Programming Machine Learning Applications

Lecture One: A Review of Machine Learning Concepts

Dr. Aleksandar Velkoski

About Me

Work for REALTORS®

Teach at DePaul

Advise Startups

Volunteer at Chicago ML

About You

Name

Degree Program

Work Experience

Interests

Course Overview

Purpose of Class

To strengthen hands-on experience developing machine learning algorithms in the context of popular applications.

Assignments

Assignment 0 - Introductions

Assignment 1 - Basic Data Processing & Analysis

Assignment 2 - k-Nearest-Neighbor Classification

Assignment 3 - Linear Regression & Clustering

Assignment 4 - PCA, ARM, & Item-Based Recommendation

Final Project

Data Analysis

Application Development

Tentative Schedule

Show schedule.

Office Hours

Tuesday 4:00pm to 5:30pm
CDM Center Building, Room 522
312-362-1279
avelkosk@cdm.depaul.edu

Machine Learning

Machine Learning?

Computer programs that learn to improve performance at a task based on experience. (Mitchell, 1997).

Supervised Learning

Learning from training data with class labels

Unsupervised Learning

Learning from training data without class labels

Semi-Supervised Learning

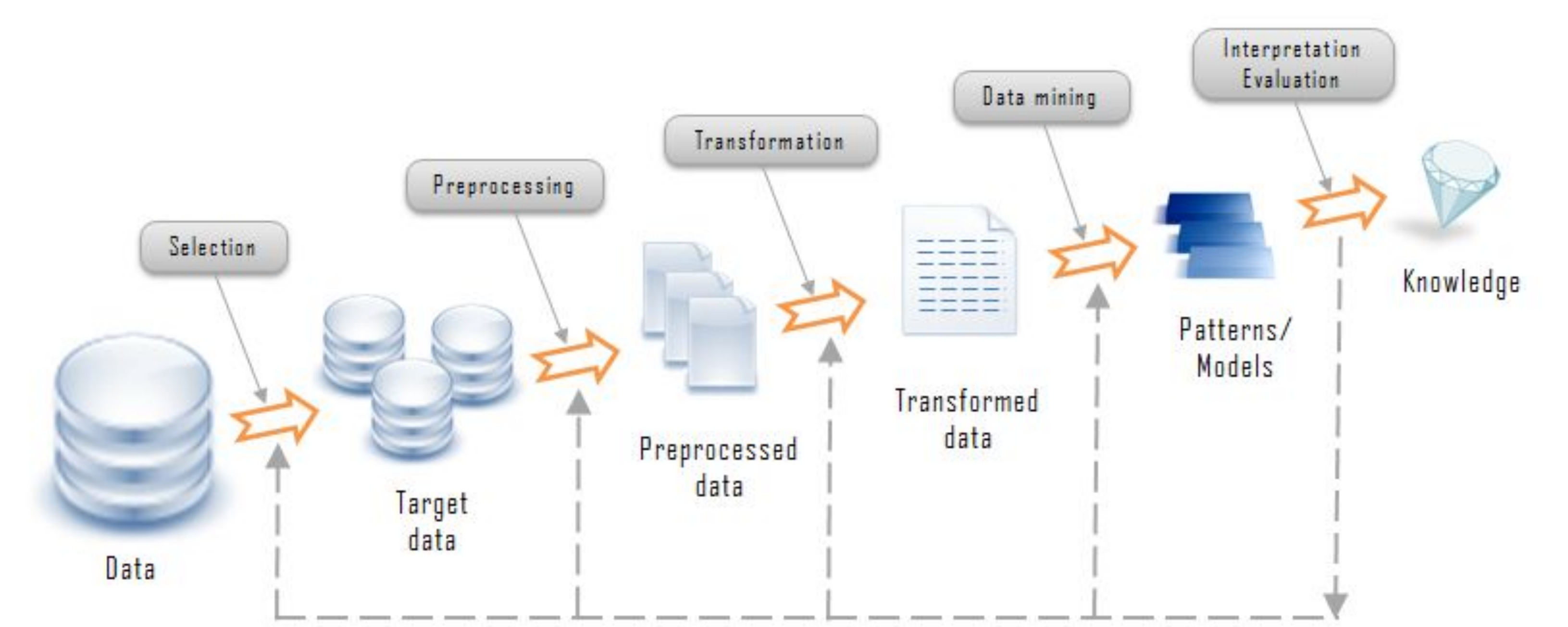
Learning from training data with and without class labels

