Session #12

CMSC 409: Artificial Intelligence

http://www.people.vcu.edu/~mmanic/

Virginia Commonwealth University,
Fall 2023,
Dr. Milos Manic
(mmanic@vcu. edu)

1

CMSC 409: Artificial Intelligence

Session # 12

Topics for today

- Announcements
- Previous session review
- Extracting models for data classification & prediction
- Two phases or data classification
- Classification
  - *Classification vs. clustering*
  - *Supervised vs. unsupervised learning*
  - *Training vs. testing data set*

2

CMSC 409: Artificial Intelligence

Session # 12

Topics for today (cont.)

- Classification Accuracy (efficiency, scalability)
  - *Accuracy estimation*
    - *holdout and random subsampling*
    - *k-fold cross-validation*
    - *bootstrap*
  - *Accuracy improvement*
    - *bagging (bootstrap aggregation)*
    - *boosting the accuracy (adaboost)*

CMSC 409: Artificial Intelligence

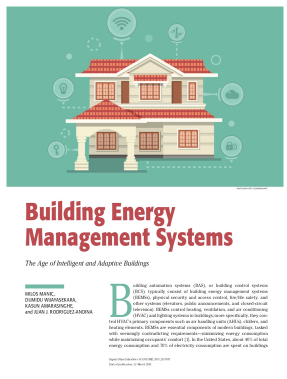
Announcements Session # 12

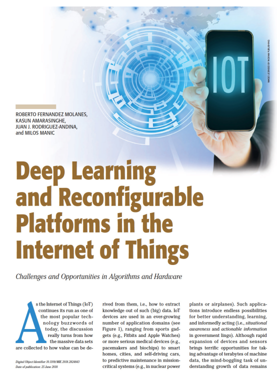
- Canvas
  - *New slides posted*
- Office hours zoom
  - *Zoom disconnects me after 45 mins of inactivity. Feel free to chat me via zoom if that happens and I will reconnect (zoom chat welcome outside of office hours as well)!*
- Project #2
  - *Deadline Oct. 3 (noon); Review a week from the deadline.*
- Midterm exam
  - *Oct. 19 (Thu)*
- Paper (optional)
  - *The 2nd draft due Oct. 10 (noon)*
  - *Literature review and updated problem description (check out the class paper instructions for the 2nd draft)*
- Subject line and signature
  - *Please use [CMSC 409] Last_Name Question*

“Smart” => Connected!

Connected = Data Exchange (can learn, can be intelligent)

But if connected, is it secure?





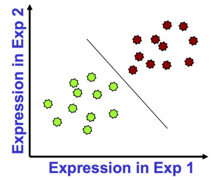
5

Classification

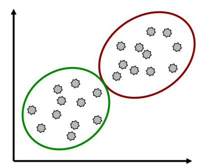
6

Clustering vs. Classification

- **Objects**
  - *Patterns, points in space, characterized by dimensions (features, attributes) Examples?*
- **Classification**
  - ***Has labels** for points, has predefined classes*
  - *Can use a rule that assigns labels to new points*
  - *Sometimes referred to **supervised** learning*



- **Clustering**
  - ***No labels**, no predefined classes*
  - *Group points into clusters based on interrelationships, structure in data*
  - *Some metrics of similarity*
  - *Sometimes referred to as **unsupervised** learning*



7

Extracting models for data classification & prediction

Two phases or data classification

Classification

- *Supervised learning*
- *Unsupervised learning*

Classification Accuracy (efficiency, scalability)

- *Training and testing*
- *Accuracy*
- *Boosting the accuracy*

8

Classification & Prediction

Extracting models for classification & prediction

- **Data Classification**
  - *Performs categorical, discrete, unordered categorization. Examples?*
- **Data Prediction**
  - *Models ordered values, continuous valued functions, also known as **numeric prediction** (also predictive classification). Examples?*
- **Two phases of data classification -**
  - *Training phase*
    - *Creating a model*
    - *Learning from a training set*
    - *Creation of classification rules*
  - *Testing phase*
    - *Sometimes used validation phase (model sel.); testing (of final model)*
    - *done on different data set than training set; sometimes 50/25/25*
  - *Execution phase (can be adaptive)*
    - *Acting based on the rules learned*

