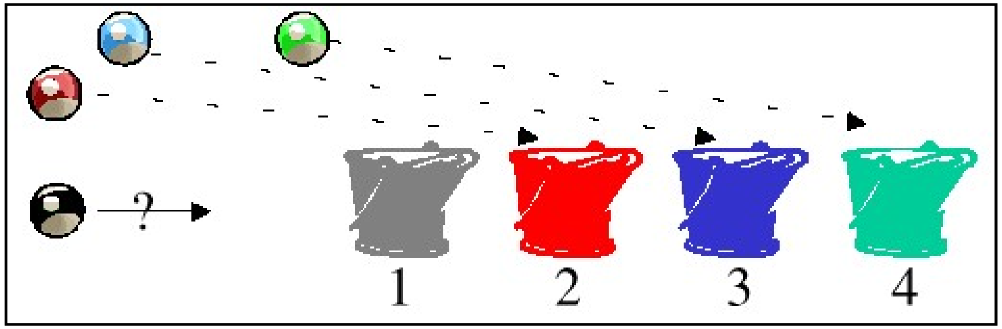
Classification and Prediction: Basic Concepts
CSE450: Data Mining
Summer 2018

SAH @ DIU

What Is Classification?

- **The goal of data classification is to organize and categorize data in distinct classes**
  - A model is first created based on the data distribution
  - The model is then used to classify new data
  - Given the model, a class can be predicted for new data
- **Classification = prediction for discrete and nominal values (e.g., class/category labels)**
  - Also called “Categorization”



Prediction, Clustering, Classification

- **What is Prediction/Estimation?**
  - The goal of prediction is to forecast or deduce the value of an attribute based on values of other attributes
  - A model is first created based on the data distribution
  - The model is then used to predict future or unknown values
  - Most common approach: regression analysis

- **Supervised vs. Unsupervised Classification**
  - Supervised Classification = Classification
    - We know the class labels and the number of classes
  - Unsupervised Classification = Clustering
    - We do not know the class labels and may not know the number of classes

Classification Task

- **Given:**
  - A description of an instance, $x \in X$, where X is the *instance language* or *instance* or *feature space*.
    - Typically, x is a row in a table with the instance/feature space described in terms of features or attributes.
  - A fixed set of class or category labels: $C=\{c_1, c_2, \ldots c_n\}$

- **Classification task is to determine:**
  - The class/category of x: $c(x) \in C$, where $c(x)$ is a function whose domain is X and whose range is C.

Learning for Classification

- **A training example is an instance $x \in X$, paired with its correct class label $c(x)$: <x, $c(x)$> for an unknown classification function, c.**

- **Given a set of training examples, D**
  - Find a hypothesized classification function, $h(x)$, such that: $h(x) = c(x)$, for all training instances (i.e., for all <x, $c(x)$> in D). This is called consistency.

Example of Classification Learning

- **Instance language: <size, color, shape>**
  - size $\in$ {small, medium, large}
  - color $\in$ {red, blue, green}
  - shape $\in$ {square, circle, triangle}
- ***C* = {positive, negative}**
- ***D*:**

Example	Size	Color	Shape	Category
1	small	red	circle	positive
2	large	red	circle	positive
3	small	red	triangle	negative
4	large	blue	circle	negative

