{Total \# of correctly classified records over k iterations}}{\text{Total \# of records in the original dataset}}$$

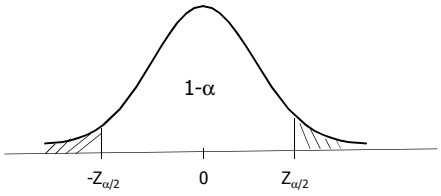
Confidence Interval

- Accuracies are *estimated*
- We want to say something like: the accuracy is in the range of 0.66±0.04 with 99% confidence
  - Confidence interval: 0.66±0.04
  - Degree of confidence (a.k.a. confidence level): 99%

Probabilistic Distribution of Accuracy



Confidence Interval of Standard Normal Distribution



1-α	0.99	0.98	0.95	0.9	0.8	0.7	0.5
$Z_{\alpha/2}$	2.58	2.33	1.96	1.65	1.28	1.04	0.67

Confidence Interval for Accuracy

$$\frac{2N \bullet Acc + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + 4N \bullet Acc - 4N \bullet Acc^2}}{2(N + Z_{\alpha/2}^2)}$$

`N`: number of testing records
`Acc`: accuracy

Confidence Interval Examples

- Accuracy: 80%
- Confidence level (i.e. `1-α`): 95%

N	Confidence Interval
50	[0.67, 0.89]
100	[0.71, 0.87]
500	[0.76, 0.83]
1000	[0.77, 0.82]

Comparing Classifiers

- Is a classifier with 72% accuracy better than one with 68% accuracy? Or in other words, is the 4% difference statistically significant?

t-test

- Test whether the means of two normally distributed populations are the same
  1. Null hypothesis: $\mu_1 - \mu_2 = 0$
  2. Choose significance level: α
  3. Calculate t statistic
  4. Reject hypothesis if $t > t_{v,\alpha}$ where v is the degree of freedom

t-test Example

- Two classifiers M_1 and M_2
- 5-fold cross validation, and at each round M_1 and M_2 use the same training/testing partition(s)

Round	1	2	3	4	5
Acc(M_1)	0.62	0.47	0.70	0.72	0.69
Acc(M_2)	0.57	0.57	0.63	0.53	0.72

t Statistic

Accuracy difference in round i: $d_i = Acc(M_1)_i - Acc(M_2)_i$

Mean accuracy difference: $\bar{d} = \frac{1}{k}\sum_{i=1}^{k} d_i$

Standard deviation of accuracy differences: $S_d = \sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(d_i - \bar{d})^2}$

t Statistic: $t = \frac{\bar{d}}{S_d / \sqrt{k}}$

t Statistic Calculation Example

Round	1	2	3	4	5
Acc(M_1)	0.62	0.47	0.70	0.72	0.69
Acc(M_2)	0.57	0.57	0.63	0.53	0.72
d_i	0.05	-0.10	0.07	0.19	-0.03

$\bar{d} = 0.036$

$S_d = 0.109$

$t = \frac{0.036}{0.109/\sqrt{5}} = 0.74$

t Value Table Lookup

α	0.1	0.05	0.025	0.01	0.005	v
	1.53	2.13	2.78	3.75	4.60	4
	1.38	1.83	2.26	2.82	3.25	9

- In our example
  - Choose significance level $\alpha = 0.05$
  - Degree of freedom $v = k - 1 = 4$

Readings

- Textbook 6.12, 6.13, and 6.15