Reinforcement Learning

Learning from actions that maximize cumulative reward

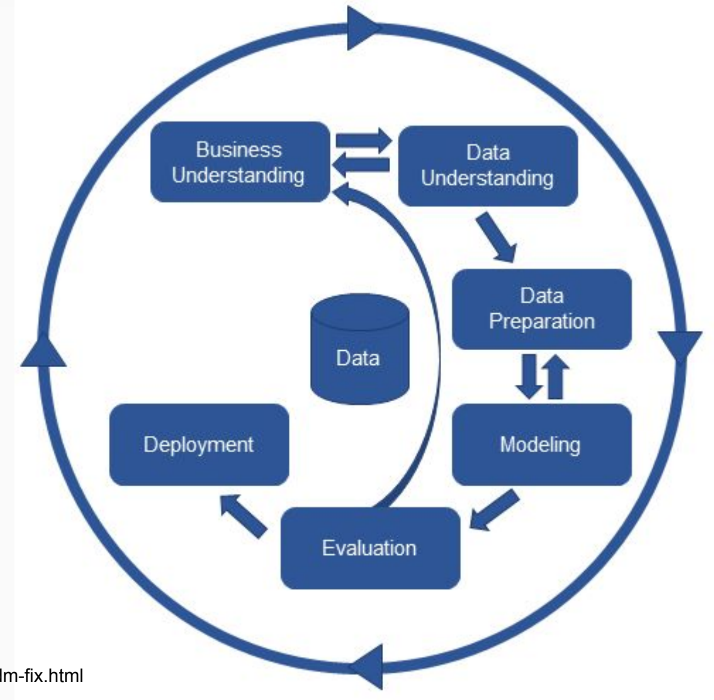
Fundamental Concepts

Knowledge Discovery Framework



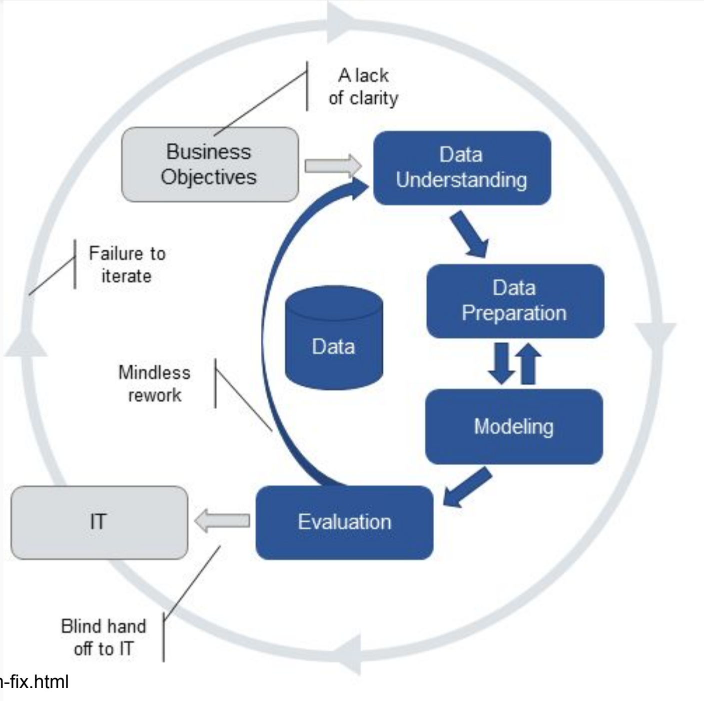
Source: http://www.rithme.eu/?m=home&p=kdprocess&lang=en

CRISP-DM Framework



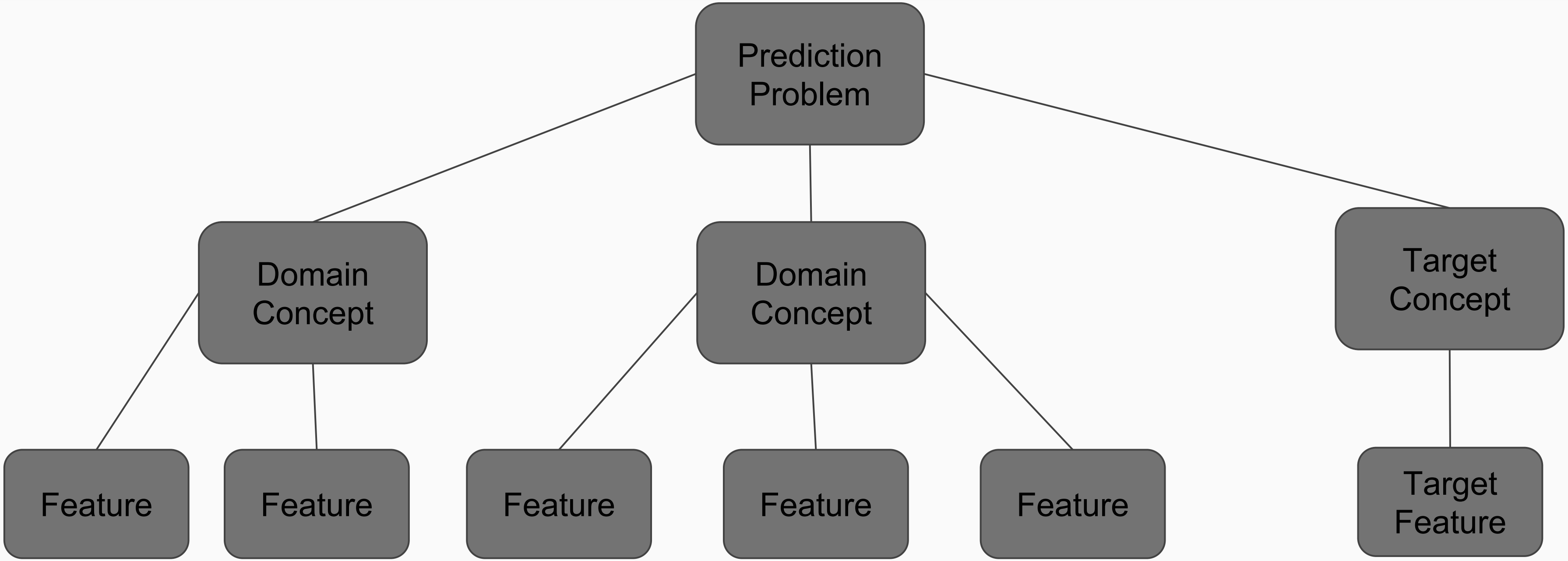
Source: http://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

