Introduction to Big Data

Day 6 and 7

Model Application, Feature Engineering, and Evaluation

YORK UNIVERSITÉ UNIVERSITY

Day 6 and 7 - Learning Objectives

During Day you will learn to:

- Identify some common application areas of analytic models
- Describe the purpose of various analytical modeling techniques
- Describe model features
- Select model features
- Describe the purpose of dimension reduction
- Evaluate model performance

Application Areas

Application Areas

Applications of Analytic Models

Common areas of functional applications for analytic models include the capability to:

- Predict
- Forecast
- Classify
- Cluster
- Associate
- Sequence
- Detect
- Diagnose
- Describe
- Prescribe

Application Areas

Predict

- estimate the likelihood that future event or condition will occur
- common form of a prediction is in the form of a probability
- probabilities may be transformed into scores and likelihood categories to support the needs of decision makers

Forecast

- estimate the future value of a continuous variable
- common form of a forecast is to position the future value(s) on a timeline

Classify

- assign an observation, condition or event to a pre-defined class or category
- the observations, conditions or events may be from the past, the present or the future.

Application Areas

Cluster

- determine what clusters or categories best describe groups that have common observed behavior
- the premise is that the algorithm determines what variables define the boundaries of a given cluster
- after the clusters are determined, then conditions or events can be classified into the appropriate cluster
- the algorithm is told how many clusters are desired and then the appropriate categories are determined

Associate

- determine what events, conditions or items are expected to exist together in combination, given a related context or circumstance

Sequence

- determine how events or conditions based on their dependencies exist on a common timeline

Application Areas

Detect

- the capability to identify and report an abnormal event or condition

Diagnose

- the capability to identify and estimate root causes to a event or condition
- root causes are translated into counter measures that can be implemented to reduce the symptoms of a given problem or to prevent the symptoms from re-occurring

Describe

- the capability to quantify and describe the behaviors, conditions, activities and outcomes within the context of the domain being studied.
- descriptions are further summarized based on descriptive statistics to estimate the measures of central tendency and dispersion within a set of observations

Prescribe

- the capability to recommend a course of action based on rules of thumb, simulation methods or optimization techniques

Some tribulations

- A big objection to data mining was that it was looking for so many vague connections that it was sure to find things that were bogus
- The Rhine Paradox: a great example of how not to conduct scientific research.
- David Rhine was a parapsychologist in the 1950’s who hypothesized that some people had Extra-Sensory Perception (ESP).
- He devised an experiment where subjects were asked to guess 10 hidden cards --- red or blue.
- He discovered that almost 1 in 1000 had ESP --- they were able to get all 10 right!

Example(con’t)

- He told these people they had ESP and called them in for another test of the same type.
- Alas, he discovered that almost all of them had lost their ESP.
- What did he conclude?

You shouldn’t tell people that they have ESP: it causes them to lose it

Example (con't)

- What has really happened:

There are 1024 combinations of red and blue combinations of red and blue of length 10.

Thus with probability 0.98 at least one person will guess the sequence of red blue correctly

Machine Learning is not Data Mining

- Machine Learning design systems that can learn in the process of processing data
- Checkers program designed by one of the scientist eventually learned to play better than the program designer
- Data Mining incorporates the Machine learning methods but also benefits from the methods of other disciplines such as database and statistic

What is Data Mining

- Data Mining major task is to find all and only interesting patterns in a set of data sources
- Find all interesting patterns means – Completeness
  - Can it be done
  - Heuristic vs Exhaustive search
- Find only interesting patterns – Consistency
  - Is it possible
  - Approaches: Generate all patterns and filter out uninteresting patterns; generate only patterns that are interesting

Potential Applications

- Database analysis and decision support
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management
- Other Applications
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

Market Analysis and Management (1)

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

Market Analysis and Management (2)

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
  - identifying the best products for different customers
  - use prediction to find what factors will attract new customers
- Provides summary information
  - various multidimensional summary reports
  - statistical summary information (data central tendency and variation)

Corporate Analysis and Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning:
  - summarize and compare the resources and spending
- Competition:
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection and Management (1)

- Applications
  - widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Examples
  - auto insurance: detect a group of people who stage accidents to collect on insurance
  - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - medical insurance: detect professional patients and ring of doctors and ring of references

Fraud Detection and Management (2)

- Detecting inappropriate medical treatment
- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
- Retail
  - Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Application Areas – Video Content

The following videos provide examples of analytic model applications

Please view these videos to gain a richer perspective

Video from Ted Talks describing how humans play a critical role that complements analytical models
https://www.ted.com/talks/tricia_wang_the_human_insights_missing_from_big_data

Video describing some analytics applications from the energy sector
http://www.mastersindatascience.org/industry/energy/

Applications of Photograph Interpretation from Ted Talks
https://www.youtube.com/watch?v=40riCqvRoMs

Categories of Analytic Models

Categories of Analytic Models

Categories of Models

Analytic models can be classified according some of their key properties
The following groups are commonly used.

- Empirical vs Mechanistic
  - Empirical models are based on relationships from observed data
  - Mechanistic models are based on relationships based on theory and existing knowledge
- Deterministic vs Stochastic
  - Deterministic models do not consider randomness or uncertainty
  - Stochastic models do consider elements of randomness and uncertainty
- Static vs Dynamic
  - Static models do not consider the impact of time and the output is generally considered to be at a “point in time”
  - Dynamic models do consider the impact of time on the output variables.

Categories of Analytic Models

Machine learning, data mining and data science focus on developing empirical models, based on observed data.

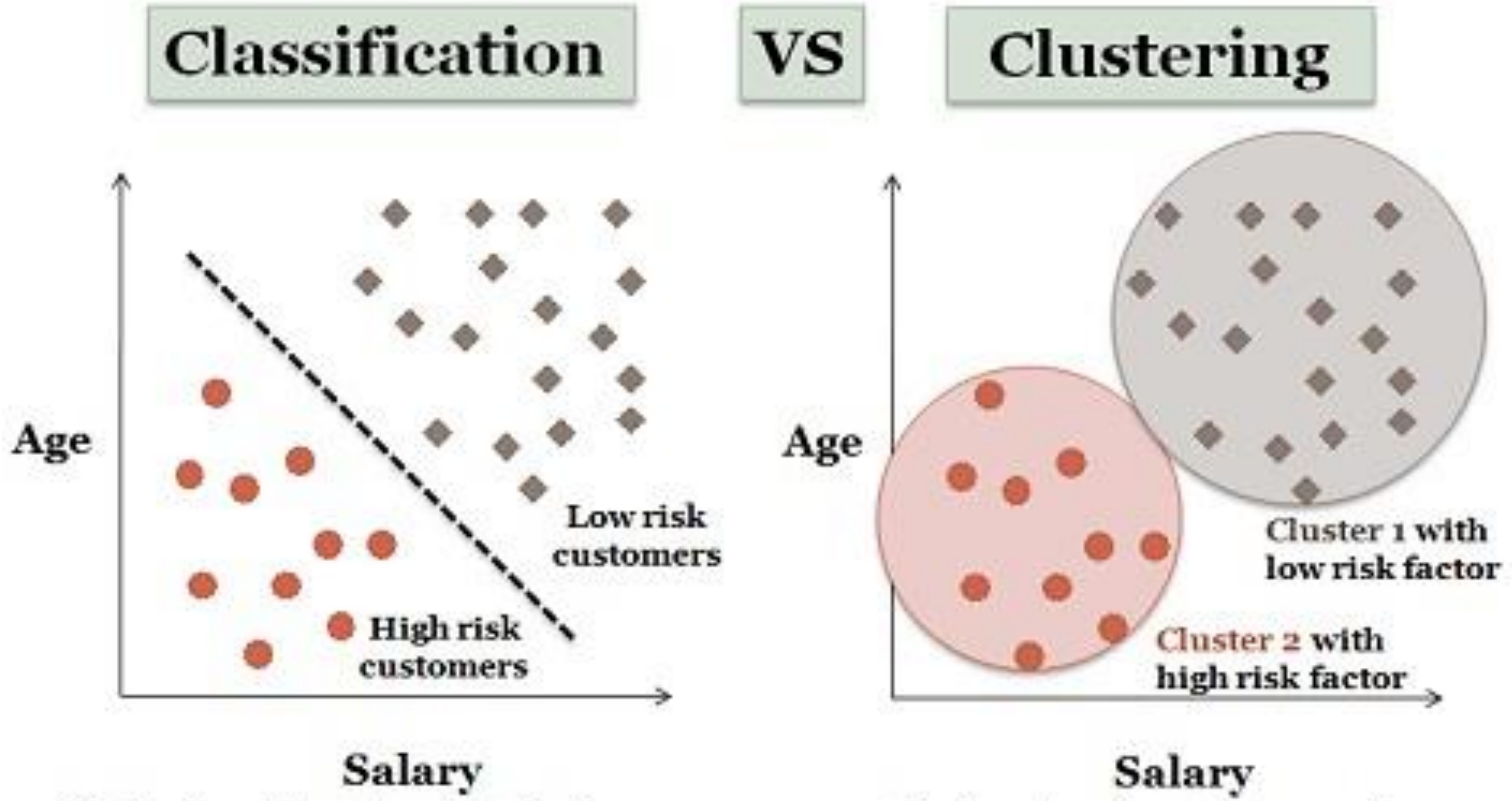
Examples of Empirical Modeling Techniques

Supervised Learning Models

- Supervised learning models are generated based on defined input variables and a defined output variable, sometimes called a label. The relationships are discovered that best fits a model between the input variables and the relate output variable.
- Examples of Supervised Learning Techniques
    - Regression
    - Decision Trees
    - Classification

Unsupervised Learning Models

- Unsupervised learning models are generated without a known or defined output variable of label. The algorithms define related variables or features the describe how in combination they provide a group having some common observed behavior.
- Examples of Unsupervised Learning Techniques
    - Clustering
    - Association
    - Dimensionality Reduction



Risk classification for the loan payees on the basis of customer salary

Classification

- Given a set of classes, distribute the data into a given set of classes so that a newly arrived data will be with the high probability will fall into one of the classes.
- Credit Card example: 4 classes: authorize; request more info; do not authorize; contact police
- Data is a set of credit card applications that contain Name, age, credit score, address, income, own or rent primary residence, etc.

