

ACP: DSPA

Strategic Business Analysis Using Predictive Analytics



Dr. Ash Pahwa

Lesson 5: Data Exploration

Lesson 5.1: Data Problems in Machine Learning



Outline

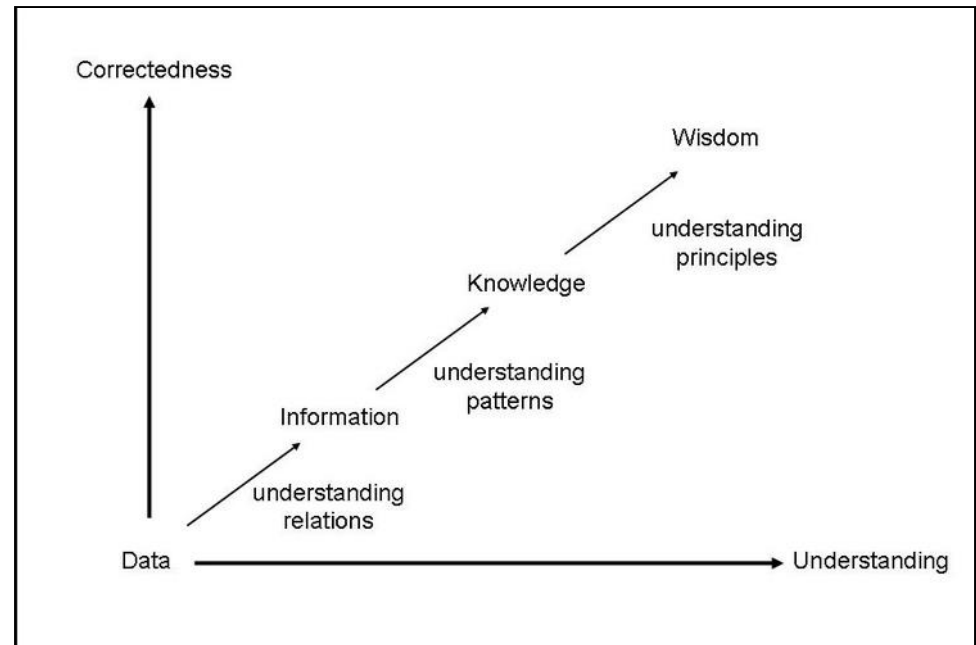
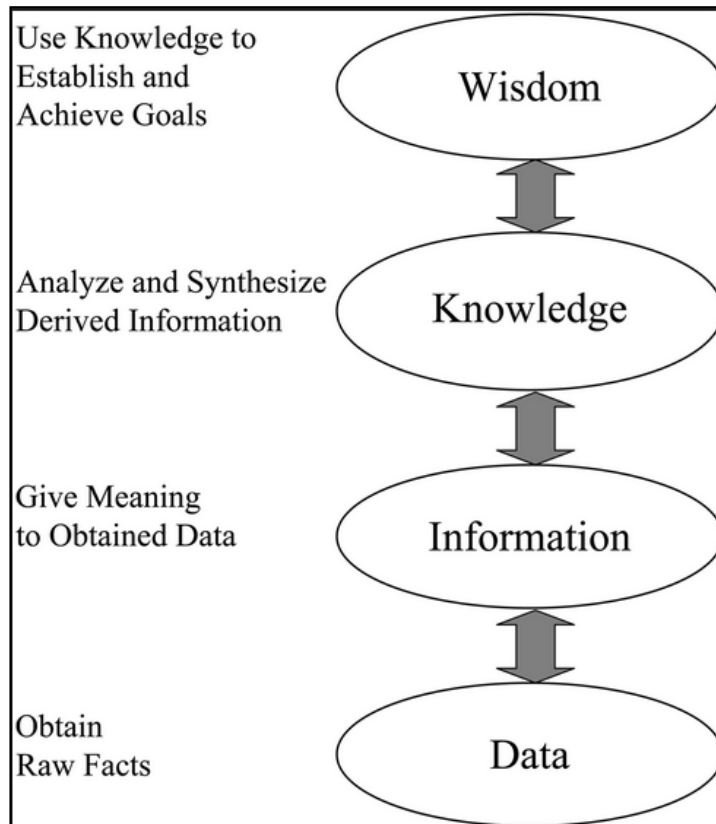
- Data, Information, Knowledge
- Data Characteristics for ML
- Data Problems
- Data Preparation