- **Hypotheses? circle → positive? red → positive?**

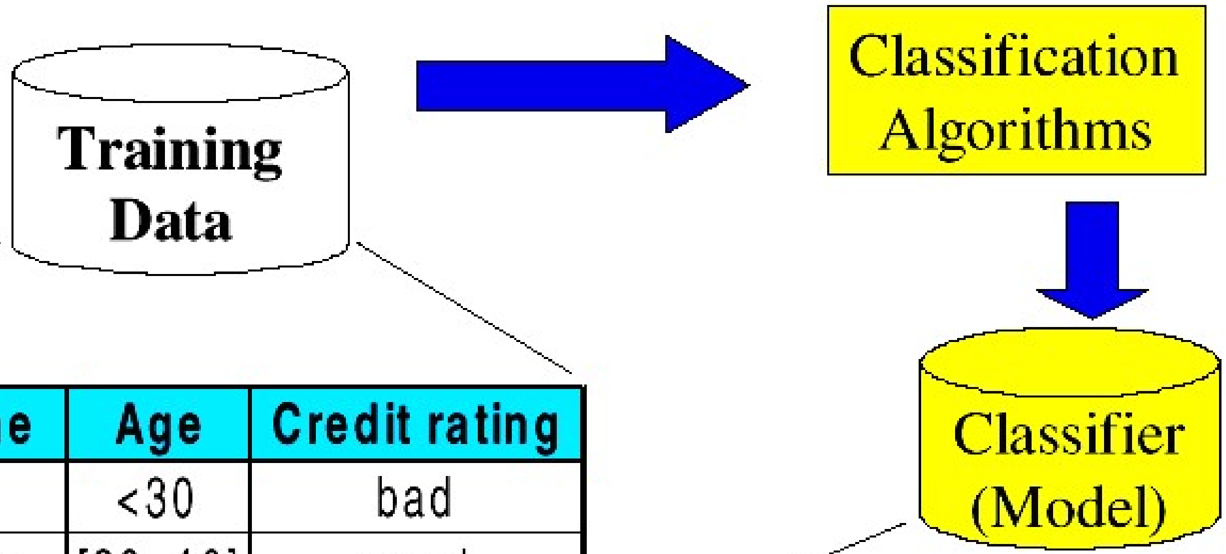
General Learning Issues
(All Predictive Modeling Tasks)

- **Many hypotheses can be consistent with the training data**
- **Bias: Any criteria other than consistency with the training data that is used to select a hypothesis**
- **Classification accuracy (% of instances classified correctly)**
  - Measured on independent test data
- **Efficiency Issues:**
  - Training time (efficiency of training algorithm)
  - Testing time (efficiency of subsequent classification)
- **Generalization**
  - Hypotheses must generalize to correctly classify instances not in training data
  - Simply memorizing training examples is a consistent hypothesis that does not generalize
  - Occam's razor: Finding a simple hypothesis helps ensure generalization
    - Simplest models tend to be the best models
    - The KISS principle

Classification: 3 Step Process

- **1. Model construction (Learning):**
  - Each record (instance, example) is assumed to belong to a predefined class, as determined by one of the attributes
    - This attribute is call the target attribute
    - The values of the target attribute are the class labels
  - The set of all instances used for learning the model is called training set
  - The model may be represented in many forms: decision trees, probabilities, neural networks, ….
- **2. Model Evaluation (Accuracy):**
  - Estimate accuracy rate of the model based on a test set
  - The known labels of test instances are compared with the predicts class from model
  - Test set is independent of training set otherwise over-fitting will occur
- **3. Model Use (Classification):**
  - The model is used to classify unseen instances (i.e., to predict the class labels for new unclassified instances)
  - Predict the value of an actual attribute