Regression

- Regression is a process of mapping a given data to some function. Regression may be linear (mapping into a linear function the set of given data or non-linear function.
- For example, one may map saving amount to a person age as follows:

  samt = a*age+b, where constant *a* and *b* are determined by existing data
- Fitting the rest of the data into a defined function should have the least possible error

Time Series Analysis

- Given data that changes with time to predict the data behavior based on the known data
- Example: predict stock market, predict the stock price of a specific company
- Visualization is an important tool of time series analysis
- There are special operations on time series that facilitate the time series analysis

Prediction

- Differences between Classification and Prediction:
  - Classification deals with an existing data
  - Prediction deals with future events
- Mathematical Models are normally used for prediction: Weather forecast, quake forecast, etc.

Clustering

- Clustering is a process of distributing given data into several sets so that distance between different sets is larger than the distance between elements in the same set
- Difference between Clustering and Classification is that the number of clusters is not known in advance, whereas the number of classes is known in advance.
- Examples

Association Rules and Sequence Discovery

- Association rules discovery relates to uncovering unexpected relationships between data attribute values.
  - For example people who buy coffee may not buy tee, or man who buy diapers also buy beer. However, women who buy diapers do not buy beer
  - “Men buy beer and diapers on Fridays”
    - https://www.theregister.co.uk/2006/08/15/beer_diapers/
- Sequence discovery – an ability to determine sequential patterns in the data

Categories of Analytic Models

Videos about Supervised and Unsupervised Learning Methods

https://www.youtube.com/watch?v=cfj6yaYE86U

https://www.youtube.com/watch?v=Ig1nfPjrETc&t=92s

Introduction to Machine Learning
https://www.youtube.com/watch?v=IpGxLWOIZy4

Comparing Supervised and Unsupervised Learning
https://www.youtube.com/watch?v=qDbpYUbf3e0

Model Features

Model Features

Introducing Model Features

- In machine learning a feature is an individual measurable property or characteristic of a phenomenon being observed
- Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.
- Features are usually numeric, but structural features such as strings and graphs are also used.
- The concept of "feature" is related to that of explanatory variable used in statistical techniques such as linear regression.
- The initial set of raw features can be redundant and too large to be managed.
- A preliminary step in machine learning consists of selecting a subset of features or constructing a new and reduced set of features to facilitate learning, and to improve generalization
- Extracting or selecting features is a combination of art and science; developing systems to do so is known as feature engineering.

- Source Wikipedia

Model Features - Examples

The following videos provide a conceptual introduction to the concept of model features. Please view the videos to gain this perspective

Introduction to Model Features
https://www.youtube.com/watch?v=yVICmUvy060

What makes a good feature?
https://www.youtube.com/watch?v=N9fDIAflCMY

Feature Engineering

Feature Engineering

The art and science of selecting and/or generating the columns in a data table for a machine learning model.

- Not all columns are useful in their raw form.
- There are three sub categories of feature engineering, ie. feature selection, dimension reduction and feature generation.

Feature Selection

- Process of ranking the attributes by their value to predictive ability of a model.
- Algorithms such as decision trees automatically rank the attributes in the data set.
- Regression type models usually employ methods such as forward selection or backward elimination to select the final set of attributes for a model.

Dimension Reduction

- Sometimes called feature extraction.
- A classic example of dimension reduction is principle component analysis or PCA
- PCA allows us to combine existing attributes into a new data frame consisting of a much reduced number of attributes by utilizing the variance in the data.
- The attributes which "explain" the highest amount of variance in the data form the first few principal components and we can ignore the rest of the attributes if data dimensionality is a problem from a computational standpoint.
- PCA results in a data table whose attributes do not look anything like the attributes of the raw dataset.

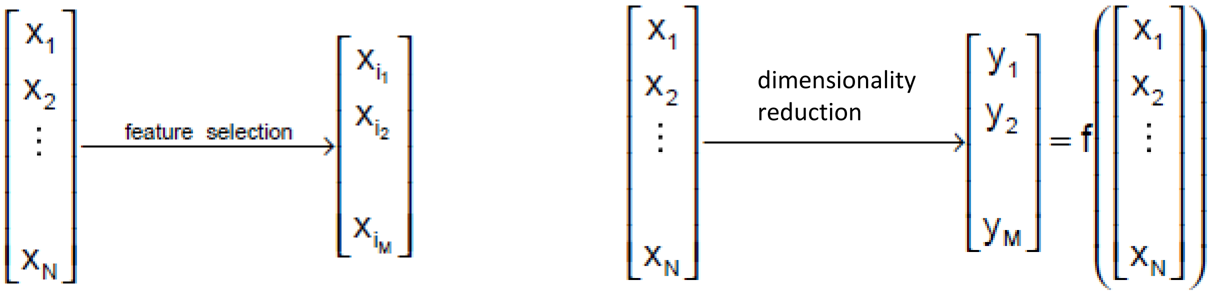
Feature Generation or Feature Construction

- Process of manually constructing new attributes from raw data.
- Involves combining or splitting existing raw attributes into new ones which have a higher predictive power.
- Examples:
  - A date stamp may be used to generate 2 new attributes such as AM and PM which may be useful in discriminating whether day or night has a higher propensity to influence the response variable.
  - Convert noisy numerical attributes into simpler nominal attributes, by calculating the mean value and determining if a given row is above or below that mean value.
  - Generate a new attribute such as number of claims a member has filed for in a given time period, by combining date attribute and a nominal attribute such as claim_filed (Y/N)

- Source Wikepedia

Feature Selection

- Given a set of **n** features, the goal of **feature selection** is to select a subset of **d** features (**d** < **n**) in order to minimize the classification error.



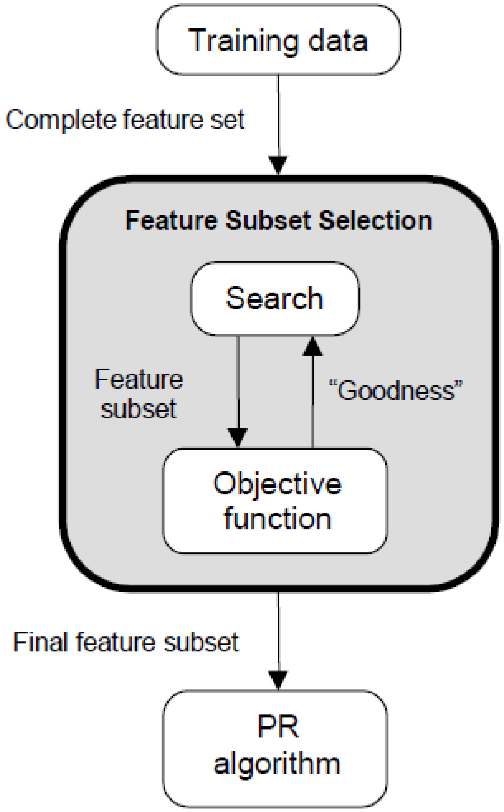
- Fundamentally different from dimensionality reduction (e.g., PCA or LDA) based on feature combinations (i.e., **feature extraction**).

Feature Selection vs Dimensionality Reduction

- Dimensionality Reduction
  - When classifying novel patterns, all features need to be computed.
  - The measurement units (length, weight, etc.) of the features are lost.
- Feature Selection
  - When classifying novel patterns, only a small number of features need to be computed (i.e., faster classification).
  - The measurement units (length, weight, etc.) of the features are preserved.

Feature Selection Steps

- Feature selection is an **optimization** problem.
  - Step 1: Search the space of possible feature subsets.
  - Step 2: Pick the subset that is optimal or near-optimal with respect to some objective function.



Feature Engineering

The following videos provide Python tutorials on basic ideas related to data preparation and model feature engineering

Please view these videos to learn about feature engineering in Python

Using Jupyter Notebook to show examples of Features and Feature Engineering
https://www.youtube.com/watch?v=V0u6bxQOUJ8

Data Agnostic Approach to Feature Engineering without Domain Expertise
https://www.youtube.com/watch?v=bL4b1sGnILU&t=5s

Dimension Reduction

Dimension Reduction

Dimension Reduction

- every potential feature that can be used to predict a dependent variable is defined as a “dimension”
- a series of techniques in machine learning and statistics to reduce the number of independent variables or features to consider
- works to minimize the loss of information contained in the candidate set of features while creating a simpler model
- involves feature selection and feature extraction
- simplifies the modeling effort and the models produced

Dimension Reduction

The mathematical details of dimension reduction are beyond the scope of this course

The primary technique used for dimension reduction is called Principal Components Analysis (PCA)

It is used to reduce the number of initial features considered for the model into a smaller set

A component is a weighted sum of the original features that accounts for a given amount of the variation in the dependent variable.

The “principal” components are the set of components that account for an acceptable amount of the variation within the dependent variable

Why Dimensionality Reduction?

