LEC 5: Dimensionality Reduction | PCA

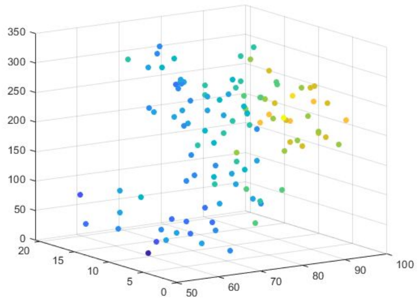
Feb 19, 2020

Quiz

Link:

https://docs.google.com/forms/d/e/1FAIpQLSdEPBv-DPozTx8Q-ZICA1uDpsMK0eJPbDfk8nXI-jfkHmhL4w/viewform?usp=sf_link

Motivation for dimensionality reduction



- Hard to visualize > 3 variables - e.g. When we did the housing prices prediction :(
- Faster computational time for complex things that still captures most signal especially in regression
- Overfitting: need to leave out variables that increase variance without making the model better

Technique: Principal Component Analysis

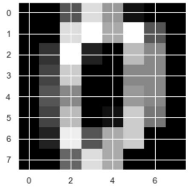
- Not a classification or regression ML technique by itself

1. Used as pre-processing for classification or regression because of the last two reasons
2. Does not assume prior distribution unlike other techniques such as LDA/QDA (assumes normal distribution)
3. Plotting PCA results can show promise of separability before pursuing ML clustering

Detailed paper: https://arxiv.org/pdf/1404.1100.pdf

Reducing complexity on handwritten digits

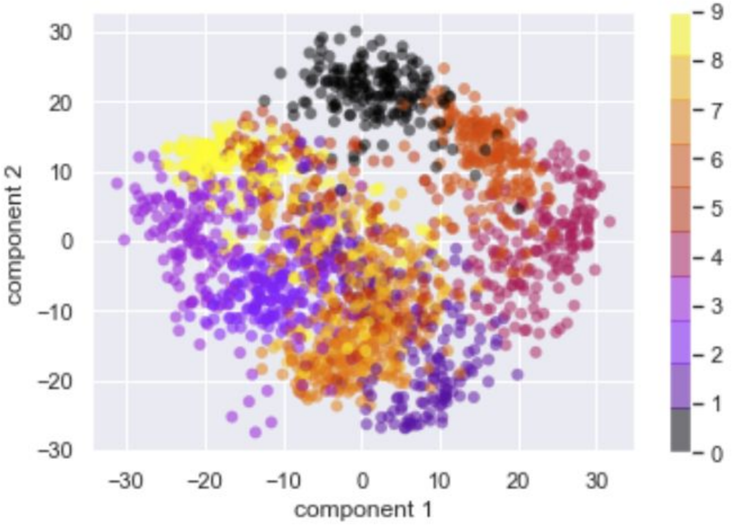
We want to be able to classify handwritten digits e.g. auction house recording



	0	1	2	3	4	5	6	7	8	9	...	54	55	56	57	58	59	60	61	62	63
0	0.0	0.0	5.0	13.0	9.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	6.0	13.0	10.0	0.0	0.0	0.0
1	0.0	0.0	0.0	12.0	13.0	5.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	11.0	16.0	10.0	0.0	0.0
2	0.0	0.0	0.0	4.0	15.0	12.0	0.0	0.0	0.0	0.0	...	5.0	0.0	0.0	0.0	0.0	3.0	11.0	16.0	9.0	0.0
3	0.0	0.0	7.0	15.0	13.0	1.0	0.0	0.0	0.0	8.0	...	9.0	0.0	0.0	0.0	7.0	13.0	13.0	9.0	0.0	0.0
4	0.0	0.0	0.0	1.0	11.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	2.0	16.0	4.0	0.0	0.0

5 rows × 64 columns

	0	1
0	-1.259466	21.274884
1	7.957612	-20.768695
2	6.991923	-9.955989
3	-15.906105	3.332464
4	23.306867	4.269065



Derivation overview

1. Normalization
2. Redundancy, variance, and covariance matrix
3. Finding the best basis
   a. Dual of reducing mean squared error and maximizing variance along axes
4. Change of basis
5. Faster method: Singular value decomposition

Data

n x d matrix of data

If you were to choose which features to plot on a 2x2 plot for regression, which ones would you choose?