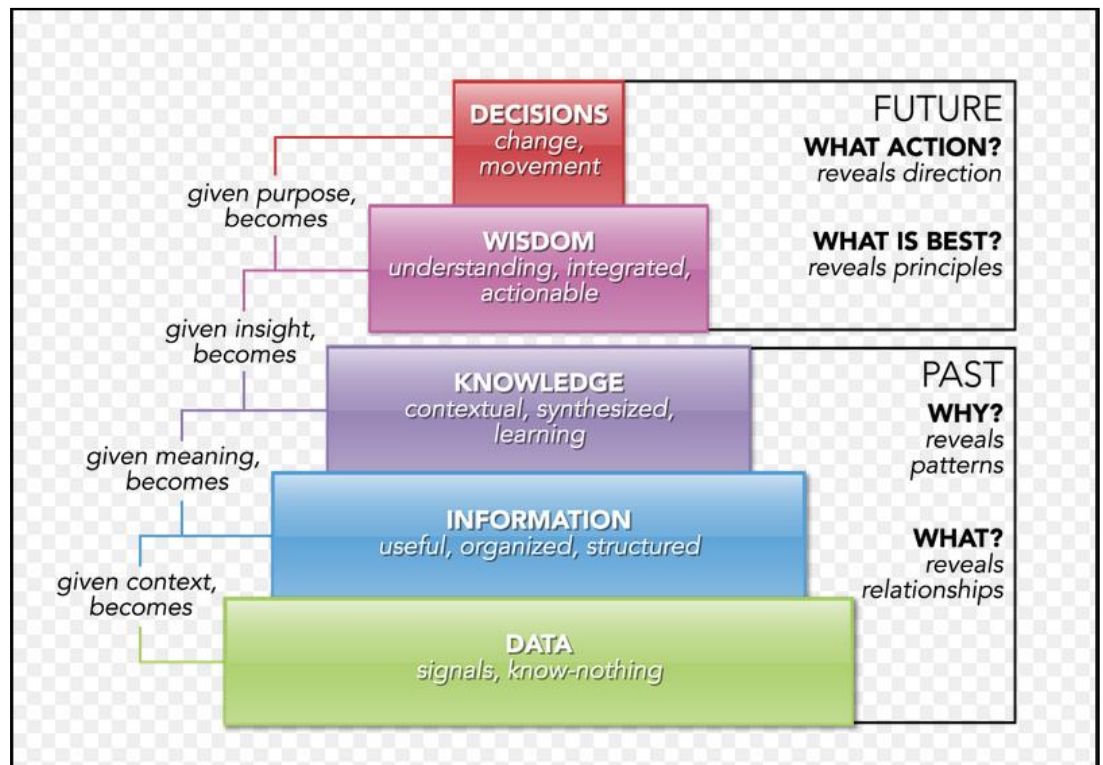
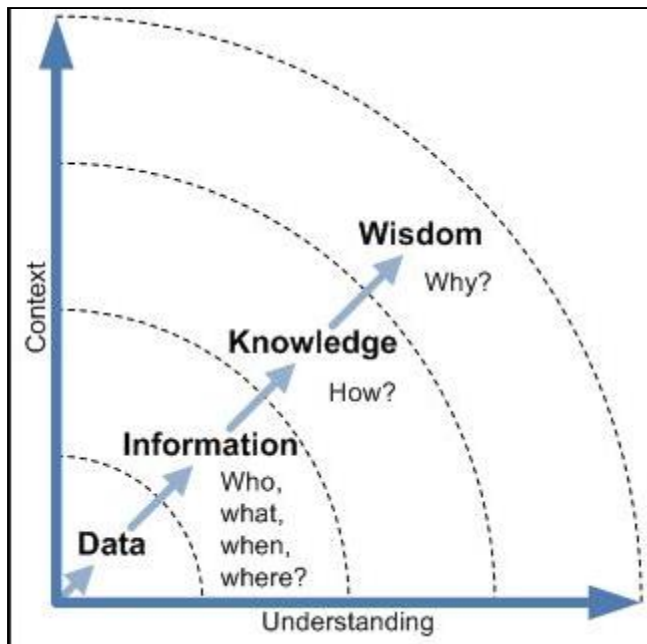
What is Data?