Classification & Prediction

Classification

- Training set
  - *instances, samples, examples, data points, objects*
  - *attribute or feature vector (measurements) is an n-tuple:*

$$X = (x_1, x_2, \ldots, x_n)$$

  - *on database attributes*

$$A = (A_1, A_2, \ldots, A_n)$$

  - *Class label attributes*
    - *Database of discrete-valued, unordered values*

Extracting models for data classification & prediction

Two phases or data classification

Classification

- *Supervised learning*
- *Unsupervised learning*

Classification Accuracy (efficiency, scalability)

- *Training and testing*
- *Accuracy*
- *Boosting the accuracy*

Classification & Prediction

Classification *(in literature, you may find reference to both…)*

- **Two types**
  - Supervised learning
    - *Teacher*
    - *We know the desired outcome*
      - *Number of classes,*
      - *Description of classes*
      - *Behavior of classes*
  - Unsupervised learning
    - *The knowledge about the outcome does not exist*
      - *Learning on the fly*
      - *Discovering of new classes*
      - *Dynamic, adaptive learning*
  - Classification
    - *Mapping of DB measurement tuple to class label*
    - *Classification rule, decision trees, ANNs,…*

Extracting models for data classification & prediction

Two phases or data classification

Classification

- *Supervised learning*
- *Unsupervised learning*

Classification Accuracy (efficiency, scalability)

- *Training and testing*
- *Accuracy*
- *Boosting the accuracy*

Classification & Prediction

Classification Accuracy (efficiency, scalability)

- **Training and testing** data sets
  - *Need to be* ***different*** *but* ***consistent***
  - *Risks: overfitting*
- **Accuracy**
  - *The percentage of test set tuples correctly classified*
  - *Error measures; confidence limits;*
  - *Techniques:*
    - *Cross-validation*
    - *Bootstrap*
    - *Comparison of different techniques*
  - *Unsupervised learning?*
- **Boosting the accuracy**
  - *Bagging*
  - *Boosting*

Classification & Prediction

Classification & Prediction

- **Issues (cont.)**
  - **Data preparation**
    - *Data cleaning – important!*
      - *Noise*
      - *Missing values (fill out with some value)*
    - *Relevancy analysis*
      - *Correlation analysis*
      - *Attribute subset selection*
        - *Relevant, irrelevant, redundant attributes*
  - **Data transformation**
    - *Normalization (scaling)*
    - *Generalization to higher level concepts*
      - *Day, month, year; Low, medium, high;*
      - *Concept hierarchies; fuzzy logic;*
      - *Results in data compression, storage/comput. savings, accuracy*

Classification & Prediction

Extracting models for classification & prediction

Classification & Prediction Algorithms:

- *Decision trees*
- *Bayesian belief networks*
- *Rule-based*
- *ANNs, fuzzy logic, genetic algorithms*
- *Association rules*
- *Support vector machines (SVM)*
- *Lazy learners (learning from neighbors)*
- *Regression*

Accuracy
Estimation and Improvement

Techniques for accuracy estimation

- ❑ Holdout and random subsampling
- ❑ k-fold cross-validation
- ❑ Bootstrap

Techniques for accuracy improvement

- ❑ *Bagging*
- ❑ *Boosting*

Accuracy & Error Measures

Evaluating the accuracy of a classifier or predictor

- *Holdout and random subsampling*
- *k-fold cross-validation*
- *Bootstrap to measure predictor accuracy?*

Holdout method

- The data set is partitioned into two sets
  - *2/3 into training, 1/3 into test set*
  - *testing is carried out only after the classifier has been fully trained*

Random subsampling

- Holdout method is repeated *k* times
  - *Overall accuracy estimate is the average of accuracies from each iteration*
  - *Multiple random train and test experiments are performed on training and testing samples. The overall data set is also split into two sets. The random train and test experiments are carried out simultaneously over a number of iterations.*