Why? What's the heuristic you used?

x1	x2	x3	x4	x5
2.5	9.0	9.0	9.0	3.0
2.0	3.5	10.0	9.0	7.0
2.0	1.0	1.0	9.0	6.5
6.5	11.0	2.0	10.0	1.0
8.0	2.5	1.5	10.0	10.0
7.5	8.5	10.5	10.5	10.0
10.0	4.0	1.0	9.5	1.0
10.5	7.5	11.0	11.0	1.0

Normalization

Some types of normalization

1. Centering
2. Standardization
3. Min-Max

$$z = \frac{\boxed{x_i - \mu}}{\sigma}$$

$$z = \frac{x - min(x)}{[\max(x) - min(x)]}$$

Which to use?

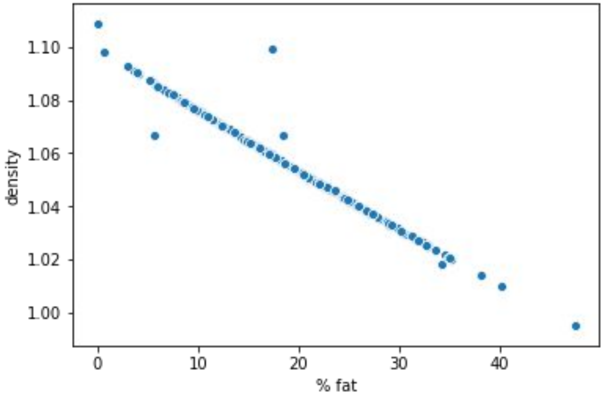
	x1	x2	x3	x4	x5
count	8.00	8.00	8.00	8.00	8.00
mean	0.00	0.00	0.00	0.00	0.00
std	3.52	3.58	4.72	0.75	3.93
min	-4.12	-4.88	-4.75	-0.69	-3.94
25%	-3.75	-2.62	-4.38	-0.69	-3.94
50%	0.88	-0.12	-0.25	-0.19	-0.19
75%	2.38	2.75	4.38	0.44	2.81
max	4.38	5.12	5.25	1.31	5.06

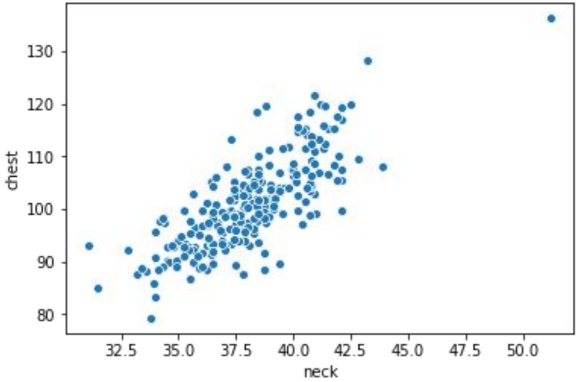
	x1	x2	x3	x4	x5
count	8.00	8.00	8.00	8.00	8.00
mean	0.00	0.00	-0.00	-0.00	0.00
std	1.00	1.00	1.00	1.00	1.00
min	-1.17	-1.36	-1.01	-0.91	-1.00
25%	-1.06	-0.73	-0.93	-0.91	-1.00
50%	0.25	-0.03	-0.05	-0.25	-0.05
75%	0.67	0.77	0.93	0.58	0.72
max	1.24	1.43	1.11	1.74	1.29