- A collection of facts usually obtained as a result of
 - Experience
 - Observations
 - Experiments
- Data: Lowest level of abstraction
 - From which information or knowledge are derived
- Data: Facts and figures collected, and analyzed, and summarized for presentation and interpretation

What is the difference between Information and Knowledge



What is the difference between Information and Knowledge





Data Characteristics for Machine Learning

- 90% of a successful outcome is contingent on having good data. What does "good" mean in this context?
 - **Reliable** - the effects can be reproduced.
 - Contrary example:
 - Using sales data from March-April to predict November-December data.
 - **Valid** - the data measures what you want it to measures.
 - Contrary example:
 - Number of page visits shows how popular the page is.
 - How well a job candidate does on a brainteaser indicates whether he will be a good hire.
 - Other data issues are small.



Data Problems

- Data Problems
 - Missing Data
 - Sparse Data
 - Inaccurate data
 - Inconsistent Data
 - Redundant Data
- Data Preparation is needed before building a model

Data Issues

- **Missing data** - A small amount of missing data is not a problem, since the process of statistical modeling will smooth over these bumps.
 - However a large amount of missing data will invalidate the model
 - Missing not at random: this can be very serious.
 - Not to be confused with sparse data. Recommendation systems, such as Netflix, have sparse data.
- **Sparse data** - An example is a matrix with users on one dimension and movies on the other. The entries are ratings. Most people haven't seen most of the movies, so the matrix will be mostly empty.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country	January	February	March	April	May	June	July	August	September	October	November	December
2	Angola	11027	4608	9715	8381	5881	10286	9311	2775	3354	9935	7317	3973
3	Burundi	4988	6822			13264	4150	8469	2212	6620	7883	11637	10439
4	Chad	8711	8189	12772	13158	4491	2458	13498			7975	7945	8441
5	Congo	7139	13737	5152	12218		4390	6764		11791	12331	8443	5769
6	Egypt	9474	1278	8594	13938	1354	10004	10240	14054	2510	1360	3830	14432
7	Ethiopia	10590	10572	4933	6058	13121	10772	10602	11251	8648	9382	14768	14022
8	Gabon	7003	2462	5243	8851				13258	7038	14862	4153	9738
9	Ivory Coast				6354	12059	3591	5827	14469	13454	4326	11593	14240
10	Kenya	7234	9542	12112	14368	12704	10580	7558	5355	7864	11395	3114	8491
11	Libya	10099	11447	12909	7378	12713	13599	13203	8052	6800	12004	2028	6341
12	Madagascar	2013	6180		6785	13269	11403	5693	8438	8088	7647	5806	
13	Morocco	5463	10133	4515	6198	13884	14120	2120			5933	8445	12781
14	Namibia	7810	7507	9088	14838	12787	6350	7306	13090	1911	8260	12865	7507
15	Somalia	7505	13614	2231	5130	10979	12869	1449	4973	3609	14918	4638	3051
16	Swaziland	8035	12596	5948	6166	6198	13517	9700	9470	5804	9105	12671	3518
17	Zambia	7042	7553	6785	7141	6368	2492	2557	5078	14694	5739	4812	14291

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Country	January	February	March	April	May	June	July	August	September	October	November	December
2	Angola												
3	Burundi					11491							
4	Chad												
5	Congo												
6	Egypt												
7	Ethiopia												
8	Gabon												
9	Ivory Coast												
10	Kenya						4417						
11	Libya												
12	Madagascar												
13	Morocco										5284	9331	
14	Namibia												
15	Somalia												
16	Swaziland												
17	Zambia					2767							



Data Issues



■ Inaccurate data

- If a metric is measured by 3 different systems, you are likely to have 3 different values.
- It matters when the values are way off, or you have reason to believe the differences are systematic.