- It is so easy and convenient to collect data
  - An experiment
- Data is not collected only for data mining
- Data accumulates in an unprecedented speed
- Data preprocessing is an important part for *effective* machine learning and data mining
- Dimensionality reduction is an effective approach to downsizing data

Why Dimensionality Reduction?

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Query accuracy and efficiency degrade rapidly as the dimension increases.

- The intrinsic dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

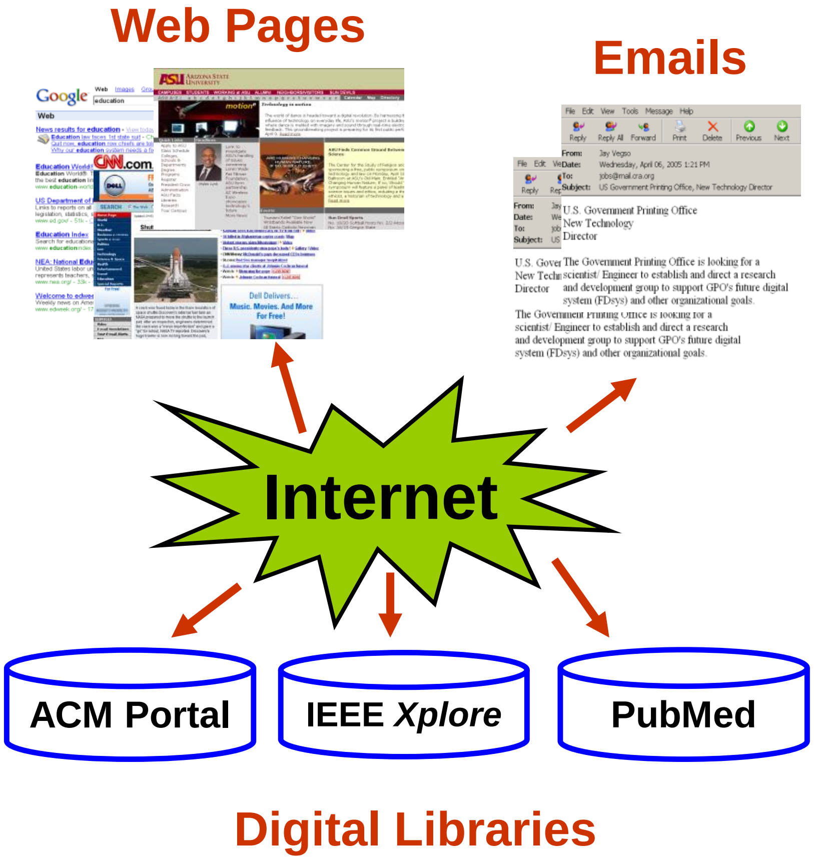
Why Dimensionality Reduction?

- Visualization: projection of high-dimensional data onto 2D or 3D.
- Data compression: efficient storage and retrieval.
- Noise removal: positive effect on query accuracy.

Application of Dimensionality Reduction

- Customer relationship management
- Text mining
- Image retrieval
- Microarray data analysis
- Protein classification
- Face recognition
- Handwritten digit recognition
- Intrusion detection

Document Classification



Terms

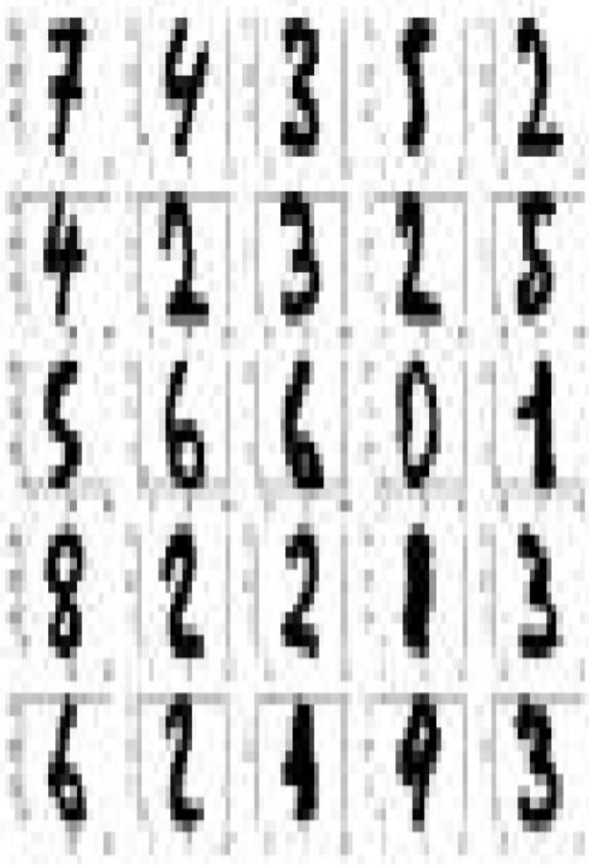
Documents

	T_1	T_2	....	T_N	C
D_1	12	0	....	6	Sports
D_2	3	10	....	28	Travel
⋮	⋮			⋮	⋮
D_M	0	11	....	16	Jobs

- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply dimensionality reduction

Other Types of High-Dimensional Data

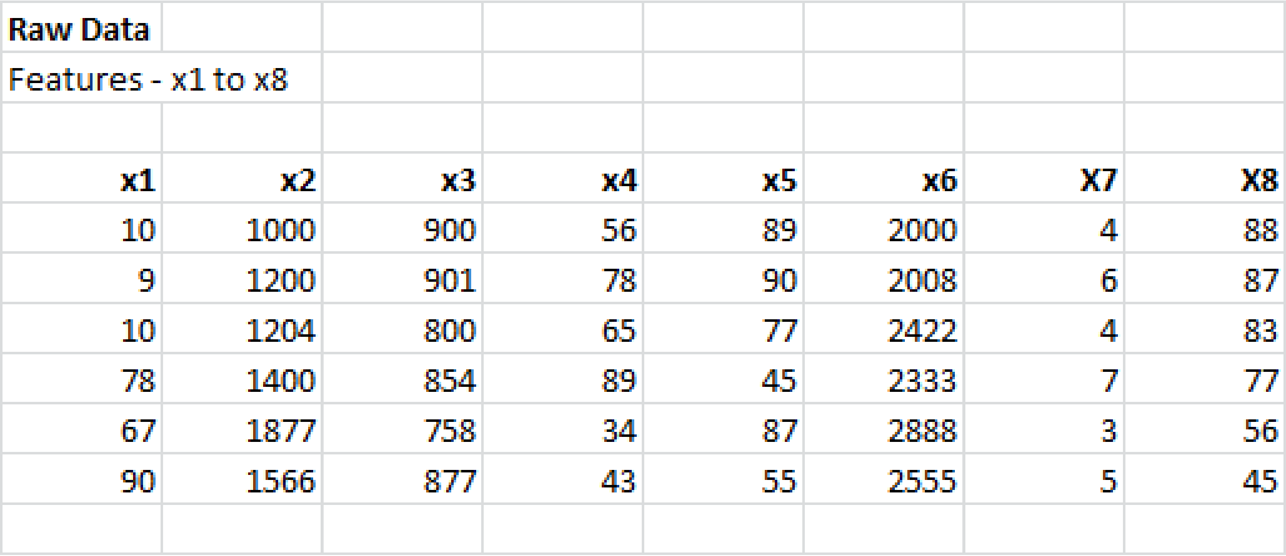
Face images



Handwritten digits

Dimension Reduction – Simple Example

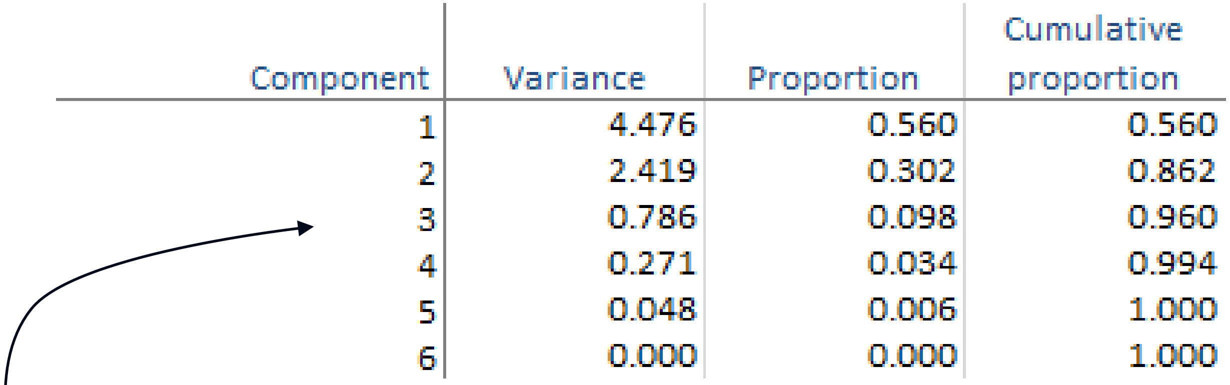
Original Data Set – 8 Potential Features



Raw Data							
Features - x1 to x8							
x1	**x2**	**x3**	**x4**	**x5**	**x6**	**X7**	**X8**
10	1000	900	56	89	2000	4	88
9	1200	901	78	90	2008	6	87
10	1204	800	65	77	2422	4	83
78	1400	854	89	45	2333	7	77
67	1877	758	34	87	2888	3	56
90	1566	877	43	55	2555	5	45