Techniques for accuracy estimation

- ❑ Holdout and random subsampling
- ❑ k-fold cross-validation
- ❑ Bootstrap

Techniques for accuracy improvement

- ❑ *Bagging*
- ❑ *Boosting*

Accuracy & Error Measures

Cross-validation

- *Model validation techniques for assessing the "generalization" capability to an independent data set (example: EBaLM-OTR paper).*

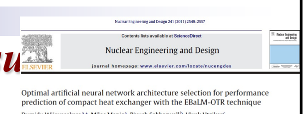
k-fold cross-validation

- Initial data set is randomly partitioned in *k* mutually exclusive subsets (folds) of approximately equal size ($D_1, D_2, \ldots, D_k$)
  - ***training and testing is performed k times in following fashion:***
    - *$D_2, \ldots, D_k$ used as a training set, and D_1 used as testing set*
    - *$D_1, D_3, \ldots, D_k$ used as a training set, and D_2 used as testing set … …*
    - *$D_1, D_2, \ldots, D_{k-1}$ used as a training set, and D_k used as testing set*
  - *each fold is used the same number of times for training and testing.*
  - *10 fold recommended for estimating accuracy*
- *Accuracy: overall number of correct estimates divided by total number of data patterns*
- *Prediction: total loss (o-d) divided by total number of data patterns*

http://www.sgi.com/tech/mlc/util/util/node6.html

D. Wijayasekara, M. Manic, P. Sabharwall, V. Utgikar, "Optimal artificial neural network architecture selection for performance prediction of compact heat exchanger with the EBaLM-OTR technique," in Nuclear Engineering and Design, vol. 241, no. 7, pp. 2549–2557, July 2011

Accuracy & Error Measures



Cross-validation

- *Model validation techniques for assessing the "generalization" capability to an independent data set (example: EBaLM-OTR).*

(prediction of cold sided pressure drop for compact printed circuit heat exchangers (PCHE) in nuclear power plants)

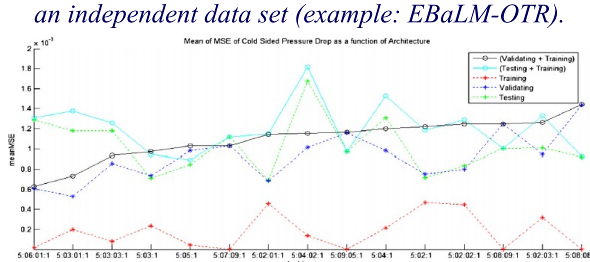


Fig. 6. Conforming capability of network architectures for cold sided pressure drop.

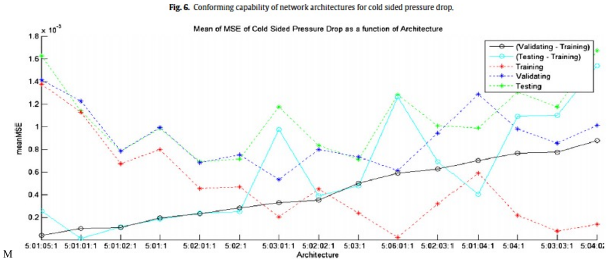


Fig. 7. Generalization capability of network architectures for cold sided pressure drop.

- *5-fold validation*
- *5:3:1:1 (2nd/7th)*
- *5:6:1:1 (1st/10th), in fact, 5:3:1:1 (2nd/7th) – slightly worse conformance much better generalization*
- *5:2:1:1 (7th/5th) optimal…*

Accuracy & Error Measures

Exhaustive cross-validation

- *Cross-validation methods which learn and test on all possible ways to divide the original sample into training and validation data sets*

Non-exhaustive cross-validation

- *Not all ways of splitting original data set*
- *Variants of leave-p-out cross validation*

Leave-one-out

- *Special case of k-fold cross-validation (k = number of initial data patterns)*
- *One sample is left out (validation/test set)*
- *Leave-p-out*

Stratified cross-validation

- *Folds are stratified so that they contain approximately the same proportions of labels as the original dataset.* ***Very important one!***