Issues with CRISP-DM in Practice



Source: http://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

From Domain Concepts to Features

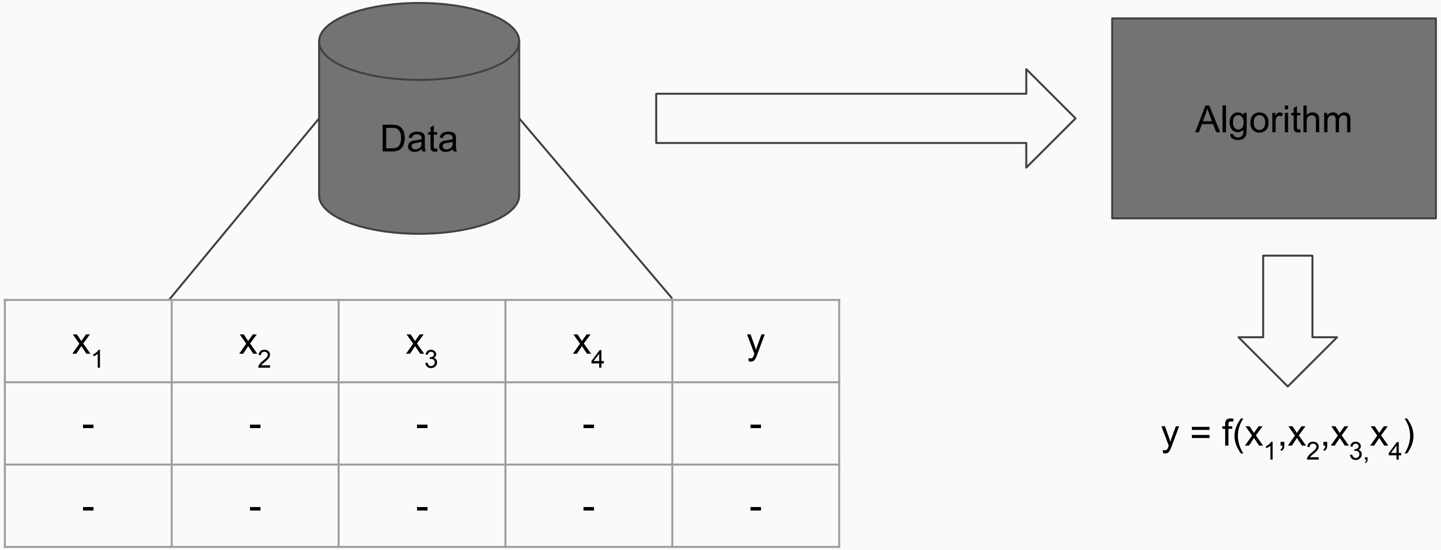


Instances and Features

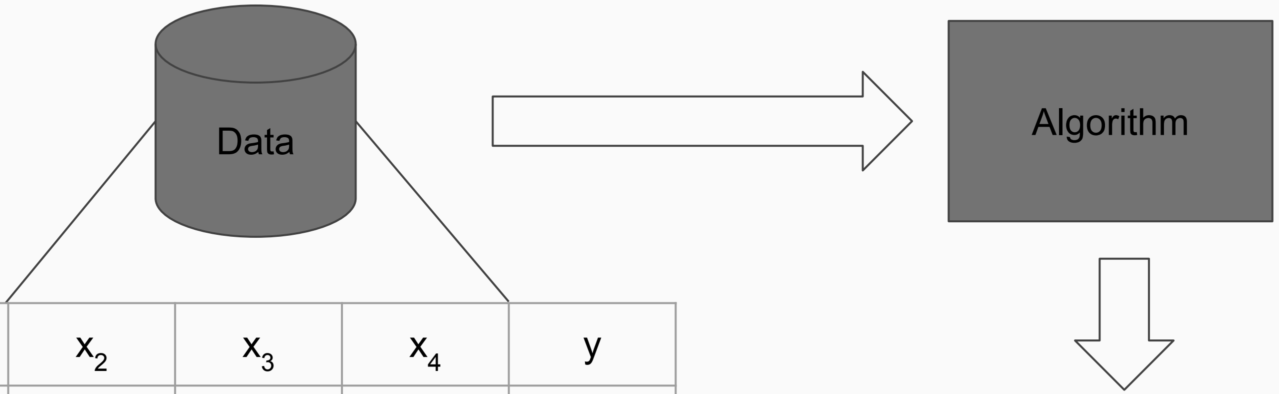
Database rows → instances
Database columns → features

loan_id	rate	fico	dti	loan_term
1	8.725	625	32.5	360
2	6.000	690	10.0	360
3	9.500	550	38.7	360
4	4.950	795	15.0	180