Dimension Reduction – Simple Example

Results of Principal Components Analysis



Component	Variance	Proportion	Cumulative proportion
1	4.476	0.560	0.560
2	2.419	0.302	0.862
3	0.786	0.098	0.960
4	0.271	0.034	0.994
5	0.048	0.006	1.000
6	0.000	0.000	1.000

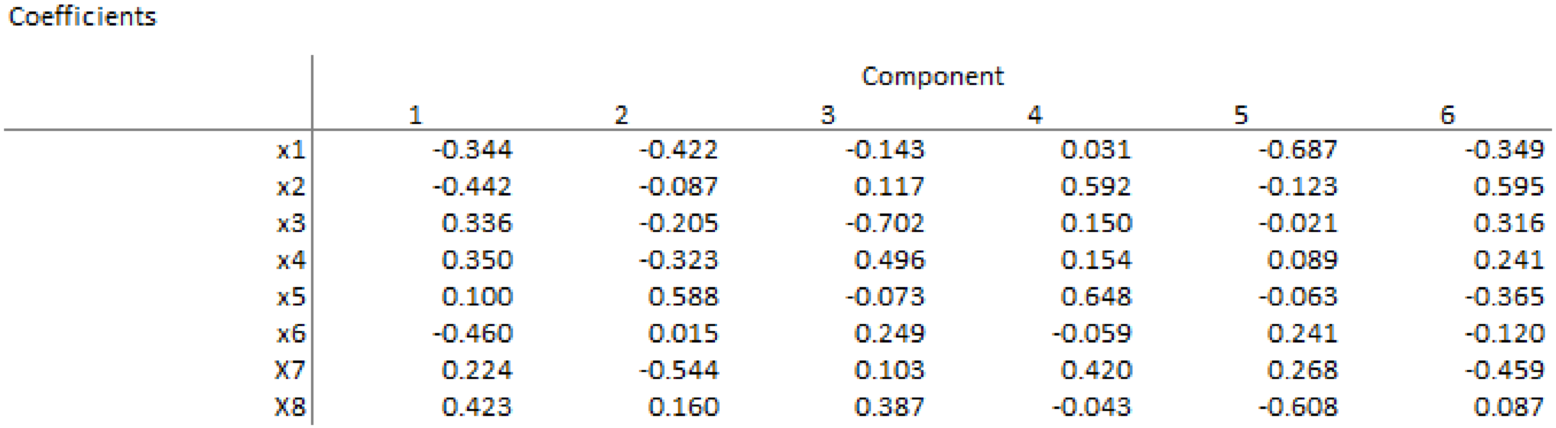
Components 1 to 6

Components 1 to 4 account for 99.4% of the variation

Four components can be
used as features instead of six
This means there are four principal components

Dimension Reduction – Simple Example

Description of Each Component



Coefficients

	Component					
	1	2	3	4	5	6
x1	-0.344	-0.422	-0.143	0.031	-0.687	-0.349
x2	-0.442	-0.087	0.117	0.592	-0.123	0.595
x3	0.336	-0.205	-0.702	0.150	-0.021	0.316
x4	0.350	-0.323	0.496	0.154	0.089	0.241
x5	0.100	0.588	-0.073	0.648	-0.063	-0.365
x6	-0.460	0.015	0.249	-0.059	0.241	-0.120
X7	0.224	-0.544	0.103	0.420	0.268	-0.459
X8	0.423	0.160	0.387	-0.043	-0.608	0.087

Note:
Component 1 = Coeff 1 * X1 + Coeff 2 * X2 + Coeff 3 * X3…… + Coeff n * Xn

Using transformations, selected components can be used as model features

Based on 4 Principal Components, a simpler model emerges

Dimension Reduction

Introduction to the need for Dimension Reduction
https://www.youtube.com/watch?v=9fVSJVp11xc&index=1&list=PLnnr1O8OWc6aVexn2BY0qjklobY6TUEIy

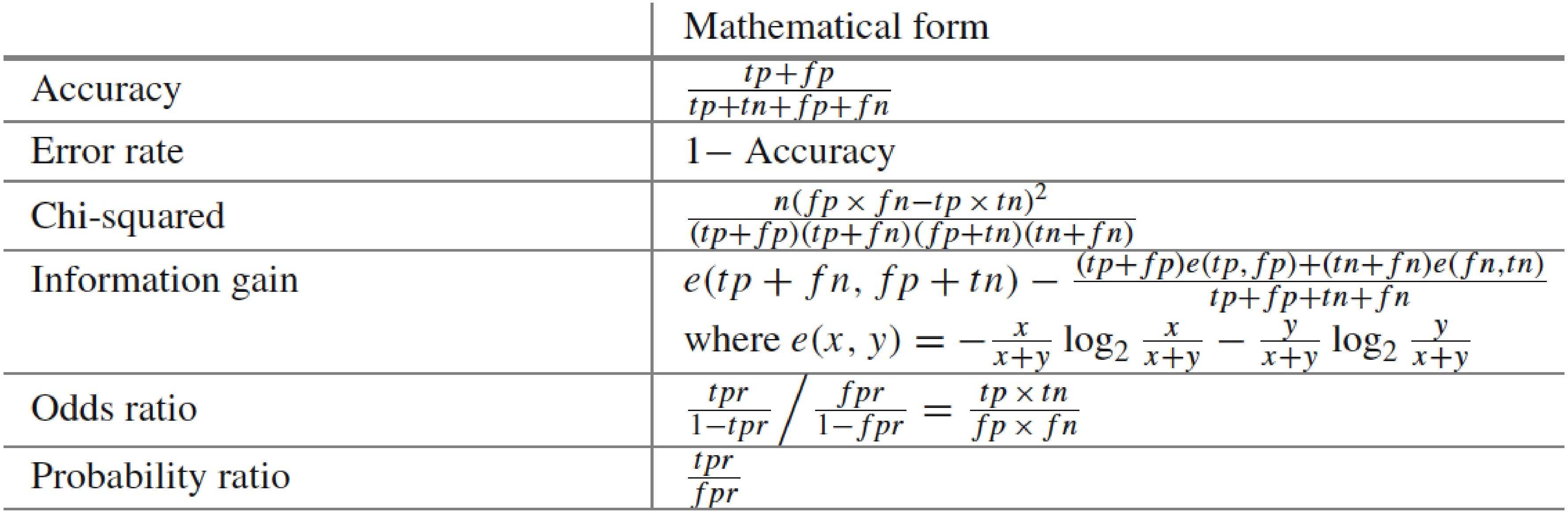
How visualization benefits from Dimension Reduction
https://www.youtube.com/watch?v=MR8sWzoUUwM&index=2&list=PLnnr1O8OWc6aVexn2BY0qjklobY6TUEIy

Perspectives:
Selection Criteria

- Consistency Measures.
    - They attempt to find a minimum number of features that separate classes as the full set of features can.
    - They aim to achieve **P(C|FullSet) = P(C|SubSet)**.
    - An inconsistency is defined as the case of two examples with the same inputs (same feature values) but with different output feature values (classes in classification).

Perspectives: Selection Criteria

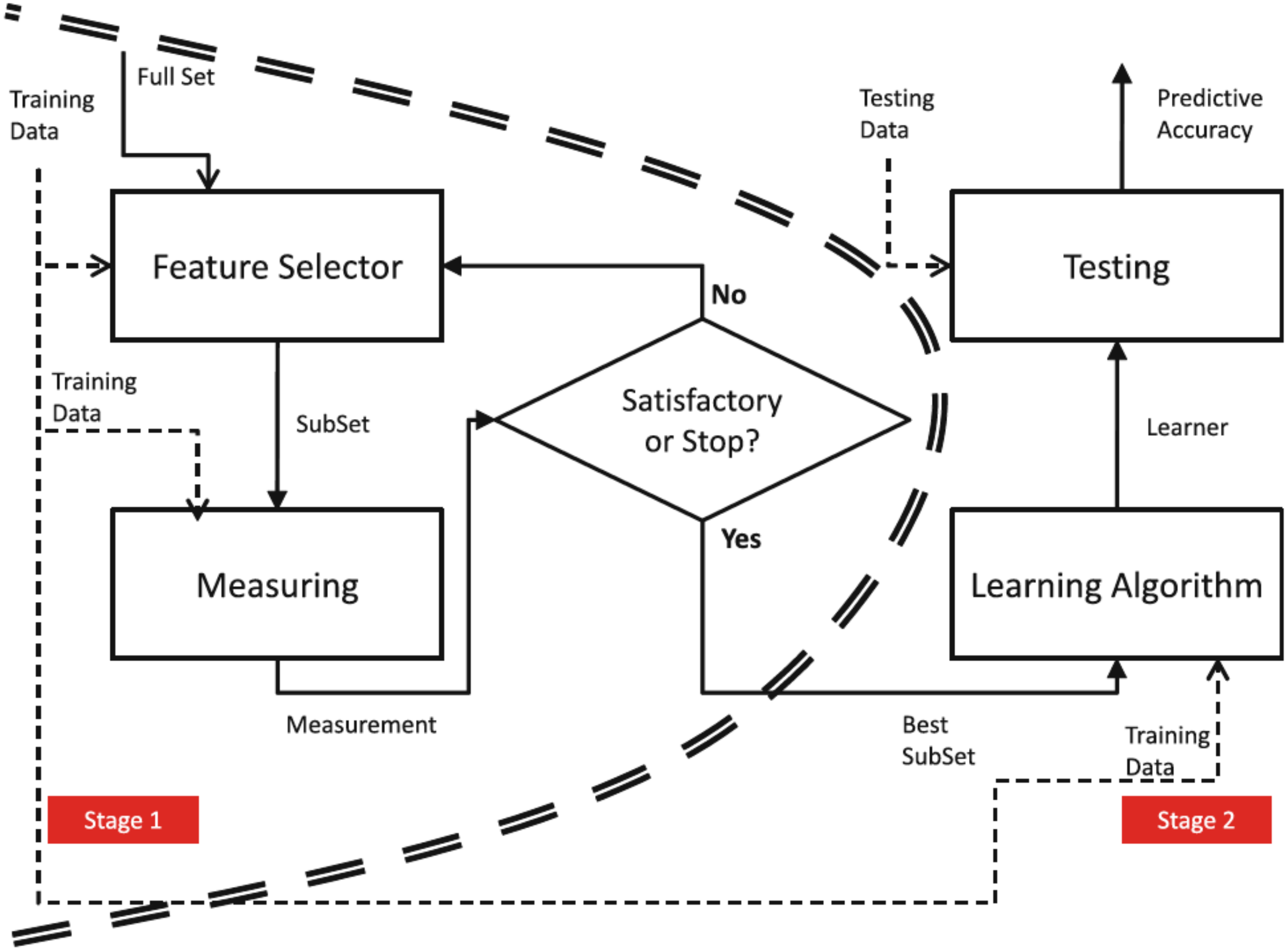
- Accuracy Measures.
  - This form of evaluation relies on the classifier or learner. Among various possible subsets of features, the subset which yields the best predictive accuracy is chosen



