

دوره‌ی آموزشی «علم داده»
Data Science Course



جلسه سی و یکم
نکاتی تكمیلی در خصوص
پیش‌پردازش

مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

PREPROCESSING

هرگونه دستکاری روی دیتاست قبل از دادن آون به مدل رو
پیش‌پردازش می‌گن

Any manipulation of the dataset before
running it through the model



اما چرا این کار رو انجام می دیم؟

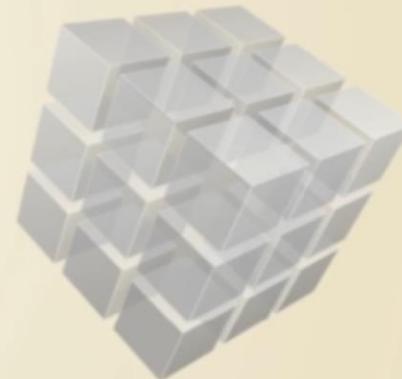
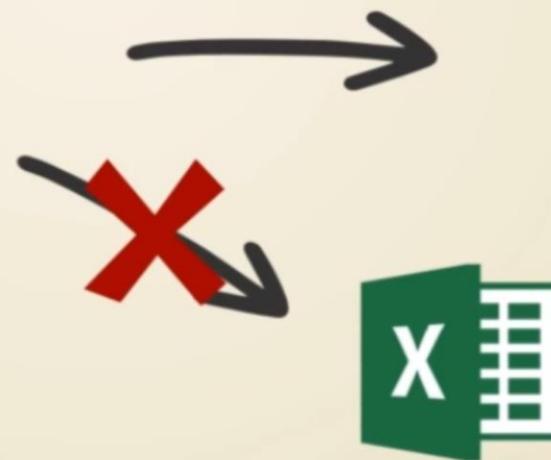
PREPROCESSING

motivation

- Compatibility



TensorFlow



PREPROCESSING

motivation

- Compatibility

Exchange rate

~1

Volume

~100,000

- Orders of magnitude

Mathematically 1 is negligible w.r.t. 100,000

$$(\sim 100,000) + (\sim 1) = \sim 100,000$$

PREPROCESSING

motivation

- Compatibility

- Orders of magnitude

- Generalization



Same model,
different issue



خب. با این مقدمه بدیم سراغ موضوع و بحث
اصلی‌مون

Relative metrics are especially useful when we have time-series data

apple stock price

All Finance News Books Videos More Tools

About 952,000,000 results (0.69 seconds)

Apple Inc

141.50 USD NASDAQ: AAPL

-1.33 (0.93%) ↓ today

Closed: Mehr 8, 20:06 EDT · Disclaimer
After hours 141.84 +0.34 (0.24%)

1D | 5D | 1M | 6M

The screenshot shows a Google search result for "apple stock price". The top navigation bar includes All, Finance, News, Books, Videos, More, and Tools. Below the search bar, it says "About 952,000,000 results (0.69 seconds)". The main result is for Apple Inc., showing a current price of 141.50 USD on the NASDAQ (AAPL). It indicates a decrease of -1.33 (0.93%) from the previous day. The "Closed" price was 142.83 USD at 20:06 EDT. An "After hours" price of 141.84 USD is shown with a 0.34 (0.24%) increase. A 1-day chart is displayed, showing the price fluctuating between 141 and 144 USD throughout the day. The chart has a legend for "Previous close" at 142.83. The chart area is partially covered by a large blue callout box containing Persian text.

در دنیای بورس و بازارهای مالی، دوتا مقدار وجود دارد. مقدار مطلق و مقدار نسبی (Relative Value). گاهاً بدلایل مختلف از مقدار نسبی استفاده می‌کنند که این باعث میشه دیتاست در ظاهر پراکندگی‌های زیادی در بخش‌هاییش بوجود بیاد. اما خیلی ساده از یک تبدیل \log (اسلاید بعد) استفاده می‌کنیم و این میزان زیادی از پراکندگی‌ها رو کاهش میده. پس دوتا متراینج وجود داشت. مترا نسبی و مترا مطلق.