Inconsistent Data

Number, Date, Time, Zip Code,
Phone Number, SS#, Scientific

- Data in different Format
 - Date
 - Month/Date/Year
 - Date/Month/Year
 - Time
 - Zip Code

Inconsistent Data

Number, Date, Time, Zip Code,
Phone Number, SS#, Scientific

Format Cells

Number Alignment Font Border Fill

Category:

- General
- Number
- Currency
- Accounting
- Date
- Time
- Percentage
- Fraction
- Scientific
- Text
- Special
- Custom

Sample

Decimal places: 2

☐ Use 1000 Separator (,)

Negative numbers:

- 1234.10
- 1234.10
- (1234.10)
- (1234.10)

Format Cells

Number Alignment Font Border Fill

Category:

- General
- Number
- Currency
- Accounting
- Date
- Time
- Percentage
- Fraction
- Scientific
- Text
- Special
- Custom

Sample

Type:

- *3/14/2012
- *Wednesday, March 14, 2012
- 3/14
- 3/14/12
- 03/14/12
- 14-Mar
- 14-Mar-12

Locale (location):

English (United States)

Format Cells

Number Alignment Font Border Fill

Category:

- General
- Number
- Currency
- Accounting
- Date
- Time
- Percentage
- Fraction
- Scientific
- Text
- Special
- Custom

Sample

Type:

- *1:30:55 PM
- 13:30
- 1:30 PM
- 13:30:55
- 1:30:55 PM
- 30:55.2
- 37:30:55

Locale (location):

English (United States)

Format Cells

Number Alignment Font Border Fill

Category:

- General
- Number
- Currency
- Accounting
- Date
- Time
- Percentage
- Fraction
- Scientific
- Text
- Special
- Custom

Sample

Type:

- Zip Code
- Zip Code + 4
- Phone Number
- Social Security Number

Locale (location):

English (United States)

Inconsistent Data

Text Data

- Text Data
 - Padded with extra blanks

	A	B	C	D	E	F	G	H	I
1	Add 1	Add2	City	State	ZIP	COMPANY	COUNTRY	CREATEDBY	CREATED
4	6169 Honey Hill		Blowing Rocks	CA	95580	Acme Corp	USA	DAVID	20060517
5	8601 Broad Route	Suite 3890	Truckhaven	Cali	95580	Allied Biscuit	USA	ANNA	20060517
8	16 Rockey Mountain		Handuras	Calif	95580	Extensive Enterprise	USA	ANNA	20060523
10	9597 Stony Log Swale	Bldg A	Asquith	CA	95580	Galaxy Corp	USA	ANNA	20060524
11	3047 Merry Glade		Storthoaks	Calif	95580	Globex Corporation	USA	PHIL	20060512
12	3982 Gentle Treasure Avenue		Red Coat Woods	Cali	95580	Globo-Chem	USA	PHIL	20060620
18									

Redundant Data

- Storing Age and date-of-birth together
- Storing Duration and end-date together