Techniques for accuracy estimation

- ❑ Holdout and random subsampling
- ❑ k-fold cross-validation
- ❑ Bootstrap

Techniques for accuracy improvement

- ❑ *Bagging*
- ❑ *Boosting*

Accuracy & Error Measures

Bootstrapping

- The name *bootstraping*
  - *The name "bootstrapping" comes from the phrase, "To lift himself up by his bootstraps." This refers to something that is preposterous and impossible. Try as hard as you can, you cannot lift yourself into the air by tugging at pieces of leather on your boots.*
- Statistical technique
  - (Resampling) to estimate a population.
  - Used to determine the value of a parameter of a population

http://statistics.about.com/od/Applications/a/What-Is-Bootstrapping.htm

© M. Manic, CMSC 409: Artificial Intelligence, F23

Accuracy & Error Measures

Bootstrapping

- Statistical technique
  - (Resampling) to estimate a population.
  - Used to determine the value of a parameter of a population
- Example: ***chocolate factory***
  - *Need to guarantee mean weight of candy bars*
    - *Cannot weigh every candy produced, so use sampling; randomly choose 100 candy bars, find the mean (if mean falls within margin of tolerance, good).*
  - *Few months later…*
    - *Try to find with greater accuracy what the mean bar weight was on the day when we sampled?*
    - *Now everything different (atmospheric conditions, employees, batches of milk, sugar, cocoa); No time machine; But we still have those 100 weights!*
  - *Use Boostraping:*
    - *Sample with replacements from 100 known weights (boostrap sample)*
    - *Thousands of boostrap samples can be constructed in short period of time*

http://statistics.about.com/od/Applications/a/What-Is-Bootstrapping.htm

© M. Manic, CMSC 409: Artificial Intelligence, F23

Accuracy & Error Measures

Bootstrap

- Samples for the training patterns are chosen uniformly with replacements
  - *pattern (tuple) is equally likely to be selected again and trained upon*
- .632 bootstrap
  - *Data set is sampled* ***d*** *times with replacements -> bootstrap (training) set of* ***d*** *samples; The patterns that were not used for training are used for the testing set.*
  - *63.2% of the original patterns will end up in the bootstrap, and 36.8% in test set. The probability of one pattern being chosen is 1/d; if process repeated d times and with* ***d*** *large, then…:*

$$\lim_{d\to\infty}\left(1-\frac{1}{d}\right)^d = e^{-1} \approx .368$$

  - If sampling procedure is repeated k times where in each iteration accuracy estimate is made on the current test set, the overall accuracy is:

$$Accuracy(M) = \sum_{i=1}^{k}\left(0.632 \cdot Accuracy(M_i)_{test\_set} + 0.368 \cdot Accuracy(M_i)_{train\_set}\right)$$

(accuracy of the model obtained with bootstrap sample i when applied to test set i, and accuracy of the model obtained with bootstrap sample i when applied to the original train. set of data patterns)

Techniques for accuracy estimation

- ❑ Holdout and random subsampling
- ❑ k-fold cross-validation
- ❑ Bootstrap

Techniques for accuracy improvement

- ❑ *Bagging*
- ❑ *Boosting*

Accuracy Improvement

Bagging & Boosting

- Ensemble methods for accuracy improvement (meta algorithms)
- Ensemble methods – combination of *k* learned models (classifiers or predictors) with the idea of forming an improved composite model.

Bagging (bootstrap aggregation)

- E.g. majority vote on best diagnosis made by a large group of doctors
- Algorithm
  - *D - total set of patterns;*
  - *In each iteration i (i=1,k),*
    - *D_i training set of d patterns is sampled w/ replacements from a total set of patterns D (each training set is a bootstrap sample, w/ some patterns occurring more often than others)*
    - *A classifier model M_i is trained on set D_i*
  - *When fed an unseen pattern X, each classifier M_i returns its prediction (one vote)*
  - *Finally, the bagged classifier M_* makes a decision based on majority of votes.*

29

Accuracy Improvement

Bagging (bootstrap aggregation)