Basic Learning Model



Example of Basic Learning Model



x_1	x_2	x_3	x_4	y
Mike	P	7	Y	Y
Nancy	AP	3	N	N

IF x_2 = “P”
OR $x_3 > 6$
OR x_4 = “Y”
THEN y = “Y”

Ill-posed Problem

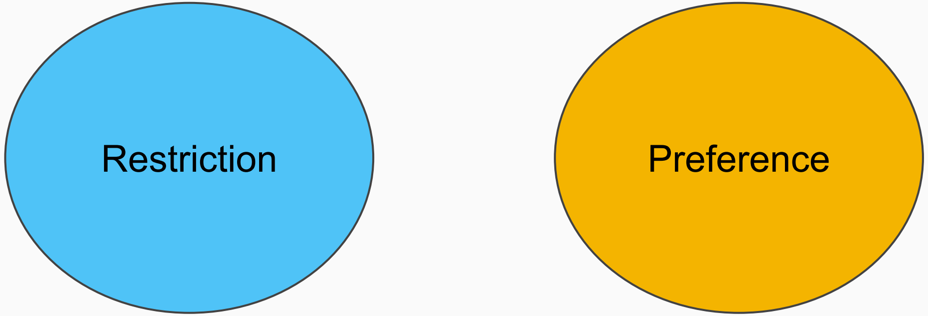
Unique solution cannot be determined using only the available training data.

x_1	x_2	x_3	x_4	y
0	0	1	0	0
0	1	0	0	0
0	0	1	1	1
1	0	0	1	1
0	1	1	0	0
1	1	0	0	0

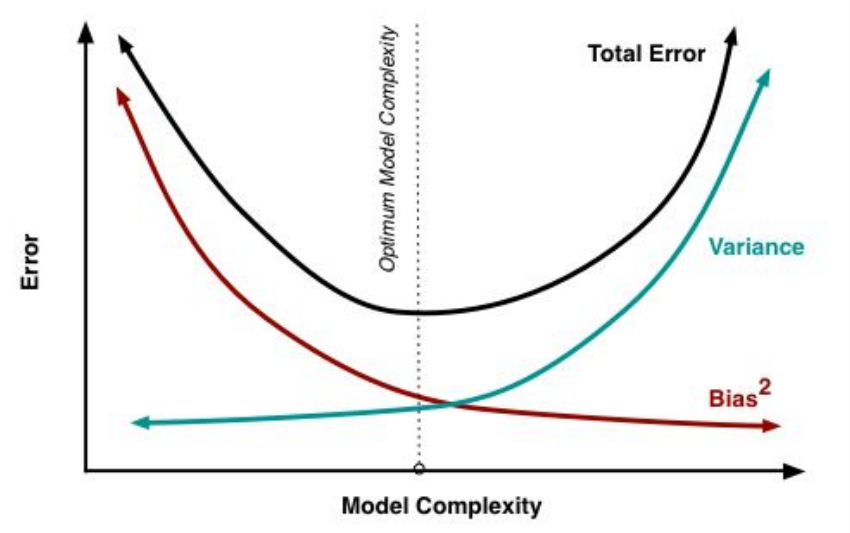
Source: https://web.engr.oregonstate.edu/~tgd/classes/534/slides/part1.pdf

Inductive Bias

A set of assumptions that defines model selection criteria.