10:00 11:00 12:00 13:00 14:00 15:00 16:00 17:00 18:00 19:00 20:00

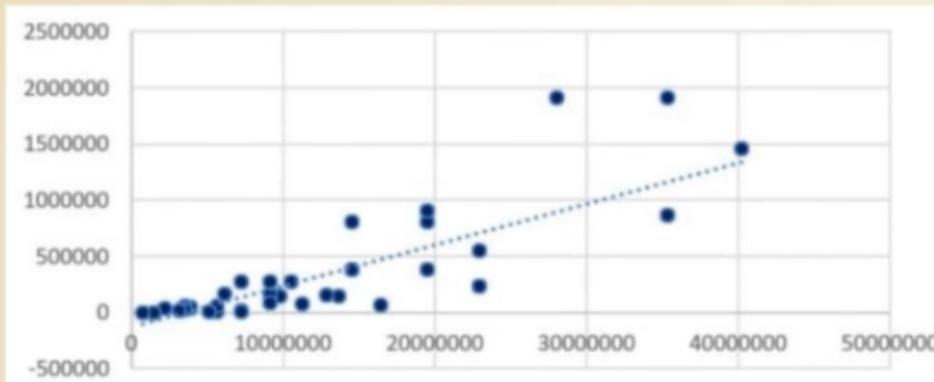
141 142 143 144 145 146

Previous close 142.83

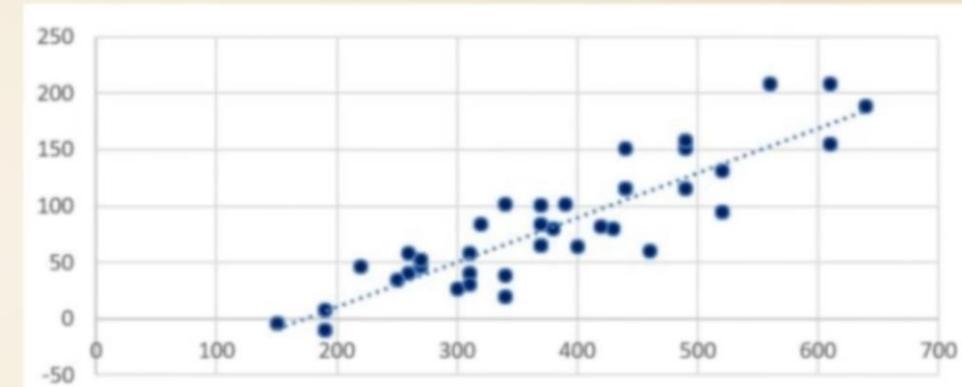
Data Science Course, By Dr. Mohammad Fozouni. <https://www.m-fozouni.ir/>