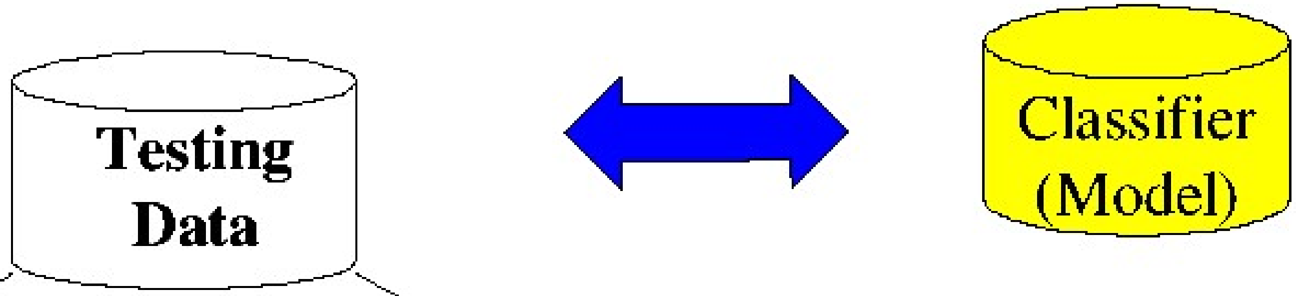
Model Construction



Name	Income	Age	Credit rating
Bruce	Low	<30	bad
Dave	Medium	[30..40]	good
William	High	<30	good
Marie	Medium	>40	good
Anne	Low	[30..40]	good
Chris	Medium	<30	bad

IF Income = ‘High’
OR Age > 30
THEN CreditRating = ‘Good’

Model Evaluation

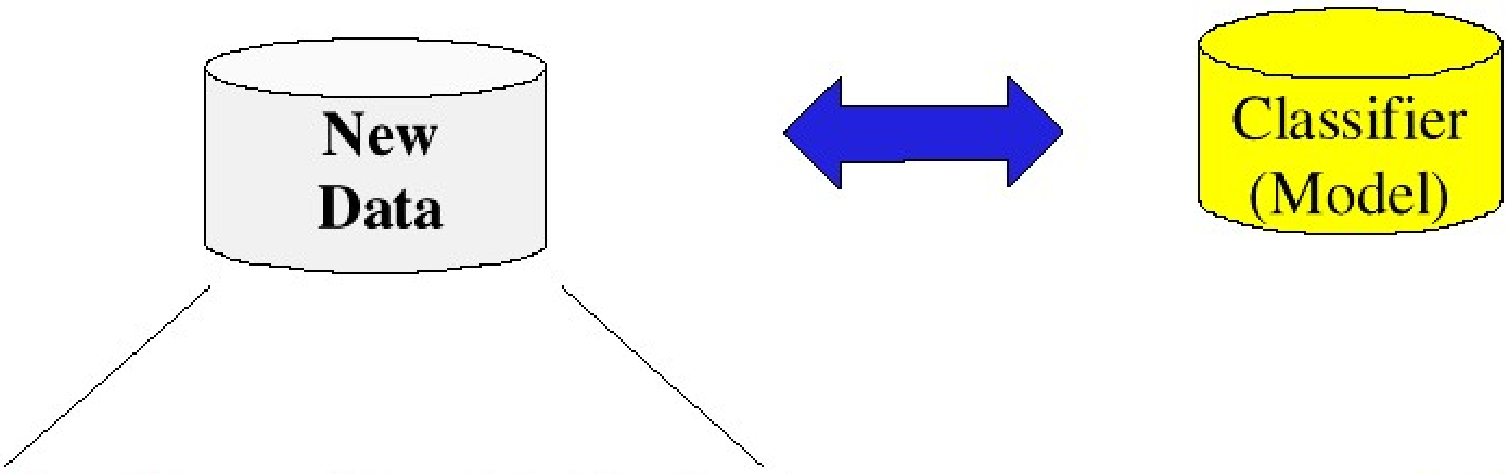


Name	Income	Age	Credit rating
Tom	Medium	<30	bad
Jane	High	<30	bad
Wei	High	>40	good
Hua	Medium	[30..40]	good

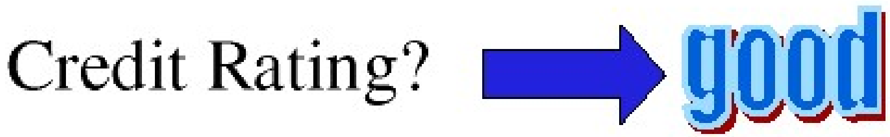
How accurate is the model?