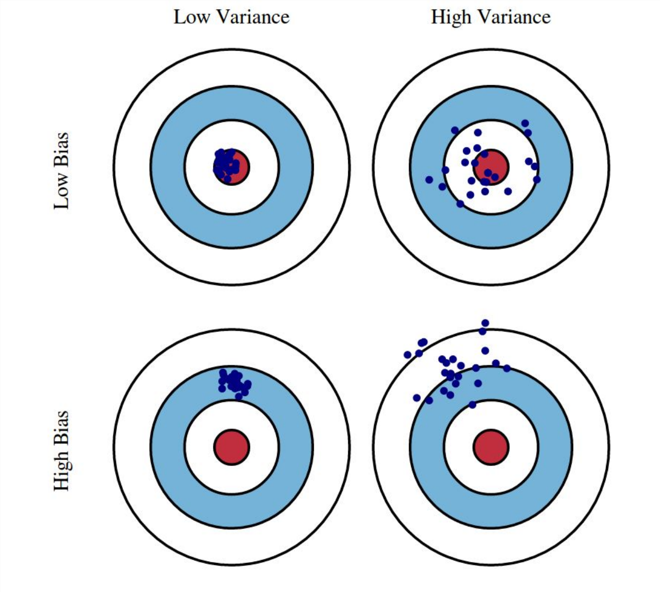


Model Complexity



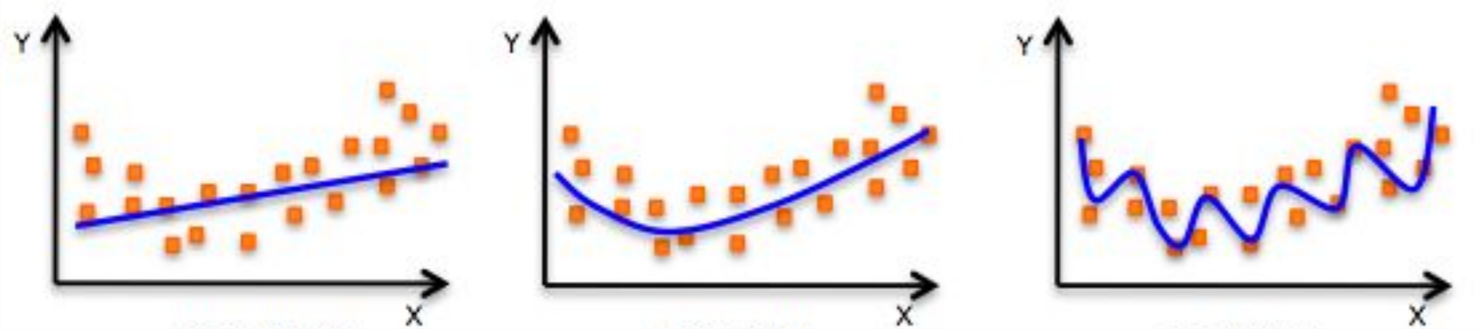
Source: http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

Bias-Variance Tradeoff



Source: http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote12.html

Underfitting and Overfitting



Source: https://web.stanford.edu/class/cs46n/day_two_readings.htm

Getting to Know your Data

Too Many Algorithms!

Knowing the problem, and your data, can help you with choosing the right approach.

Invalid Data?

Real world data is often *incomplete*, *inconsistent*, and *noisy*.

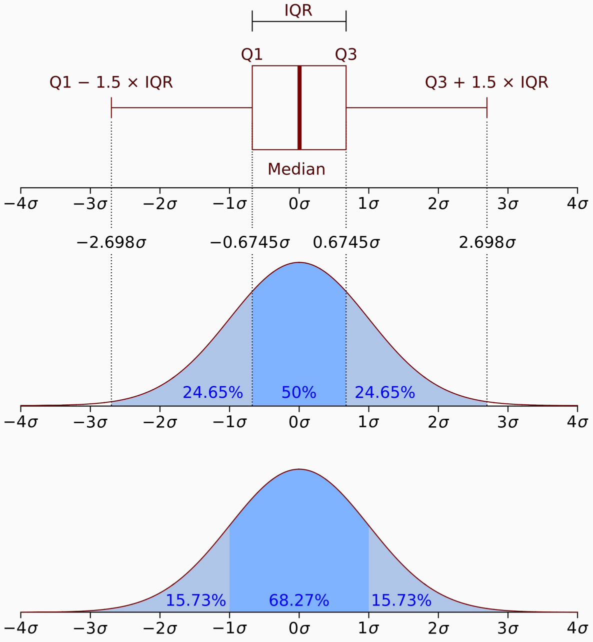
Trust?

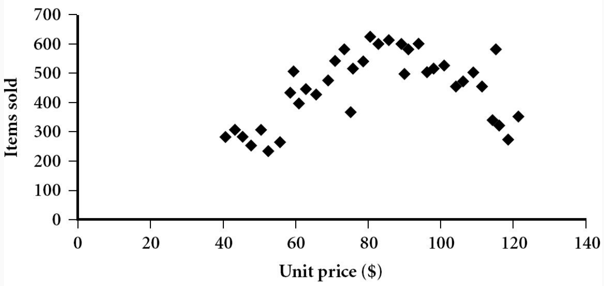
If you can’t trust the data, can you trust the decision?

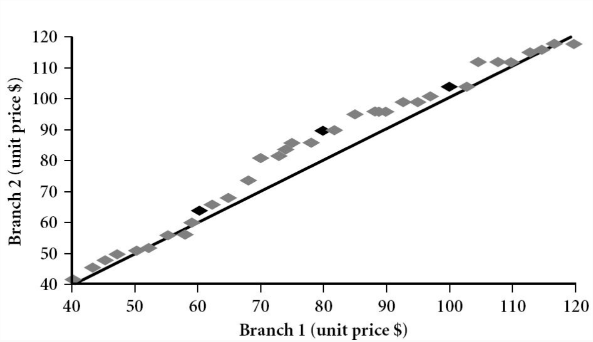
Analytic Base Table

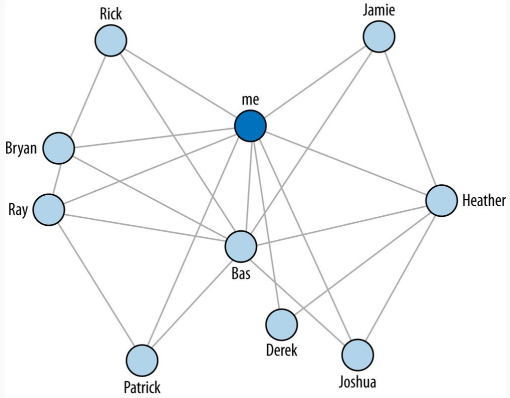
loan_id	rate	fico	dti	loan_term
1	8.725	625	32.5	360
2	6.000	690	10.0	360
3	9.500	550	38.7	360
4	4.950	795	15.0	180