LOGARITHMS

Financial data

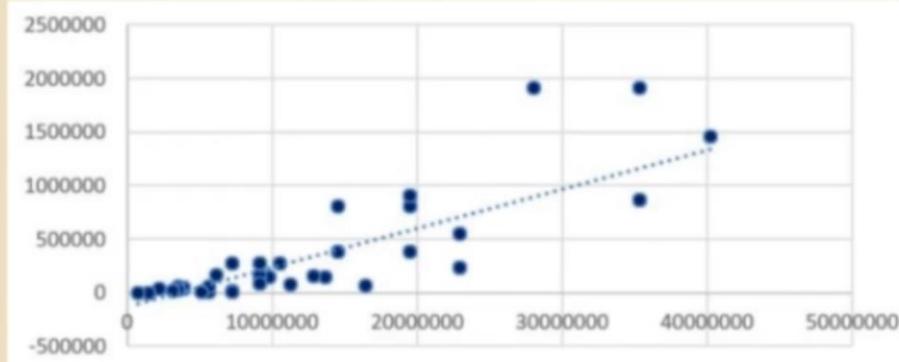


Log transformed financial data

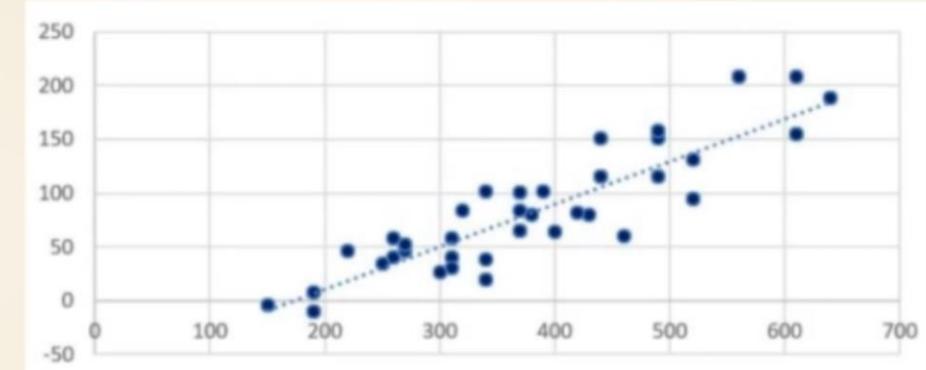


LOGARITHMS

Financial data



Log transformed financial data



Advantages:

Faster computation

Lower order of magnitude

Clearer relationships

Homogeneous variance

STANDARDIZATION

FEATURE SCALING

NORMALIZATION



STANDARDIZATION

FEATURE SCALING

The process of transforming data
into a standard scale

STANDARDIZATION

FEATURE SCALING

$$\text{standardized variable} = \frac{x - \mu}{\sigma}$$

original variable

mean of original variable

standard deviation of original variable

AACD	68.754	0.00%	0.00%	0.304	2.03%	16,310	0.00
ABCO	4.344	0.304	0.87%	0.304	0.87%	NA	0.00
ABCO	55.543	0.130	0.60%	0.304	2.03%	NA	0.00
ABCO	63.446	0.304	0.89%	0.304	2.03%	NA	0.00
ABCO	32.304	0.304	2.03%	0.304	2.03%	NA	0.00
ABCO	12.324	0.304	0.89%	0.304	2.03%	NA	0.00
ABCO	38.785	0.304	0.89%	0.304	2.03%	NA	0.00
ABCO	21.734	0.304	2.03%	0.304	2.03%	NA	0.00

STANDARDIZATION

FEATURE SCALING

day

Exchange rate

Daily trading volume

1

1.3

110,000

STANDARDIZATION

FEATURE SCALING

	Open	High	Low	Close	Change	% Change	Volume	MA	RSI
ABC1	90.754	90.98	90.70	90.84	0.084	0.09%	16,310	0.00	
ABC2	4.344	4.364	4.327	4.364	0.024	0.57%	N/A	0.00	
ABC3	55.543	55.130	54.867	55.304	0.161	0.29%	N/A	0.00	
ABC4	61.446	61.084	60.920	61.304	0.158	0.25%	N/A	0.00	
ABC5	32.304	32.304	32.203	32.304	0.004	0.01%	N/A	0.00	
ABC6	12.324	12.304	12.280	12.304	0.004	0.03%	N/A	0.00	
ABC7	32.795	32.984	32.897	32.804	0.004	0.03%	N/A	0.00	
ABC8	81.734	81.304	80.237	81.304	0.004	0.03%	N/A	0.00	

day

Exchange rate

Daily trading volume

1

1.3

110,000

2

1.34

98,700

ABC	66.754	0.384	2.83%	0.384	2.83%	16,310	0.00
ABCD	4.344	0.384	0.87%	0.384	0.87%	NA	0.00
ABCO	55.543	0.384	0.80%	0.384	2.83%	NA	0.00
ABCO	63.446	0.384	2.83%	0.384	2.83%	NA	0.00
ABCO	32.394	0.384	2.03%	0.384	2.03%	NA	0.00
ABCO	12.324	0.384	2.03%	0.384	2.03%	NA	0.00
ABCO	32.795	0.384	2.83%	0.384	2.83%	NA	0.00
ABCO	21.734	0.384	2.83%	0.384	2.83%	NA	0.00