	Mathematical form
Accuracy	$\frac{tp+fp}{tp+tn+fp+fn}$
Error rate	$1-$ Accuracy
Chi-squared	$\frac{n(fp \times fn - tp \times tn)^2}{(tp+fp)(tp+fn)(fp+tn)(tn+fn)}$
Information gain	$e(tp+fn, fp+tn) - \frac{(tp+fp)e(tp,fp)+(tn+fn)e(fn,tn)}{tp+fp+tn+fn}$ where $e(x,y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$
Odds ratio	$\frac{tpr}{1-tpr} \Big/ \frac{fpr}{1-fpr} = \frac{tp \times tn}{fp \times fn}$
Probability ratio	$\frac{tpr}{fpr}$

Perspectives

- Filters:

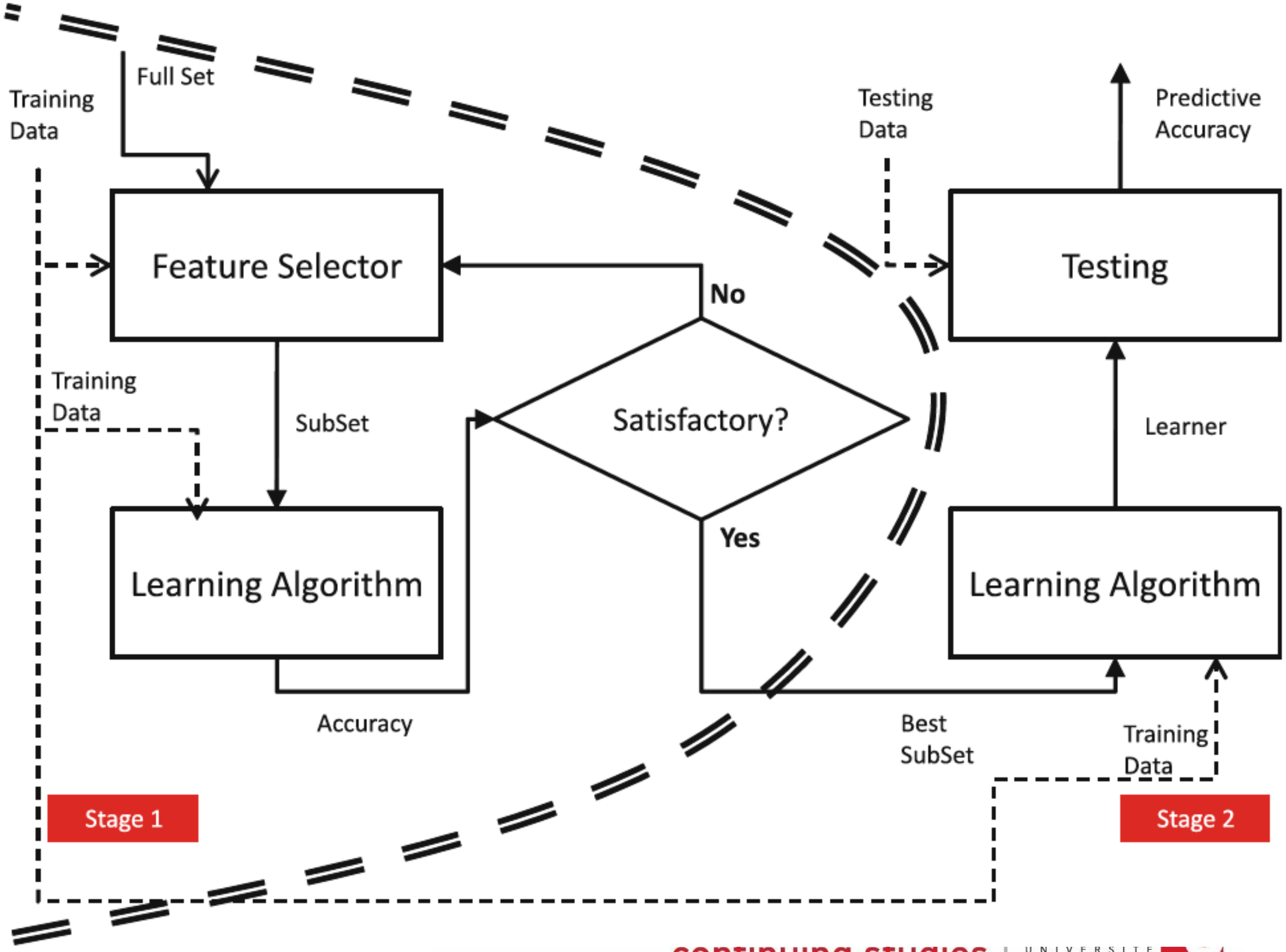


Perspectives

- Filters:
  - measuring uncertainty, distances, dependence or consistency is usually cheaper than measuring the accuracy of a learning process. Thus, filter methods are usually faster.
  - it does not rely on a particular learning bias, in such a way that the selected features can be used to learn different models from different DM techniques.
  - it can handle larger sized data, due to the simplicity and low time complexity of the evaluation measures.

Perspectives

- Wrappers:



Perspectives

- Wrappers:
  - can achieve the purpose of improving the particular learner’s predictive performance.
  - usage of internal statistical validation to control the overfitting, ensembles of learners and hybridizations with heuristic learning like Bayesian classifiers or Decision Tree induction.
  - filter models cannot allow a learning algorithm to fully exploit its bias, whereas wrapper methods do.

Perspectives

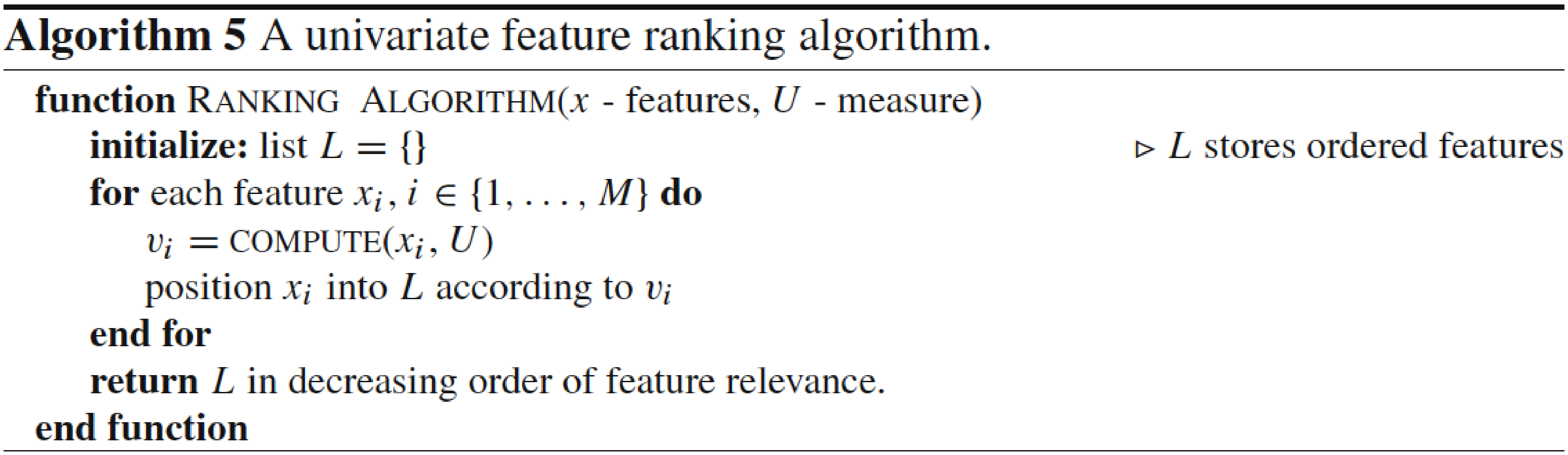
- Embedded FS:
  - similar to the wrapper approach in the sense that the features are specifically selected for a certain learning algorithm, but in this approach, the features are selected during the learning process.
  - they could take advantage of the available data by not requiring to split the training data into a training and validation set; they could achieve a faster solution by avoiding the re-training of a predictor for each feature subset explored.