- E.g. majority vote on best diagnosis is made by a large group of doctors
- Bagged classifier (theoretically):
  - *Has greater accuracy than single classifier trained on the original training set D*
  - *Is more robust to noisy data*
  - *Composite model reduces the variance of individual classifiers*
  - *For prediction – bagged predictor always has improved accuracy over a single predictor derived from D.*

30

Accuracy Improvement

Boosting

- E.g. of doctors and diagnosis
  - *Final diagnosis is a combination of weighted diagnosis (diagnosis is weighted higher based on the accuracies of diagnoses that doctor made in the past)*
- Algorithm
  - *Weights assigned to each training pattern*
  - *k classifiers iteratively trained:*
    - *After each classifier model M_i is trained, "harder" to train patterns (patterns that were misclassified) are assigned higher weights.*
  - *Classifier M_* combines the votes of each classifier (weight of each vote is a function of its accuracy)*

31

Accuracy Improvement

Boosting

- E.g. of boosting algorithm: ***Adaboost***
  - *Used in case the classification method needs boosting of its accuracy*
  - *D - total set of labeled patterns*
    - *(X_1, y_1),…. (X_d, y_d), where y_i is the class label of pattern X_i*
    - *each training pattern is assigned weight 1/d*
  - *In each iteration i (i=1,k),*
    - *D_i training set of d patterns is sampled w/ replacements from total set D*
      - *each training set is a bootstrap sample, w/ some patterns occurring more often than others*
      - *the chance of a pattern being selected is based on its weight*
    - *A classifier model M_i is trained on set D_i*
      - *its error is calculated using D_i as a **test set** now*
      - *weights of training patterns are adjusted (if incorrectly classified, its weight is increased)*
      - *pattern's weight therefore indicates the degree of difficulty of classifying it*
    - *go back and build another classifier M_i.*

32

Accuracy Improvement

Boosting

- E.g. of boosting algorithm: ***Adaboost***
  - *weights of patterns created in previous iteration are used for training of a classifier in next iteration*
  - *some classifiers may be better at classifying difficult patterns than others*
  - *a series of classifiers that complement each other is being built.*
- *Error rate of model M_i*

$$error(M_i) = \sum_{j}^{d} w_j \cdot err(X_j), \quad err(X_j) = \begin{cases} 1, & \text{if } X_j \text{ misclassified} \\ 0, & \text{otherwise} \end{cases}$$

where w_j is pattern weight.

  - *If $error(M_i)>0.5$, discard the classifier M_i as poor. Create a new M_i trained on new set D_i.*
  - *Weights of patterns are updated.*

Accuracy Improvement

Boosting

- E.g. of boosting algorithm: ***Adaboost***
- *Error rate of model M_i*

$$error(M_i) = \sum_{j}^{d} w_j \cdot err(X_j), \quad err(X_j) = \begin{cases} 1, & \text{if } X_j \text{ misclassified} \\ 0, & \text{otherwise} \end{cases}$$

where w_i is pattern weight.

  - *if $error(M_i)>0.5$, discard the classifier M_i as poor. Create a new M_i trained on new set D_i.*
  - *Weights of patterns updated*
    - *if the pattern in iteration i is correctly classified, its weight is multiplied by*

$$\frac{error(M_i)}{1 - error(M_i)}$$

    - *normalize weights*
      - *norm. weight = (weight * Sum_of_old_weights) / Sum_of_new_weights*
    - *weights of misclassified increased, and weights of correctly classified patterns are decreased*

Accuracy Improvement

Boosting

- E.g. of boosting algorithm: ***Adaboost***

- *Ensemble of classifiers used to predict a class label* X_i
    - *Weight assigned to each classifier's vote*
    - *The lower the classifier's error rate is, the higher its voting weight is*

    - *Weight of classifier's* M_i *vote is:*

$$\log \frac{1 - error(M_i)}{error(M_i)}$$

    - *The class with the highest sum of weights of classifiers wins – that is the class prediction for pattern* X_i *.*

© M. Manic, CMSC 409: Artificial Intelligence, F23