IF Income = 'High'
OR Age > 30
THEN CreditRating = 'Good'

Model Use: Classification



Name	Income	Age	Credit rating
Paul	High	[30..40]	?



Classification Methods

- Decision Tree Induction
- Bayesian Classification
- K-Nearest Neighbor
- Neural Networks
- Support Vector Machines
- Association-Based Classification
- Genetic Algorithms
- Many More ….

- Also Ensemble Methods

Evaluating Models

- **To train and evaluate models, data are often divided into three sets: the training set, the test set, and the evaluation set**
- **Training Set**
  - is used to build the initial model
  - may need to "enrich the data" to get enough of the special cases
- **Test Set**
  - is used to adjust the initial model
  - models can be tweaked to be less idiosyncrasies to the training data and can be adapted for a more general model
  - idea is to prevent "over-training" (i.e., finding patterns where none exist).
- **Evaluation Set**
  - is used to evaluate the model performance

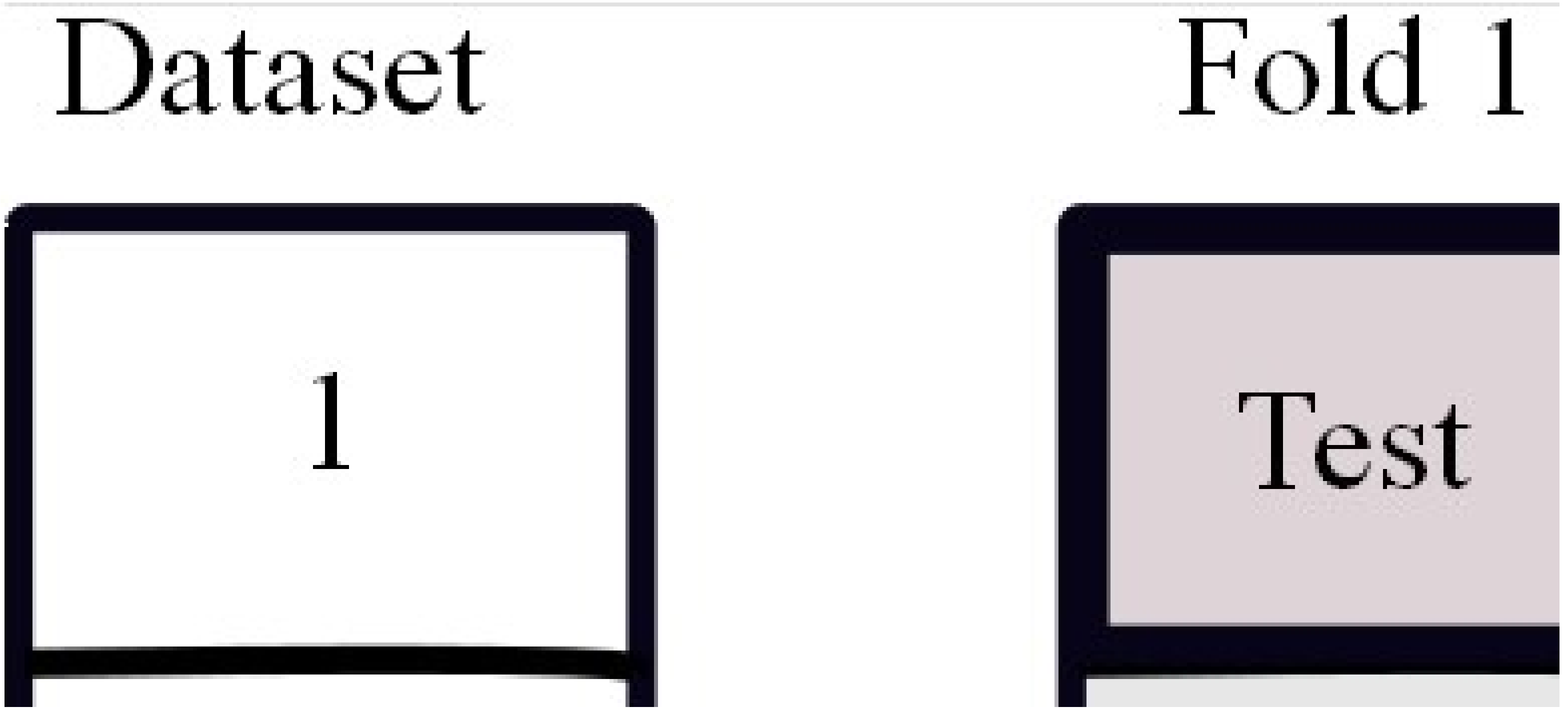
Test and Evaluation Sets

- **Reading too much into the training set (overfitting)**
  - common problem with most data mining algorithms
  - resulting model works well on the training set but performs poorly on unseen data
  - test set can be used to "tweak" the initial model, and to remove unnecessary inputs or features
- **Evaluation Set is used for final performance evaluation**
- **Insufficient data to divide into three disjoint sets?**
  - In such cases, validation techniques can play a major role
    - Cross Validation
    - Bootstrap Validation

Cross Validation

- **Cross validation is a heuristic that works as follows**
  - randomly divide the data into *n* ***folds***, each with approximately the same number of records
  - create *n* models using the same algorithms and training parameters; each model is trained with *n*-1 folds of the data and tested on the remaining fold
  - can be used to find the best algorithm and its optimal training parameter
- **Steps in Cross Validation**
  - 1. Divide the available data into a training set and an evaluation set
  - 2. Split the training data into *n* folds
  - 3. Select an algorithm and training parameters
  - 4. Train and test *n* models using the *n* train-test splits
  - 5. Repeat step 2 to 4 using different algorithms / parameters and compare model accuracies
  - 6. Select the best model
  - 7. Use all the training data to train the model
  - 8. Assess the final model using the evaluation set