Visualizing Data









Covariance

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

Suppose stocks A and B have these values: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4

E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6

Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4

Correlation

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

Data Quality Reports

Feature	Missing	Mean	...	Max
-	-	-	-	-
-	-	-	-	-

Feature	Missing	Mode	...	Mode Freq.
-	-	-	-	-
-	-	-	-	-

Data Quality Plan

Feature	Issue	Strategy
-	-	-
-	-	-

Dealing with Quality Issues

Addressing Missing Cases

Ignore

Fill in manually

Global constant

Feature mean

Feature mean by class

Infer

Irregular Cardinality

Ignore

Fill in manually

Global constant

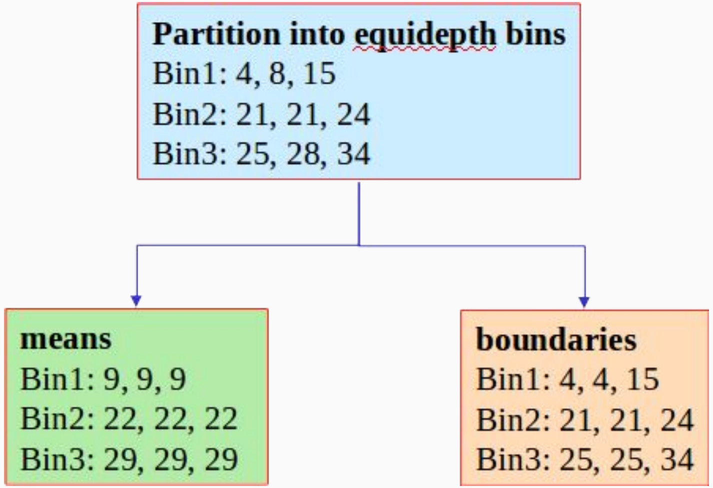
Feature mode

Feature mode by class

Infer

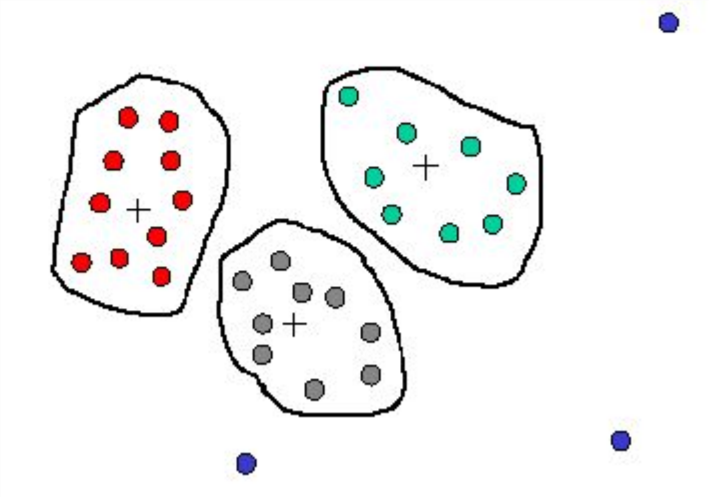
Smoothing Noise via Binning

Original Data: 4, 8, 15, 21, 21, 24, 25, 28, 34

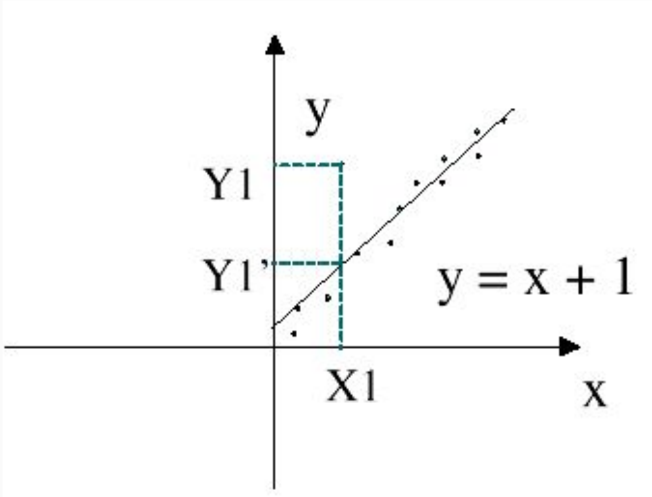


Smoothing Noise via Other Methods

Clustering



Regression



Let's Smooth Temperature

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE

ID	Temperature	
7	58	Bin1
6	65	
5	68	
9	69	Bin2
4	70	
10	71	
8	72	Bin3
12	73	
11	75	
14	75	Bin4
2	80	
13	81	
3	83	Bin5
1	85	

Let’s Smooth Temperature

ID	Temperature	
7	58	Bin1
6	65	
5	68	
9	69	Bin2
4	70	
10	71	
8	72	Bin3
12	73	
11	75	
14	75	Bin4
2	80	
13	81	
3	83	Bin5
1	85	