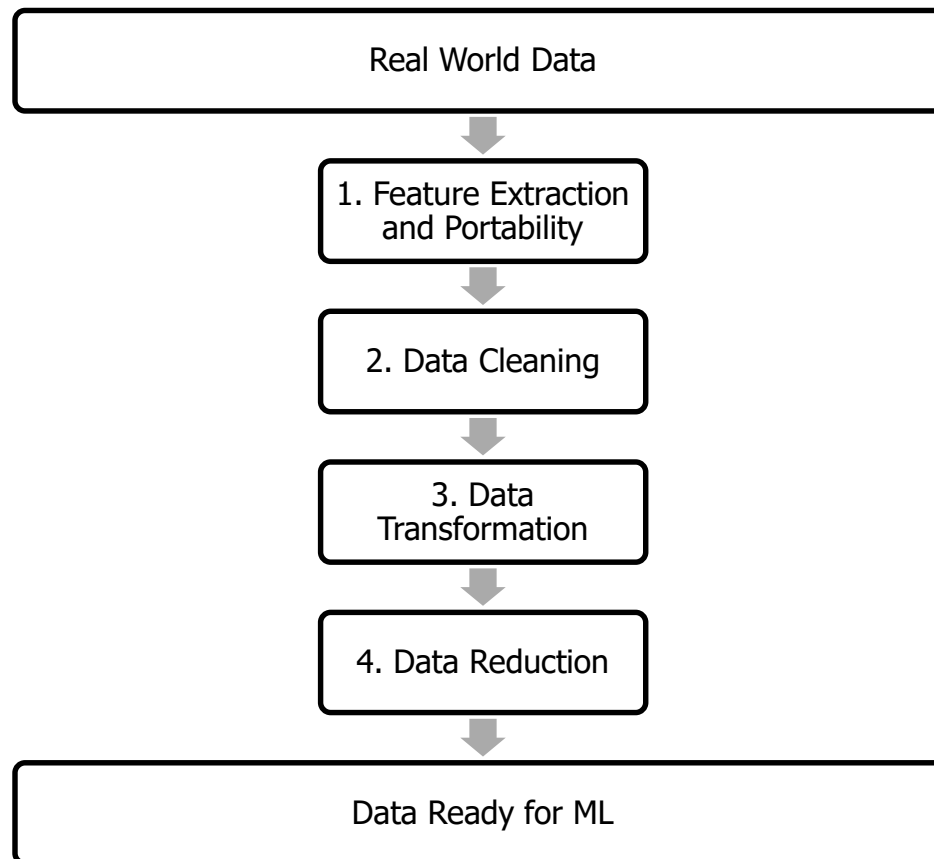
Patient Id	Name	D.o.B	Gender	Phone	Doctor Id	Doctor	Room	
134	Jeff	4-Jul-1993	Male	7876453	01	Dr Hyde	03	Duplicate
178	David	8-Feb-1987	Male	8635467	02	Dr Jekyll	06	
198	Lisa	18-Dec-1979	Female	7498735	01	Dr Hyde	03	Duplicate
210	Frank	29-Apr-1983	Male	7943521	01	Dr Hyde	03	Duplicate
258	Rachel	8-Feb-1987	Female	8367242	02	Dr Jekyll	06	



Data Preparation



Data Preparation





1. Feature Extraction

- Feature Selection and Extraction
 - What is a good feature?
 - Ability to Predict – Informative
 - Independent – low correlation with other features
 - Simple – Easy to understand



1. Data Portability

Source Data Type	Destination Data Type	Methods
Numeric	Categorical	Discretization
Categorical	Numeric	Binarization
Text	Numeric	Latent Semantic Analysis (LSA)

- Machine Learning models needs data from more than one data source
- Data is stored in different format in different data sources

Example of Discretization

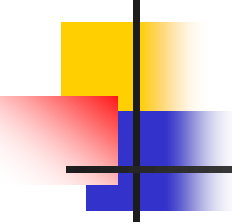
Numeric to Categorical

Student's Earned Points Converted to Grades

- Total 30 students
- Score
 - From 1 to 100

G		H
Points		Grade
91 - 100		A
81 - 90		B
71 - 80		C
61 - 70		D
Less than 60		F

	A	B	
1		Student Score	
2		56	
3		43	
4		81	
5		78	
6		78	
7		93	
8		65	
9		84	
10		80	
11		89	
12		62	
13		75	
14		83	
15		55	
16		59	
17		92	
18		72	
19		55	
20		44	
21		67	
22		87	
23		63	
24		73	
25		63	
26		53	
27		93	
28		54	
29		83	
30		58	
31		72	
32			