STANDARDIZATION

FEATURE SCALING

day	Exchange rate	Daily trading volume
1	1.3	110,000
2	1.34	98,700
3	1.25	135,000

ABC	80.754	0.00%	0.00%	0.304	2.03%	16,310	0.00
ABCD	4.344	0.004	0.87%	0.304	0.97%	NA	0.00
ABCD	55.543	0.130	0.60%	0.304	2.03%	NA	0.00
ABCD	83.446	0.004	0.00%	0.304	2.03%	NA	0.00
ABCD	32.304	0.304	2.03%	0.304	2.03%	NA	0.00
ABCD	12.324	0.004	0.00%	0.304	2.03%	NA	0.00
ABCD	32.765	0.004	0.00%	0.304	2.03%	NA	0.00
ABCD	21.734	0.304	2.03%	0.304	2.03%	NA	0.00

STANDARDIZATION

FEATURE SCALING

day

Exchange rate

Daily trading volume

1

1.3

110,000

2

1.34

98,700

3

1.25

135,000

mean: 1.3

std: 0.045

	66.754	0.00%	0.00%	0.304	2.03%	16,310	0.00
ABCD	4.344	0.00%	0.00%	0.304	0.07	NA	0.00
ABCD	55.543	0.13%	0.00%	0.304	2.03%	NA	0.00
ABCD	63.446	0.00%	0.00%	0.304	2.03%	NA	0.00
ABCD	33.304	0.00%	0.00%	0.304	2.03%	NA	0.00
ABCD	12.324	0.00%	0.00%	0.304	2.03%	NA	0.00
ABCD	33.795	0.00%	0.00%	0.304	2.03%	NA	0.00
ABCD	21.734	0.00%	0.00%	0.304	2.03%	NA	0.00

STANDARDIZATION

FEATURE SCALING

$$\frac{x - \mu}{\sigma}$$

day

Exchange rate

Daily trading volume

1

0.07

(1.3)

110,000

2

0.96

(1.34)

98,700

3

-1.03

(1.25)

135,000

	Open	High	Low	Close	Change	% Change	Volume	Market Cap
AAPL	167.54	167.65	167.54	167.65	0.11	0.06%	16,310	\$190B
AMZN	4,344	4,364	4,335	4,364	0.87	0.02%	N/A	\$190B
GOOGL	95,543	95,730	95,543	95,730	1.67	0.02%	N/A	\$190B
MSFT	214.46	214.66	214.46	214.66	0.20	0.09%	N/A	\$190B
TXN	33.344	33.364	33.325	33.364	0.023	0.07%	N/A	\$190B
FB	12.324	12.368	12.324	12.368	0.044	0.35%	N/A	\$190B
INTC	33.765	33.868	33.765	33.868	0.103	0.30%	N/A	\$190B
ADBE	25.734	25.864	25.725	25.864	0.131	0.52%	N/A	\$190B

STANDARDIZATION

FEATURE SCALING

$$\frac{x - \mu}{\sigma}$$

day

Exchange rate

Daily trading volume

1

0.07

(1.3)

-0.25

(110,000)

2

0.96

(1.34)

-0.85

(98,700)

3

-1.03

(1.25)

1.1

(135,000)

ABC	80.754	0.384	0.03%	0.384	2.03%	NA	16,310	0.00
ABCD	4.344	0.384	0.87%	0.384	0.87	NA	0.00	0.00
ABCD	15.043	0.384	0.69%	0.384	2.03%	NA	0.00	0.00
ABCD	63.446	0.384	2.03%	0.384	2.03%	NA	0.00	0.00
ABCD	32.394	0.384	2.03%	0.384	2.03%	NA	0.00	0.00
ABCD	12.324	0.384	0.03%	0.384	2.03%	NA	0.00	0.00
ABCD	33.745	0.384	0.03%	0.384	2.03%	NA	0.00	0.00
ABCD	21.734	0.384	2.03%	0.384	2.03%	NA	0.00	0.00