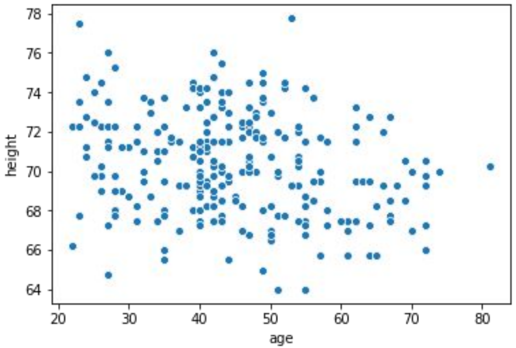
	x1	x2	x3	x4	x5
count	8.00	8.00	8.00	8.00	8.00
mean	0.49	0.49	0.48	0.34	0.44
std	0.41	0.36	0.47	0.38	0.44
min	0.00	0.00	0.00	0.00	0.00
25%	0.04	0.22	0.04	0.00	0.00
50%	0.59	0.48	0.45	0.25	0.42
75%	0.76	0.76	0.91	0.56	0.75
max	1.00	1.00	1.00	1.00	1.00

Idea of variance in multiple dimensions

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

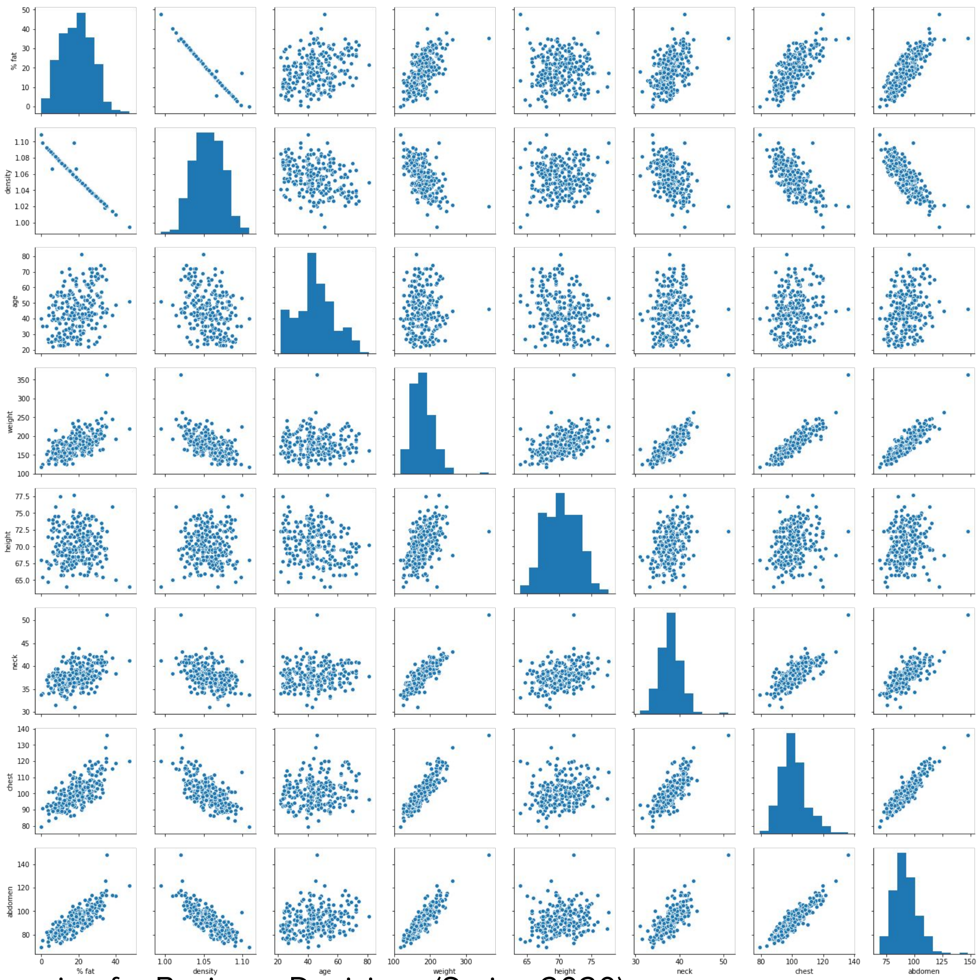






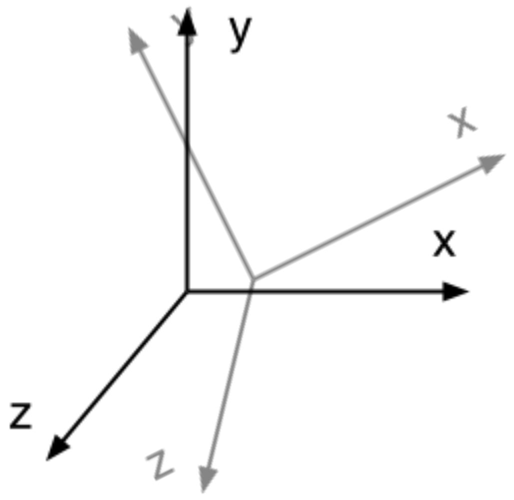
Redundancy

Plot of every pair of vars



Idea of a new coordinate system

Can we redefine the axes to go through the directions of greatest variance in the data?



PCA intuition without matrix decomposition

Mathematically, the transformation is defined by a set of *p*-dimensional vectors of weights or coefficients

$$\mathbf{w}_{(k)} = (w_1, \dots, w_p)_{(k)}$$

that map each row vector (data point!) x_i of data matrix **X** to a new vector of principal component ***scores*** given by

$$t_{k(i)} = \mathbf{x}_{(i)} \cdot \mathbf{w}_{(k)} \qquad \text{for} \qquad i = 1, \dots, n \qquad k = 1, \dots, l$$

In order to maximize variance, the first weight vector $\mathbf{w}_{(1)}$ thus has to satisfy

$$\mathbf{w}_{(1)} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}$$

PCA intuition without matrix decomposition (2)

In order to maximize variance, the first weight vector $\mathbf{w}_{(1)}$ thus has to satisfy

$$\mathbf{w}_{(1)} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w}\right)^2 \right\}$$

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Do you see the equation for variance here? Note that N = 1