Example of Binarization

Categorical to Numeric

	Net Worth	Categorical Variable	Binarization
1	< \$10K	Poor	0 000 001
2	\$10K – \$30K	Lower Middle Class	0 000 010
3	\$30K - \$100K	Middle Class	0 000 100
4	\$100K - \$1M	Upper Middle Class	0 001 000
5	\$1M - \$10M	Rich	0 010 000
6	\$10M - \$1B	Ultra Rich	0 100 000
7	> \$1B	Super Rich	1 000 000



2. Data Cleaning

- Missing data is interpolated
- Inaccurate data is fixed
- Inconsistent data is made consistent
- Redundant data is eliminated



3. Data Transformation

- Suppose 'x' variable is
 - NFL Quarterback's touch down
 - MLB home run
- $y = g(x)$ is a transformation
- Functions used for transformation
 - $\log_{10}(x)$ or $\log_e(x)$
 - $\frac{1}{x}$

3. Data Transformation

- Golf Player Earning
 - Top 100 golfers
- Log of Golf Player Earning

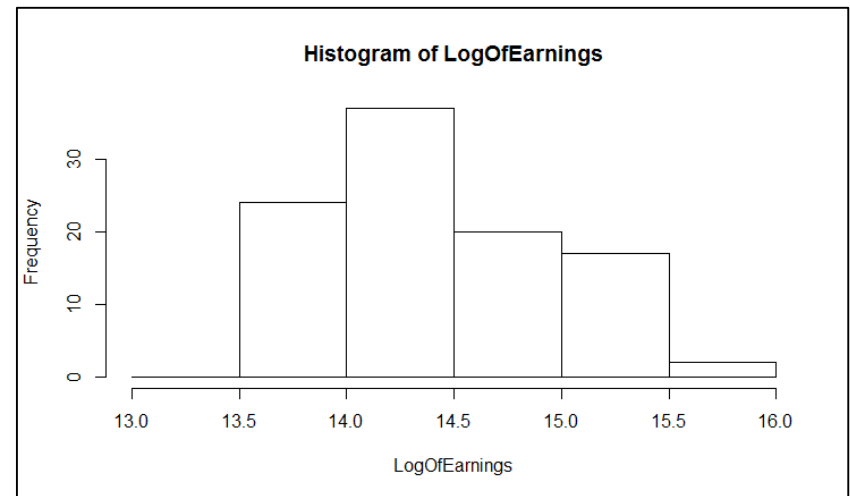
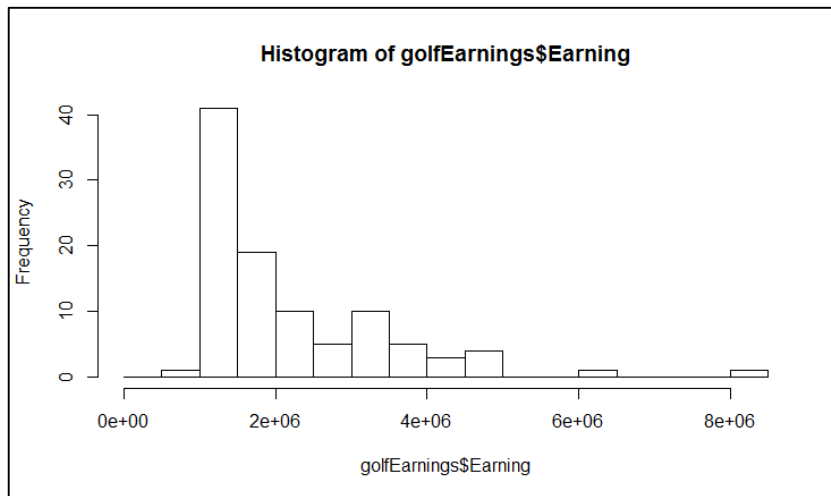
	A	B	C	D	E
1		Name	Earning	LN(Earning)	
2	1	McIlroy, Rory	8047952	15.9	
3	2	Woods, Tiger	6133158	15.63	
4	3	Snedeker, Brandt	4989739	15.42	
5	4	Dufner, Jason	4869304	15.4	
6	5	Watson, Bubba	4644997	15.35	
7	6	Johnson, Zach	4504244	15.32	
8	7	Rose, Justin	4290930	15.27	
9	8	Mickelson, Phil	4203821	15.25	
10	9	Mahan, Hunter	4019193	15.21	
11	10	Bradley, Keegan	3910658	15.18	
12	11	Kuchar, Matt	3903065	15.18	
13	12	Furyk, Jim	3623805	15.1	
14	13	Pettersson, Carl	3538656	15.08	
15	14	Donald, Luke	3512024	15.07	
16	15	Oosthuizen, Louis	3460995	15.06	
17	16	Els, Ernie	3453118	15.05	
18	17	Simpson, Webb	3436758	15.05	
19	18	Stricker, Steve	3420021	15.05	
20	19	Johnson, Dustin	3393820	15.04	
21	20	Garrigus, Robert	3206530	14.98	
22	21	Fowler, Rickie	3066293	14.94	
23	22	Watney, Nick	3044224	14.93	
24	23	Van Pelt, Bo	3043509	14.93	
25	24	Westwood, Lee	3016569	14.92	
26	25	Scott, Adam	2899557	14.88	
27	26	Moore, Ryan	2858944	14.87	
28	27	Piercy, Scott	2699205	14.81	

