Data Mining

Chapter 5. Credibility: Evaluating What's Been Learned

Evaluating how different methods work

- **Evaluation**
  - Large training set: no problem
  - Quality data is scarce.
    - Oil slicks: a skilled & labor-intensive process
    - Credit card application: 1,000 training examples
    - Electricity supply data: few days / 15 years
    - Electromechanical diagnosis: 300 examples / 20 years

Training and testing

- **Classifier's performance**
  - Error rate
  - Resubstitution error
    - Resubstituting the training instances into a classifier
    - Useful to know
  - Test set
    - Assumption : both the training data and the test data are "representative samples"

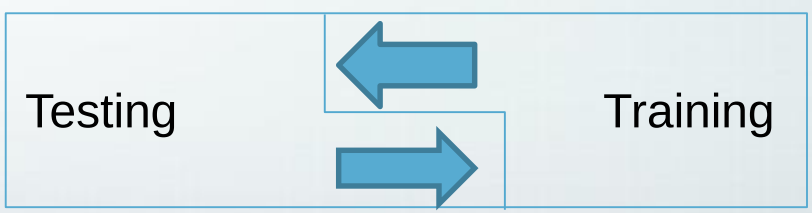
Training and testing

- Training, validation, and test data
  - training → validation → test
  - Validation data is bundled back into the training data.
  - Test data is bundled back into the training data.

A limited dataset

- Holdout procedure

To find a good classifier



To obtain a good error estimate

Predicting performance

- **Bernoulli process**
  - A succession of independent events that either succeed or fail
  - e.g.) coin tossing: an independent event
    - Head: success, Tail: failure
    - True(unknown) success rate: P
    - The number of trials: N
    - The number of successes: S
    - The observed success rate: f = $\frac{S}{N}$

Predicting performance

The Bernoulli distribution

- Mean: P (success rate)
- Variance: P(1-P)

A single Bernoulli trial

- Expected success rate: f = $\frac{S}{N}$
- Variance with N trials: $\frac{P(1-P)}{N}$

N trials

- The probability
  - A random variable : X
  - P_r [-z ≤ X ≤ z] = c where 2z is confidence range

Predicting performance

- One-tailed probability
  - $P_r [X \geq Z]$: upper tail
  - $P_r [X \leq -Z]$: lower tail

  The same

  - e.g.) $P_r [X \geq Z]$: 5%
    - There is a 5% chance that X lies more than 1.65 standard deviations above the mean (refer to Table 5.1)
    - $P_r [-1.65 \leq X \leq 1.65] = 90\%$

Predicting performance

- Bernoulli distribution
  - P_r [-z < $\frac{f-P}{\sqrt{\frac{P(1-P)}{N}}}$ < Z] = c
    - f: random variable (x or expected success rate)
    - P: mean
    - $\sqrt{\frac{P(1-P)}{N}}$ = variance with N trials
    - P = (f + $\frac{Z^2}{2N}$ $\pm$ Z $\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{Z^2}{4N^2}}$) / (1+$\frac{z^2}{N}$)

Predicting performance

- e.g.)
  - If f=75% (success rate), N=1,000 and C=80% (confidence) (z=1.28),
    then P = [0.732, 0.767].
    → 73.2% < P < 76.7%

  - If f=75%, N=100, C=80%
    then P = [0.691, 0.801]
    → 69.1% < P < 80.1%

Cross-validation

- **Cross-validation**
  - When the amount of data for training and testing is limited
  - Holdout method
    - Testing : 1/3 data
    - Training : 2/3 data
  - Repeated holdout
    - Average error rates → an overall error late!

Cross-validation

- Cross-validation
  - A fixed number of folds
    - Folds : “partitions” of data
  - e.g.) threefold cross-validation (3 parts)
    - 2/3 folds : training
    - 1/3 folds : testing

  3번 시행

- 10 fold cross-validation (10 parts)
  - 9/10 : training
  - 1/10 : testing

  10번 시행

  - A total of 10 times on different training sets
  - 10 error estimates are averaged to yield an overall error estimate

Cross-validation

Leave-one-out cross validation

- n-fold cross-validation

  where n: the number of instances in the dataset

  - Each instance in turn is left out.
  - Learning scheme is trained on all the remaining instances.
  - The results of all n judgments are averaged. → the error estimate
  - Advantages
    - The greatest possible amount of data is used for training in each case.
    - The procedure is deterministic → no random sampling
  - Disadvantages
    - High computational cost
    - No stratification: test vs training

Cross-validation

The bootstrap

- Sampling the dataset with replacement to form a training set
  - Most learning methods can use the same instance twice.

- 0.632 bootstrap
  - Being picked for the training set : 1/n probability
  - Not being picked for the training set : (1-1/n) probability
  - The number of picking opportunities : n

Cross-validation

- The chance that a particular instance will not be picked for the training set :
  - $(1-\frac{1}{n})^n \approx e^{-1} = 0.368$

    where e = 2.7183.

  - Test set : 36.8% of the instances
  - Training set : 63.2% the instances
  - Some instances will be repeated in the training set, bringing it up to a total size of n.

Cross-validation

- Bootstrap vs cross-validation
  - Bootstrap : 63%
  - 10-fold cross-validation : 90%

  - Boot strap error estimate
    - $e = 0.632 \times e_{\text{test instances}} + 0.368 \times e_{\text{training instances}}$
    - The whole bootstrap procedure is repeated several times, with different replacement samples for the training set, and the results averaged.

Comparing data mining methods

Analysis of variance

- Deciding whether observed differences among more than two sample means can be attributed to chance, or whether there are real differences among the means of the populations sampled.

- F distribution with k-1 and k(n-1) degrees of freedom
  - We reject the null hypothesis that the population means are all equal, if the value we obtain for *f* exceeds $f_{\alpha,\, k-1,\, k(n-1)}$, where α is the level of significance.

Thank You !

- **http://cis.catholic.ac.kr/sunoh**