ID	Temperature	
7	64	Bin1
6	64	
5	64	
9	70	Bin2
4	70	
10	70	
8	73	Bin3
12	73	
11	73	
14	79	Bin4
2	79	
13	79	
3	84	Bin5
1	84	

Let’s Smooth Temperature

ID	Outlook	Temperature	Humidity	Windy
1	sunny	84	85	FALSE
2	sunny	79	90	TRUE
3	overcast	84	78	FALSE
4	rain	70	96	FALSE
5	rain	64	80	FALSE
6	rain	64	70	TRUE
7	overcast	64	65	TRUE
8	sunny	73	95	FALSE
9	sunny	70	70	FALSE
10	rain	70	80	FALSE
11	sunny	73	70	TRUE
12	overcast	73	90	TRUE
13	overcast	79	75	FALSE
14	rain	79	80	TRUE

Handling Outliers

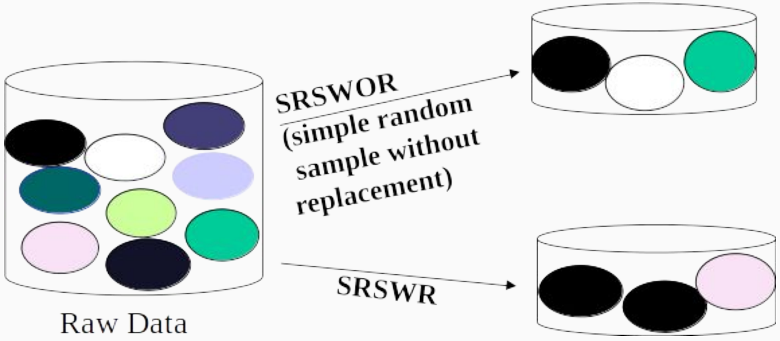
Clamp all values above an upper threshold, and below a lower threshold, to threshold values. Else, return the relevant value.

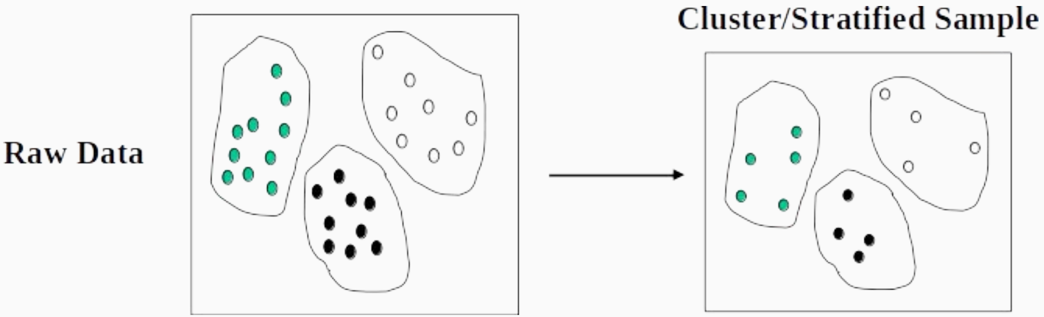
$$a_i = \begin{cases} \text{lower} & \text{if } a_i < \text{lower} \\ \text{upper} & \text{if } a_i > \text{upper} \\ a_i & \text{otherwise} \end{cases}$$

Handling Outliers

Are there any other techniques?

Sampling Techniques





Data Preparation

About Normalization

Adjusting values measured on different scales to a common scale.

Normalization Techniques

Min-max normalization: linear transformation from v to v'

v' = [(v - min)/(max - min)] x (newmax - newmin) + newmin

z-score normalization: based on mean and standard deviation

v' = (v - Mean) / StandardDeviation

Decimal scaling: moves the decimal by j positions such that j is the minimum number of positions moved so that absolute maximum value falls in [0..1]

v' = v / 10j

Min-Max Example

v’ = [(v - min)/(max - min)] x (newmax - newmin) + newmin

ID	Gender	Age	Salary
1	F	27	19,000
2	M	51	64,000
3	M	52	100,000
4	F	33	55,000
5	M	45	45,000

→

ID	Gender	Age	Salary
1	1	0.00	0.00
2	0	0.96	0.56
3	0	1.00	1.00
4	1	0.24	0.44
5	0	0.72	0.32

Z-Score Example

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Mean: 80.3
Std: 9.84

$v' = (v - \text{Mean}) / \text{StandardDeviation}$

Humidity
0.48
0.99
-0.23
1.60
-0.03
-1.05
-1.55
1.49
-1.05
-0.03
-1.05
0.99
-0.54
-0.03

Decimal Scaling

$$v' = v / 10j$$

If v in [-56 .. 9976] and j=4 $\Rightarrow$ v' in [-0.0056 .. 0.9976]

About Discretization

3 Types of attributes

- Nominal - values from an unordered set (also “categorical” attributes)
- Ordinal - values from an ordered set
- Numeric /continuous - real numbers (but sometimes also integer values)

Reduce the number of values for a given continuous features and generate concept hierarchies

For example, collecting and replacing low level concepts (e.g., numeric values for “age”) by higher level concepts (e.g., “young”, “middle aged”, “old”)