Aspects:
Output of Feature Selection

- Feature Ranking Techniques:
  - we expect as the output a ranked list of features which are ordered according to evaluation measures.
  - they return the relevance of the features.
  - For performing actual FS, the simplest way is to choose the first *m* features for the task at hand, whenever we know the most appropriate *m* value.

Aspects:
Output of Feature Selection

- Feature Ranking Techniques:



Algorithm 5 A univariate feature ranking algorithm.

```
function RANKING ALGORITHM($x$ - features, $U$ - measure)
    initialize: list $L = \{\}$                          ▷ $L$ stores ordered features
    for each feature $x_i$, $i \in \{1, \ldots, M\}$ do
        $v_i$ = COMPUTE($x_i$, $U$)
        position $x_i$ into $L$ according to $v_i$
    end for
    return $L$ in decreasing order of feature relevance.
end function
```

Aspects: Evaluation

- Goals:
  - **Inferability:** For predictive tasks, considered as an improvement of the prediction of unseen examples with respect to the direct usage of the raw training data.
  - **Interpretability:** Given the incomprehension of raw data by humans, DM is also used for generating more understandable structure representation that can explain the behavior of the data.
  - **Data Reduction:** It is better and simpler to handle data with lower dimensions in terms of efficiency and interpretability.

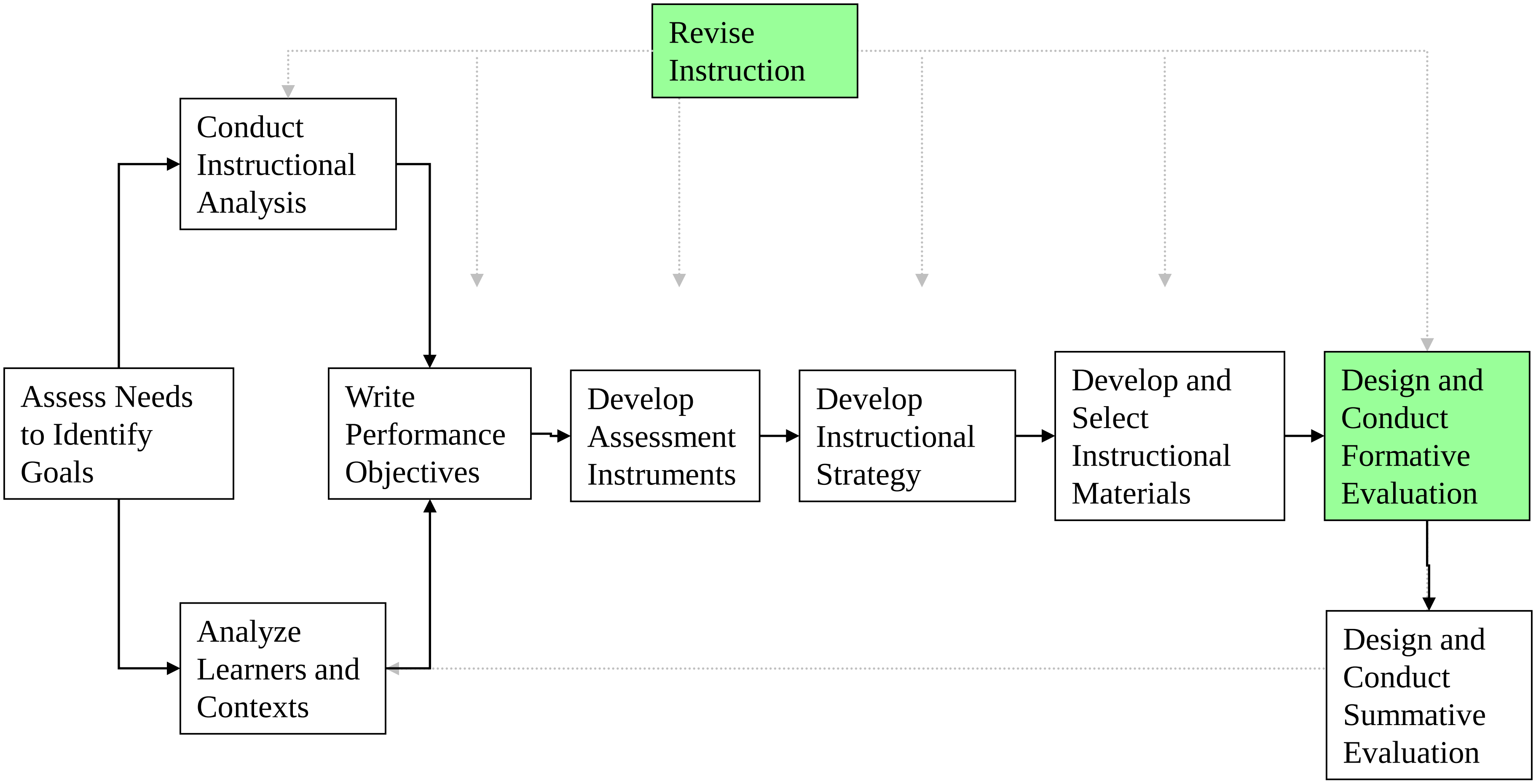
Model Evaluation

Definition

- “Evaluation models either describe what evaluators do or prescribe what they should do” (Alkin and Ellett, 1990, p.15)

When?

- Early and often
- Before it is too late


Revise Instruction
Conduct Instructional Analysis
Assess Needs to Identify Goals
Write Performance Objectives
Develop Assessment Instruments
Develop Instructional Strategy
Develop and Select Instructional Materials
Design and Conduct Formative Evaluation
Analyze Learners and Contexts
Design and Conduct Summative Evaluation

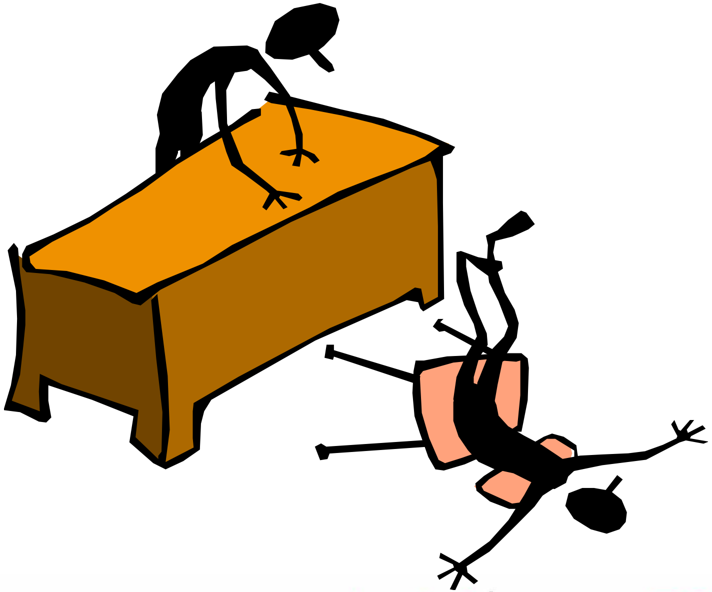
The Dick and Carey Model

What questions to be answered?

- Feasibility: Can it be implemented as it is designed?
- Usability: Can “we” actually use it?
- Appeal: Do “we” like it?
- Effectiveness: Will “we” get what is supposed to get?
  - “we” the stakeholders.

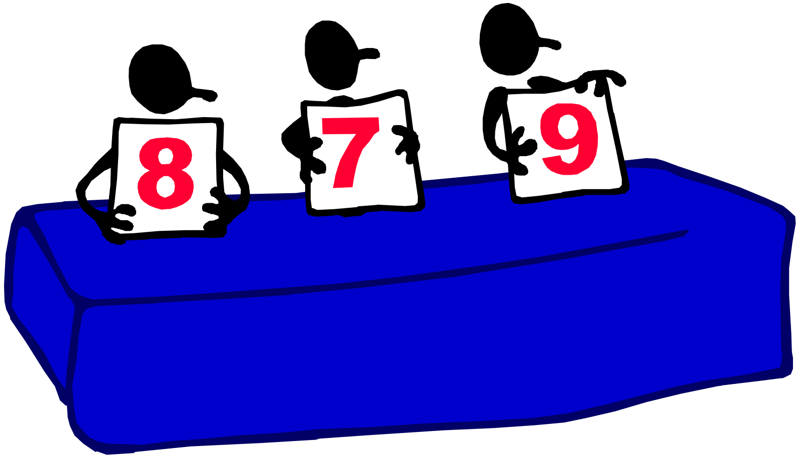
Strategies

- Expert review
  - Content experts: the scope, sequence, and accuracy of the program's content
  - Instructional experts: the effectiveness of the program
  - Graphic experts: appeal, look and feel of the program