Example – 5 Fold Cross Validation



Bootstrap Validation

- **Based on the statistical procedure of sampling with replacement**
  - data set of n instances is sampled n times (with replacement) to give another data set of n instances
  - since some elements will be repeated, there will be elements in the original data set that are not picked
  - these remaining instances are used as the test set
- **How many instances in the test set?**
  - Probability of not getting picked in one sampling = $1 - 1/n$
  - Pr(not getting picked in n samples) = $(1 - 1/n)^n = e^{-1} = 0.368$
  - so, for large data set, test set will contain about 36.8% of instances
  - to compensate for smaller training sample (63.2%), test set error rate is combined with the re-substitution error in training set:

$$\boxed{e = (0.632 * e_{\text{test instance}}) + (0.368 * e_{\text{training instance}})}$$

- **Bootstrap validation increases variance that can occur in each fold**

Measuring Effectiveness of Classification Models

- **When the output field is nominal (e.g., in two-class prediction), we use a confusion matrix to evaluate the resulting model**

- **Example**

Predicted Class

Actual Class

	T	F	Total
T	18	2	20
F	3	15	18
Total	21	17	38

  - Overall correct classification rate = (18 + 15) / 38 = 87%
  - Given T, correct classification rate = 18 / 20 = 90%
  - Given F, correct classification rate = 15 / 18 = 83%

Confusion Matrix & Accuracy Metrics

Actual class\Predicted class	C_1	$\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

- **Classifier Accuracy, or recognition rate: percentage of test set instances that are correctly classified**
  - Accuracy = (TP + TN)/All
  - Error rate: 1 – accuracy, or Error rate = (FP + FN)/All
- **Class Imbalance Problem: One class may be rare, e.g. fraud, or HIV-positive**
  - Sensitivity: True Positive recognition rate = TP/P
  - Specificity: True Negative recognition rate = TN/N

Other Classifier Evaluation Metrics

- **Precision**
  - % of instances that the classifier predicted as positive that are actually positive

$$precision = \frac{TP}{TP + FP}$$

- **Recall**
  - % of positive instances that the classifier predicted correctly as positive
  - a.k.a "Completeness"

$$recall = \frac{TP}{TP + FN}$$

- **Perfect score for both is 1.0, but there is often a trade-off between Precision and Recall**

- ***F* measure (*F_1* or *F*-score)**
  - harmonic mean of precision and recall

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

What Is Prediction/Estimation?