Discretization Techniques

Binning - Top-down split, unsupervised

Histogram analysis - Top-down split, unsupervised

Clustering analysis - Unsupervised, top-down split or bottom-up merge

Decision-tree analysis - Supervised, top-down split

Correlation analysis - Unsupervised, bottom-up merge

Discretization via Binning

Equal-width (distance) partitioning

- Divides the range into N intervals of equal size: uniform grid
- If A and B are the lowest and highest values, the width of intervals will be: $W = (B - A)/N$.
- The most straightforward, but outliers may dominate presentation
- Skewed data is not handled well

Equal-depth (frequency) partitioning

- Divides the range into N intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky

Discretization via Classification and Correlation

Classification (e.g., decision tree analysis)

- Supervised: Given class labels, e.g., cancerous vs. benign
- Using entropy to determine split point (discretization point)
- Top-down, recursive split

Correlation analysis (e.g., Chi-merge: $\chi 2$-based discretization)

- Supervised: use class information
- Bottom-up merge: merge the best neighboring intervals (those with similar distributions)
- Merge performed recursively, until a predefined stopping condition

Discretization Example

Humidity
85
90
78
96
80
70
65
95
70
80
70
90
75
80

Low = 60-69
Normal = 70-79
High = 80+

Humidity
High
High
Normal
High
High
Normal
Low
High
Normal
High
Normal
High
Normal
High

From Categories to Numbers

ID	Outlook	Temperature	Humidity	Windy
1	sunny	85	85	FALSE
2	sunny	80	90	TRUE
3	overcast	83	78	FALSE
4	rain	70	96	FALSE
5	rain	68	80	FALSE
6	rain	65	70	TRUE
7	overcast	58	65	TRUE
8	sunny	72	95	FALSE
9	sunny	69	70	FALSE
10	rain	71	80	FALSE
11	sunny	75	70	TRUE
12	overcast	73	90	TRUE
13	overcast	81	75	FALSE
14	rain	75	80	TRUE

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy
overcast	rain	sunny			TRUE	FALSE
0	0	1	85	85	0	1
0	0	1	80	90	1	0
1	0	0	83	78	0	1
0	1	0	70	96	0	1
0	1	0	68	80	0	1
0	1	0	65	70	1	0
1	0	0	64	65	1	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Data Reduction

Data is often too large; reducing data can improve performance

Data reduction consists of reducing the representation of the data set while producing the same (or almost the same) results

Data reduction includes:

- Data cube aggregation
- Dimensionality reduction
- Discretization
- Numerosity reduction

Cube Aggregations

Dimensionality Reduction

Curse of dimensionality

- When dimensionality increases, data becomes increasingly sparse
- Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
- The possible combinations of subspaces will grow exponentially

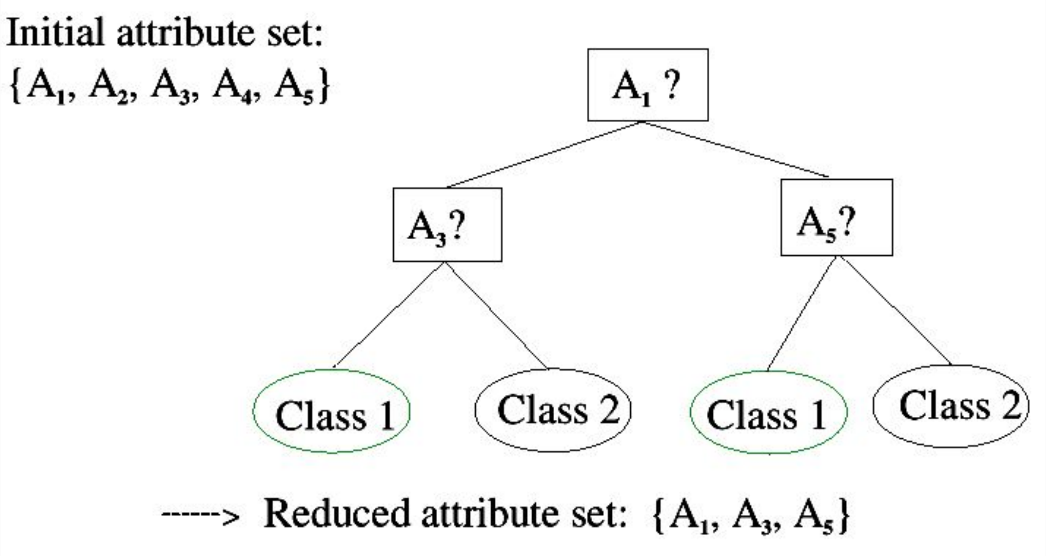
Dimensionality reduction

- Avoid the curse of dimensionality
- Help eliminate irrelevant features and reduce noise
- Reduce time and space required in data mining
- Allow easier visualization

Dimensionality reduction techniques

- Principal Component Analysis
- Feature selection (tree induction, heuristic search)
- Feature engineering

Tree Induction Example



Wrapping-up the Lecture

Questions

What is domain knowledge and why is it important?

What is meant by inductive bias?

Explain ill-posed problems.

What is the difference between normalization and discretization?