3. Data Transformation

Histogram of Golf Player Earnings

Earning histogram

- A few good golfers
- And beside these top golfers, most of them poorly
- Log of Earning histogram
 - Distribution of earning is a bell shaped curve



3. Data Transformation

Data Normalization: Standardization & Scaling





Data Standardization & Scaling

- Suppose we have 2 data items
 - Height: varies from 1 – 7 feet
 - Net Worth: \$10,000 - \$100B
- If we use both the variables in a model
 - Net Worth will dominate because it contains large values
- Solution
 - Standardize
 - Scale



Data Standardization and Scaling

- Standardization Data Variation
 - -3 to +3

$$z = \frac{\text{Data Value} - \text{Mean}}{\text{Standard Deviation}} = \frac{y - \mu}{\sigma}$$

- Scaling Data Variation
 - 0 to 1

$$y_i^j = \frac{x_i^j - \min_j}{\max_j - \min_j}$$

Example

	A	B	C	D	E	F	
1		#	Data	Standardization		Scaling	
2		1	124	-0.60		0.12	
3		2	3	-0.99		0.00	
4		3	311	0.00		0.31	
5		4	341	0.10		0.34	
6		5	298	-0.04		0.30	
7		6	136	-0.56		0.13	
8		7	23	-0.93		0.02	
9		8	75	-0.76		0.07	
10		9	5	-0.99		0.00	
11		10	51	-0.84		0.05	
12		11	822	1.65		0.82	
13		12	364	0.17		0.36	
14		13	663	1.14		0.66	
15		14	444	0.43		0.44	
16		15	999	2.22		1	
17							
18		Mean	310.60		Minimum	3	
19		StdDev	309.41		Maximum	999	
20							



Comparison of Standardization and Scaling

- In the presence of outliers in the data
 - Scaling is not effective
 - It will suppress the scaling values of other data elements



3. Other Transformation

- Haar Wavelet Transform
- Discrete Fourier Transformation



4. Data Reduction

First Name	Last Name	City	Date of visit	Amount spent
Jane	Citizen	San Francisco	09-Sep-13	\$250.00
Jane	Citizen	San Francisco	27-Sep-13	\$300.00
Jane	Citizen	San Francisco	15-Oct-13	\$120.00
Jane	Citizen	San Francisco	19-Nov-13	\$450.00
Jim	Southdown	San Diego	24-Oct-13	\$600.00
Simon	Amiet	Monterey	29-Nov-13	\$250.00
Simon	Amiet	Monterey	16-Dec-13	\$550.00
Petra	Southdown	San Diego	13-Dec-13	\$420.00

- Reduction in Number of Rows
 - Sampling
- Reduction in Number of Columns
 - Feature subset selection
 - Data Transformation
 - Rotation of axis system - Principal Component Analysis (PCA)



Summary

- Data, Information, Knowledge
- Data Characteristics for ML
- Data Problems
- Data Preparation