STANDARDIZATION

FEATURE SCALING

day

Exchange rate

Daily trading volume

1	0.07	(1.3)	-0.25	(110,000)
2	0.96	(1.34)	-0.85	(98,700)
3	-1.03	(1.25)	1.1	(135,000)

We have forced the features to appear similar

NORMALIZATION

5	100
120	-10
-1	0

normalize
→
using
L2-norm

0.042	0.995
0.999	-0.1
-0.008	0

$$\begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \rightarrow \left[\begin{array}{c|cc} -0.70710678 & 0 & 0.70710678 \\ 0 & 0.4472136 & 0.89442719 \end{array} \right]$$

$$L^2 - Norm (-1,0,1) = \sqrt{(-1)^2 + 0^2 + 1^2} = \sqrt{2}$$

$$\left(\frac{-1}{\sqrt{2}}, \frac{0}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) = (-0.70710678, 0, 0.70710678)$$

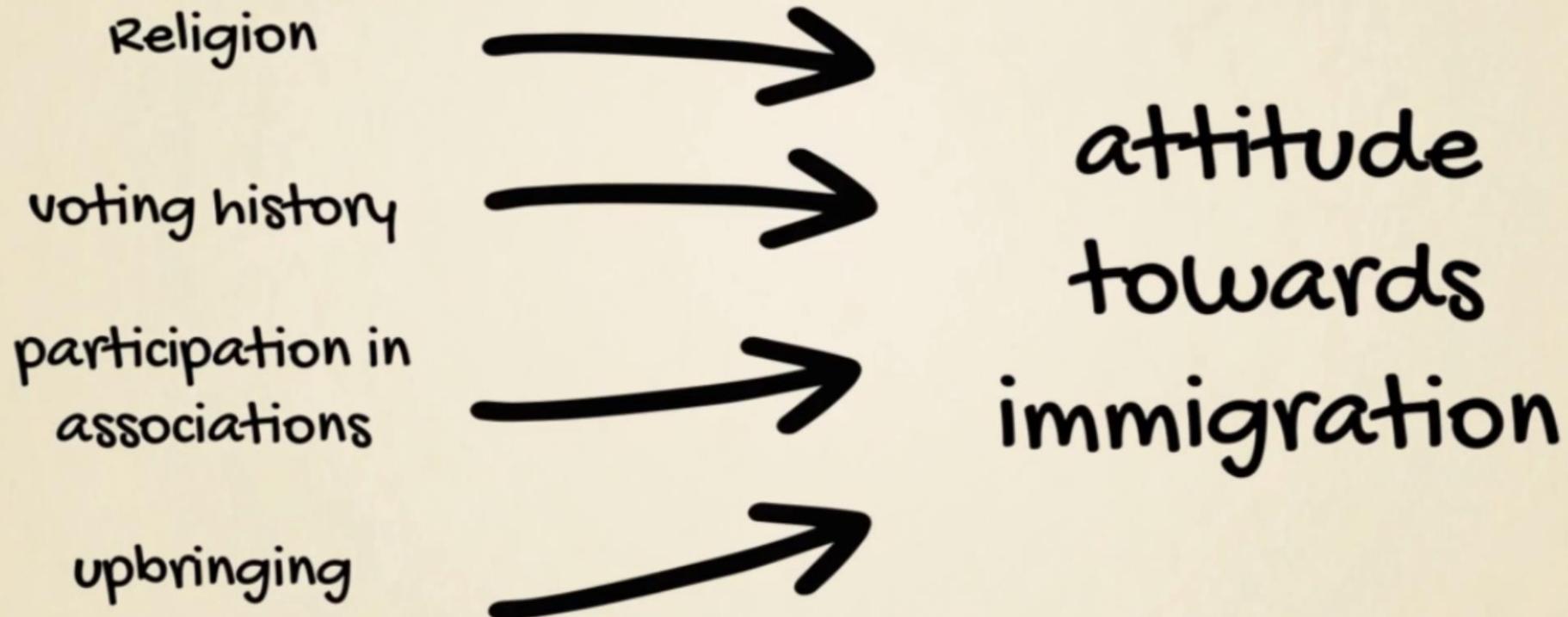
PCA

/principal components analysis/

Dimension reduction technique used to combine several variables into a bigger (latent) variable