Equivalently, writing this in matrix form gives

$$\mathbf{w}_{(1)} = \underset{\|\mathbf{w}\|=1}{\arg\max} \{\|\mathbf{X}\mathbf{w}\|^2\} = \underset{\|\mathbf{w}\|=1}{\arg\max} \{\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}\}$$

Since $\mathbf{w}_{(1)}$ has been defined to be a unit vector, it equivalently also satisfies

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T\mathbf{X}^\mathbf{T}\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \right\}$$

PCA intuition without matrix decomposition (3)

There is a way to solve for w_1 using Singular Value Decomposition (requires understanding of eigenvalues/eigenvectors)

$$\mathbf{w}_{(1)} = \arg\max \left\{ \frac{\mathbf{w}^T \mathbf{X^T X w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

How do we find subsequent principal component axes?

Constraint: They must be orthogonal

Idea: Subtract out the projection of your data onto the $w_{(1)}$ principal component axis. I.e. all data points have the same value on that axis now

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)} \mathbf{w}_{(s)}^{\mathrm{T}}$$

and then repeat to find the next set of weights which extracts the maximum variance from this new data matrix

$$\mathbf{w}_{(k)} = \underset{\|\mathbf{w}\|=1}{\arg\max} \left\{ \|\hat{\mathbf{X}}_k \mathbf{w}\|^2 \right\} = \arg\max \left\{ \frac{\mathbf{w}^T \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \mathbf{w}}{\mathbf{w}^T \mathbf{w}} \right\}$$

https://forms.gle/Uv3YfeGejQqnFXv39

https://tinyurl.com/tw7u8nd