- **(Numerical) prediction is similar to classification**
  - construct a model
  - use model to predict continuous or ordered value for a given input
- **Prediction is different from classification**
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions
- **Major method for prediction: regression**
  - model the relationship between one or more *independent* or **predictor** variables and a *dependent* or **response** variable
- **Regression analysis**
  - Linear and multiple regression
  - Non-linear regression
  - Other regression methods: generalized linear model, Poisson regression, log-linear models, regression trees

Linear Regression

- **<u>Linear regression</u>: involves a response variable y and a single predictor variable x → $\mathbf{y = w_0 + w_1 x}$**
  - w_0 (y-intercept) and w_1 (slope) are regression coefficients
- **<u>Method of least squares</u>: estimates the best-fitting straight line**

$$\boxed{w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2}}$$

$$\boxed{w_0 = \bar{y} - w_1 \bar{x}}$$

- **<u>Multiple linear regression</u>: involves more than one predictor variable**
  - Training data is of the form $(\mathbf{X_1}, y_1), (\mathbf{X_2}, y_2), \ldots, (\mathbf{X_{|D|}}, y_{|D|})$
  - Ex. For 2-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
  - Solvable by extension of least square method
  - Many nonlinear functions can be transformed into the above

Nonlinear Regression

- **Some nonlinear models can be modeled by a polynomial function**
- **A polynomial regression model can be transformed into linear regression model. For example,**

  $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

  is convertible to linear with new variables: $x_2 = x^2$, $x_3 = x^3$

  $y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$
- **Other functions, such as power function, can also be transformed to linear model**
- **Some models are intractable nonlinear (e.g., sum of exponential terms)**
  - possible to obtain least squares estimates through extensive computation on more complex functions

Other Regression-Based Models

- **Generalized linear models**
  - Foundation on which linear regression can be applied to modeling categorical response variables
  - Variance of y is a function of the mean value of y, not a constant
  - Logistic regression: models the probability of some event occurring as a linear function of a set of predictor variables
  - Poisson regression: models the data that exhibit a Poisson distribution
- **Log-linear models (for categorical data)**
  - Approximate discrete multidimensional prob. distributions
  - Also useful for data compression and smoothing
- **Regression trees and model trees**
  - Trees to predict continuous values rather than class labels

Regression Trees and Model Trees

- **Regression tree: proposed in CART system (Breiman et al. 1984)**
  - CART: Classification And Regression Trees
  - Each leaf stores a *continuous-valued prediction*
  - It is the *average value of the predicted attribute* for the training instances that reach the leaf
- **Model tree: proposed by Quinlan (1992)**
  - Each leaf holds a regression model—a multivariate linear equation for the predicted attribute
  - A more general case than regression tree
- **Regression and model trees tend to be more accurate than linear regression when instances are not represented well by simple linear models**

Evaluating Numeric Prediction

- **Prediction Accuracy**
  - Difference between predicted scores and the actual results (from evaluation set)
  - Typically the accuracy of the model is measured in terms of variance (i.e., average of the squared differences)

- **Common Metrics** (p_i = predicted target value for test instance i, a_i = actual target value for instance i)
  - **Mean Absolute Error:** Average loss over the test set

$$\boxed{MAE = \frac{\left|(p_1 - a_1) + \ldots + (p_n - a_n)\right|}{n}}$$

  - **Root Mean Squared Error**: compute the standard deviation (i.e., square root of the co-variance between predicted and actual ratings)

$$\boxed{RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}}$$