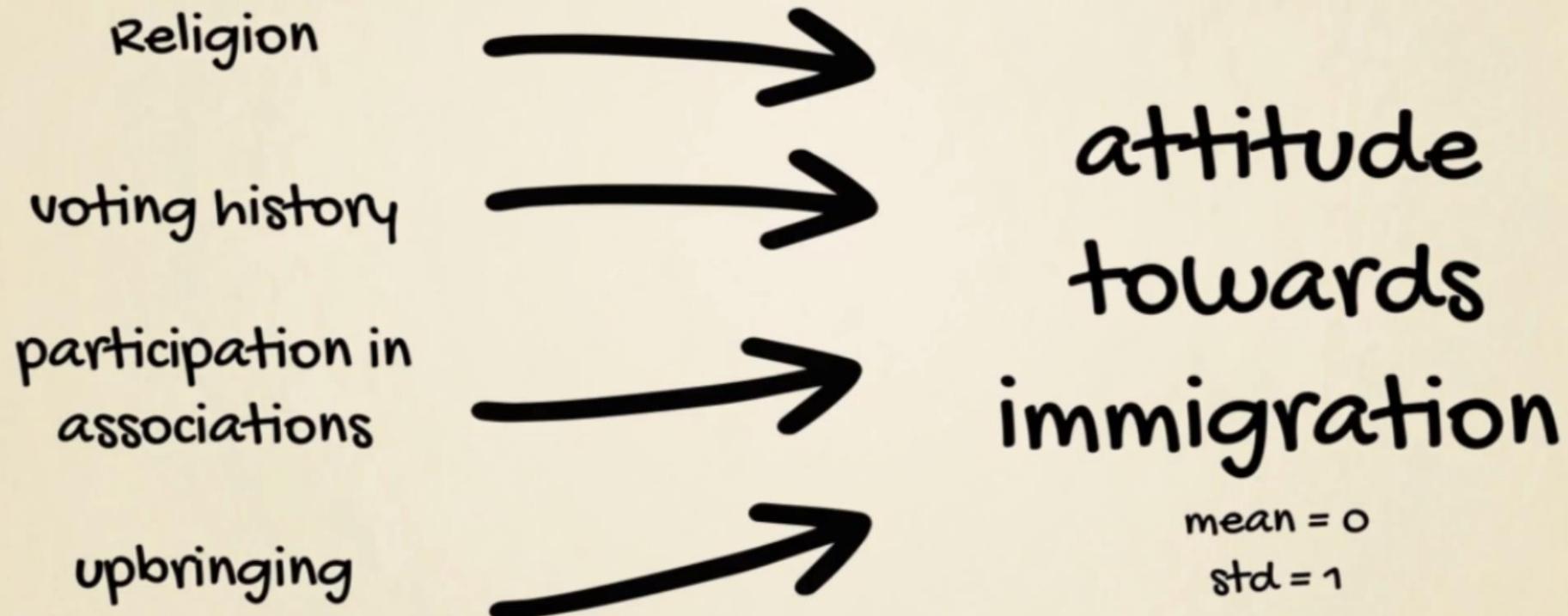
PCA

/principal components analysis/



PCA

/principal components analysis/



WHITENING

معمولًاً بعد از PCA و به منظور حذف
دادههای وابسته بکار گرفته می‌شود

attitude
towards
immigration



Uncorrelated
'attitude towards
immigration'

A whitening transformation or spherling transformation is a linear transformation that transforms a vector of random variables with a known covariance matrix into a set of new variables whose covariance is the identity matrix, meaning that they are uncorrelated and each have variance 1.

توضیح دقیق تمامی این
موارد بسیار زمان بر است و
عموماً هر کدام از اینها به
مسئله‌ی در دست وابسته
هست

اما مهمترین موضوع برای ما
در ادامه دوره، بیشتر

Standardization

من باشد

STANDARDIZATION

NORMALIZATION

PCA

WHITENING

...