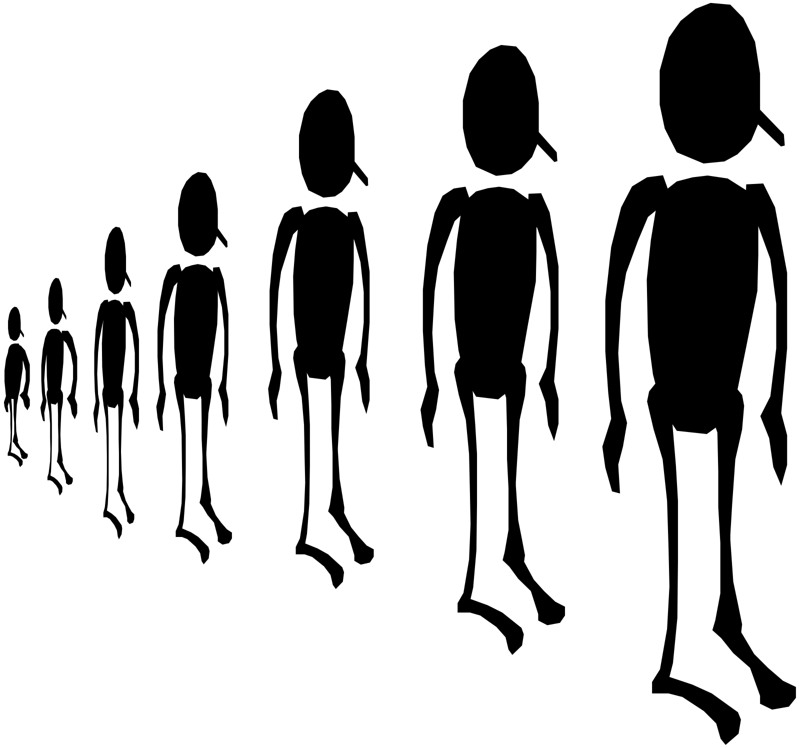
Strategies II

- User review
  - A sample of targeted learners whose background are similar to the final intended users;
  - Observations: users' opinions, actions, responses, and suggestions



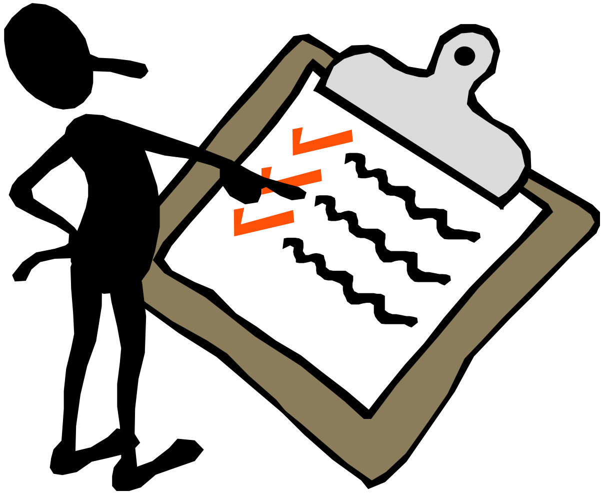
Strategies III

- Field tests
  - Alpha or Beta tests



Who is the evaluator?

- Internal
  - Member of design and development team



When to stop?

- Cost
- Deadline
- Sometimes, just let things go!

Why Evaluate?

- Multiple methods are available to classify or predict
- For each method, multiple choices are available for settings
- To choose best model, need to assess each model’s performance

Accuracy Measures (Classification)

Misclassification error

- Error = classifying a record as belonging to one class when it belongs to another class.
- Error rate = percent of misclassified records out of the total records in the validation data

Naïve Rule

Naïve rule: classify all records as belonging to the most prevalent class

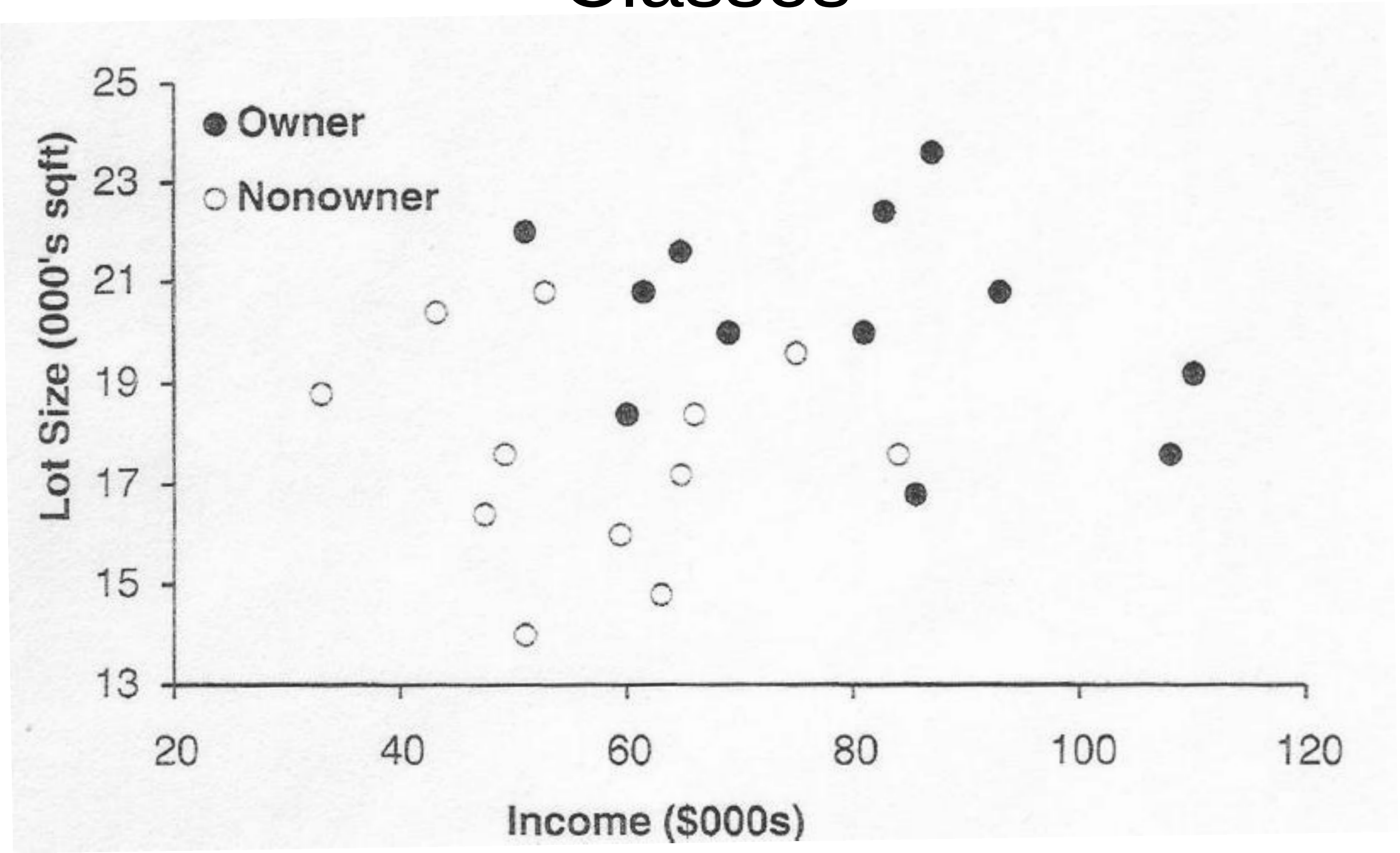
- Often used as benchmark: we hope to do better than that
- Exception: when goal is to identify high-value but rare outcomes, we may do well by doing worse than the naïve rule (see “lift” – later)

Separation of Records

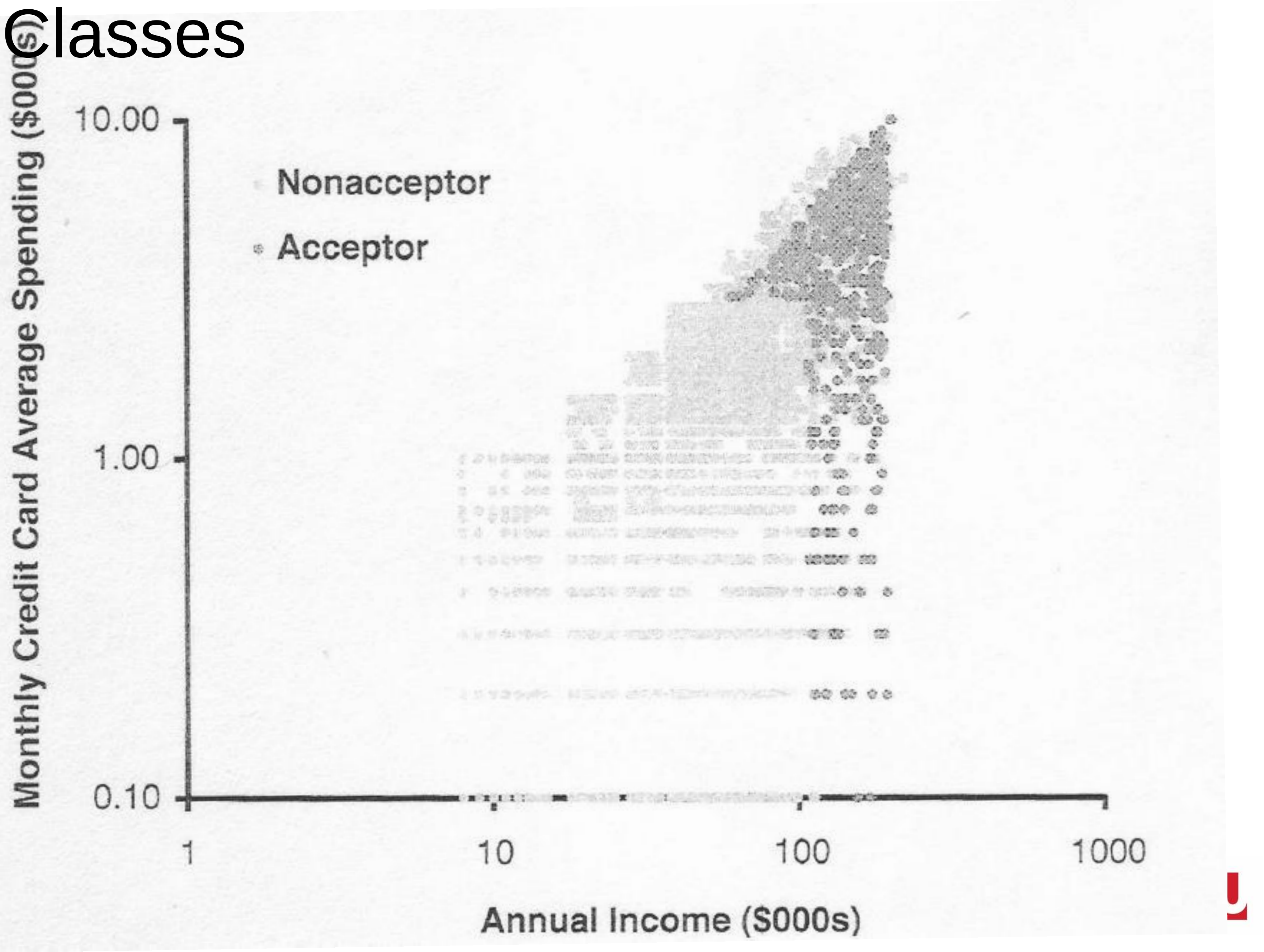
“High separation of records” means that using predictor variables attains low error

“Low separation of records” means that using predictor variables does not improve much on naïve rule

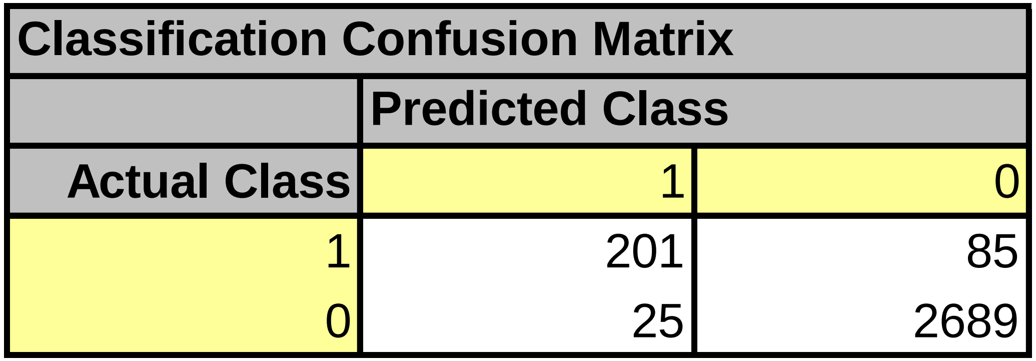
High Level of Separation Between Classes



Low Level of Separation Between Classes



Confusion Matrix



Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

201 1’s correctly classified as “1”

85 1’s incorrectly classified as “0”

25 0’s incorrectly classified as “1”

2689 0’s correctly classified as “0”

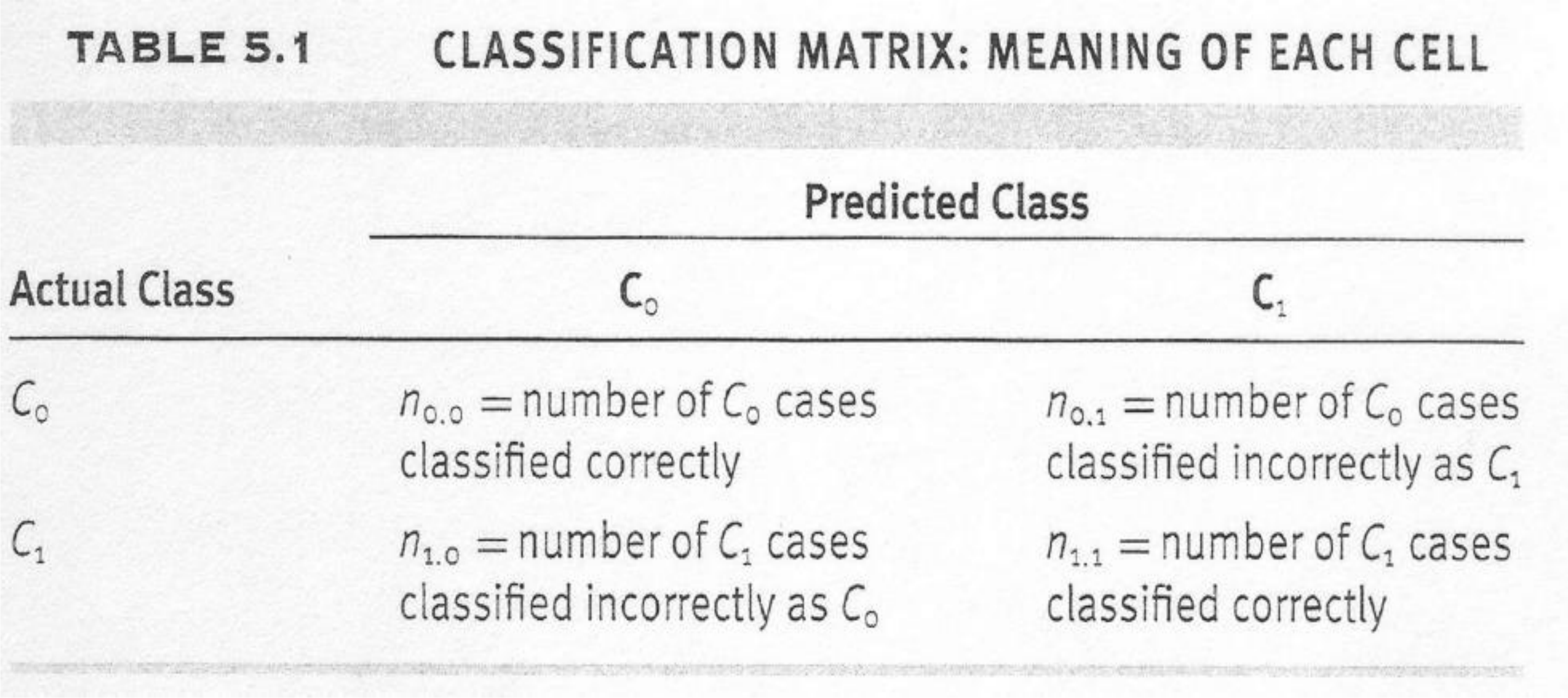


TABLE 5.1 CLASSIFICATION MATRIX: MEANING OF EACH CELL

	Predicted Class	
Actual Class	C_0	C_1
C_0	$n_{0.0}$ = number of C_0 cases classified correctly	$n_{0.1}$ = number of C_0 cases classified incorrectly as C_1
C_1	$n_{1.0}$ = number of C_1 cases classified incorrectly as C_0	$n_{1.1}$ = number of C_1 cases classified correctly

Error Rate

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

Overall error rate = (25+85)/3000 = 3.67%

Accuracy = 1 – err = (201+2689) = 96.33%

If multiple classes, error rate is:

(sum of misclassified records)/(total records)

Cutoff for classification

Most DM algorithms classify via a 2-step process:

For each record,

1. Compute **probability of belonging to class “1”**
2. Compare to cutoff value, and classify accordingly

- Default cutoff value is 0.50

  If >= 0.50, classify as “1”

  If < 0.50, classify as “0”

- Can use different cutoff values
- Typically, error rate is lowest for cutoff = 0.50

Cutoff Table

Actual Class	Prob. of "1"	Actual Class	Prob. of "1"
1	0.996	1	0.506
1	0.988	0	0.471
1	0.984	0	0.337
1	0.980	1	0.218
1	0.948	0	0.199
1	0.889	0	0.149
1	0.848	0	0.048
0	0.762	0	0.038
1	0.707	0	0.025
1	0.681	0	0.022
1	0.656	0	0.016
0	0.622	0	0.004

- If cutoff is 0.50: 13 records are classified as “1”
- If cutoff is 0.80: seven records are classified as “1”

Confusion Matrix for Different Cutoffs

Cut off Prob.Val. for Success (Updatable)	**0.25**

Classification Confusion Matrix		
	Predicted Class	
Actual Class	owner	non-owner
owner	11	1
non-owner	4	8

Cut off Prob.Val. for Success (Updatable)	**0.75**

Classification Confusion Matrix		
	Predicted Class	
Actual Class	owner	non-owner
owner	7	5
non-owner	1	11

When One Class is More Important

In many cases it is more important to identify members of one class

- Tax fraud
- Credit default
- Response to promotional offer
- Detecting electronic network intrusion
- Predicting delayed flights

In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention

Alternate Accuracy Measures

If “C_1” is the important class,

Sensitivity = % of “C_1” class correctly classified

Sensitivity = $n_{1,1} / (n_{1,0} + n_{1,1})$

Specificity = % of “C_0” class correctly classified

Specificity = $n_{0,0} / (n_{0,0} + n_{0,1})$

False positive rate = % of predicted “C_1’s” that were not “C_1’s”

False negative rate = % of predicted “C_0’s” that were not “C_0’s”

Model Evaluation

Please review the following videos to get an appreciation of how to evaluate a machine learning model.

Train /Test Splitting Your Data
https://www.youtube.com/watch?v=VcSNDMBTE-s

Testing and Error Metrics
https://www.youtube.com/watch?v=aDW44NPhNw0

Lesson Summary

Day 6 and 7 Lesson Summary

During Day 6 and 7 you learned to:

- Identify some common application areas of analytic models
- Describe the purpose of various analytical modeling techniques
- Describe model features
- Select model features
- Describe the purpose of dimension reduction
- Evaluate model performance