CATEGORICAL DATA

Categories or groups (non-numerical data)



CATEGORICAL DATA



Categories or groups (non-numerical data)

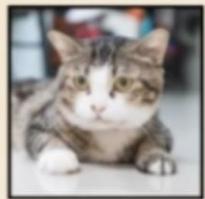


numbers →



→ numbers

CATEGORICAL DATA



Categories or groups (non-numerical data)



numbers



numbers

CATEGORICAL DATA



Categories or groups (non-numerical data)



CATEGORICAL DATA



CATEGORICAL DATA



1



2



3

CATEGORICAL DATA



We have assumed order where there isn't

1



2



3

CATEGORICAL DATA

How to encode categories in a way useful for ML

CATEGORICAL DATA

How to encode categories in a way useful for ML



One-hot encoding



Binary encoding

BINARY ENCODING

	ordinal	binary
	1	0 1
	2	1 0
	3	1 1

BINARY ENCODING

	ordinal	var 1	var 2
	1	0	1
	2	1	0
	3	1	1

BINARY ENCODING

	var1	var2
	0	1
	1	0
	1	1

By splitting the numbers in two variables we have removed the order

...but...

there are some implied correlation between them

BINARY ENCODING

	var1	var2	
	0	1	Bread is the opposite of yogurt
	1	0	Whatever is bread is not yogurt and vice versa
	1	1	

BINARY ENCODING

	var1	var2	
	0	1	Muffins is the opposite of yogurt
	1	0	Whatever is muffins is not yogurt and vice versa
	1	1	

ONE-HOT ENCODING

	bread	yogurt	muffins
	✓		
		✓	
			✓

ONE-HOT ENCODING

	bread	yogurt	muffins
	1	0	0
	0	1	0
	0	0	1

ONE-HOT ENCODING

	bread	yogurt	muffins
bread	1	0	0
yogurt	0	1	0
muffins	0	0	1

uncorrelated

unequivocal
unambiguous

CAT



DOG



HORSE



$$t_2 = [0, \text{NOT CAT} \times]$$

$$[0, \text{NOT DOG} \times]$$

$$[1, \text{HORSE} \checkmark]$$

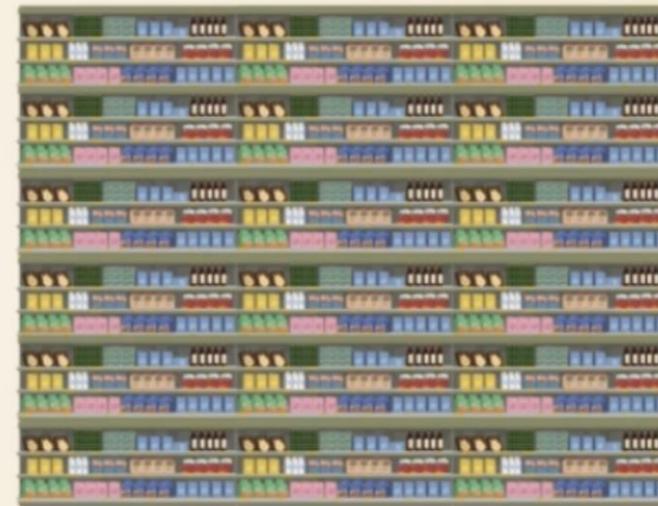
ONE-HOT ENCODING problem?



12,000
products



12,000
columns



12,000
products

	ONE-HOT	BINARY
NEW COLUMNS	12,000	16
	12,000 IN BINARY: 10111011100000	

12,000
products

	ONE-HOT	BINARY
NEW COLUMNS	12,000	16
		BETTER!!!

12,000 IN BINARY: 10111011100000

ONE-HOT



FEW
CATEGORIES

BINARY



MANY
CATEGORIES

در ویدیوی بعدی راجع به دسته‌بندی دیتاست
MNIST
صحبت خواهیم کرد

Stay Tuned and have FUN