ES114: Probability, Statistics, and Data Visualization

PCA, LLN and CLT

16^{th} Apr 2025

Recap

- Recap

Recap

- Recap
  - Markov and Chebyshev's Inequality

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix
  - Multivariate Normal

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix
  - Multivariate Normal
  - Gaussian Whitening

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix
  - Multivariate Normal
  - Gaussian Whitening
- Principal Component Analysis

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix
  - Multivariate Normal
  - Gaussian Whitening
- Principal Component Analysis
- Law of Large Numbers

Recap

- Recap
  - Markov and Chebyshev's Inequality
  - Random Vectors
  - Covariance Matrix
  - Multivariate Normal
  - Gaussian Whitening
- Principal Component Analysis
- Law of Large Numbers
- Central Limit Theorem

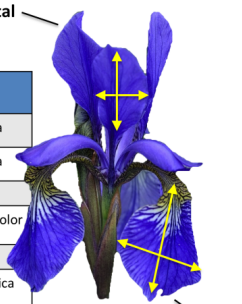
References

- PCA Explained
- PCA Vizualization
- WLLN and CLT
- WLLN and CLT 2

Iris Data

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

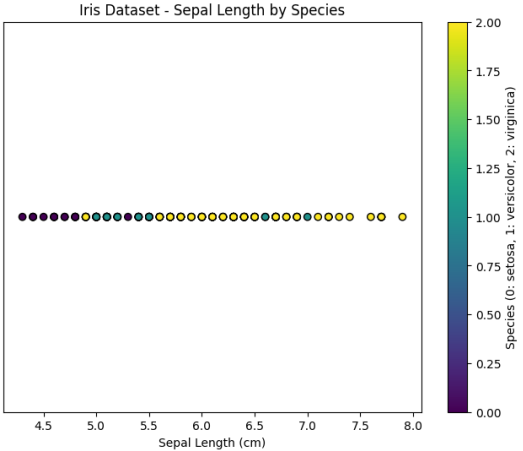
Class labels
(targets)

Iris Data (1D)

Sepal length: X^1 (Random Variable)

Iris Data (1D)

Sepal length: X^1 (Random Variable)

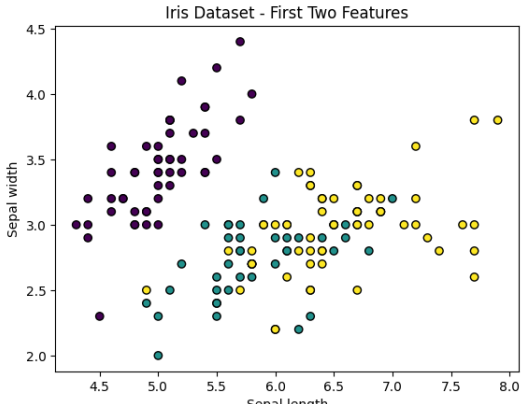


Iris Data (2D)

X^1: Sepal length, X^2: Sepal width

Iris Data (2D)

X^1: Sepal length, X^2: Sepal width

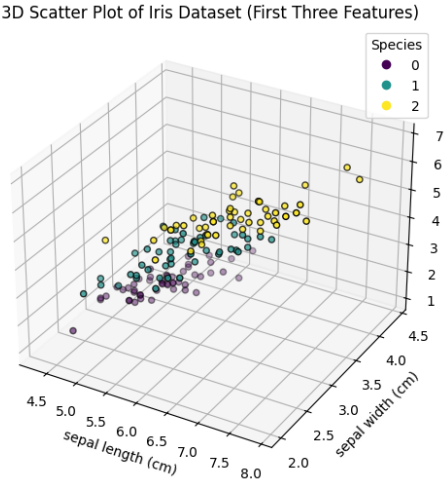


Iris Data (3D)

X^1: Sepal length, X^2: Sepal width, X^3: Petal length

Iris Data (3D)

X^1: Sepal length, X^2: Sepal width, X^3: Petal length



Iris Data (4D)

	Sepal Length	Sepal Width	Petal Length	Petal Width
	X^1	X^2	X^3	X^4
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4

Iris Data (4D)

	Sepal Length	Sepal Width	Petal Length	Petal Width
	X^1	X^2	X^3	X^4
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
5	5.4	3.9	1.7	0.4

How would you visualize it in 2D?

Images

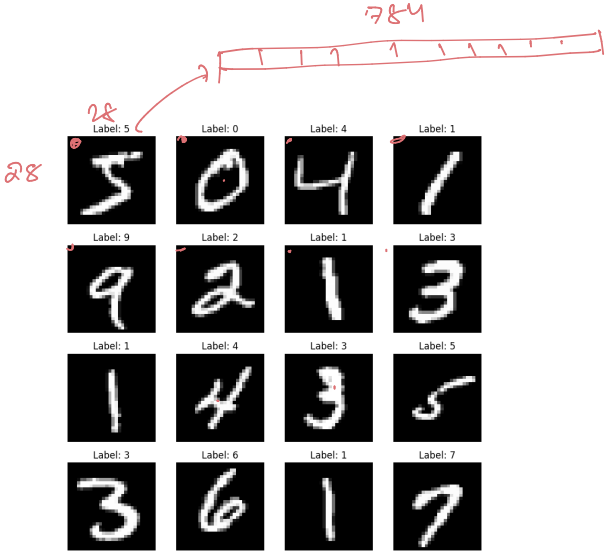

784
28
28
Label: 5
Label: 0
Label: 4
Label: 1
Label: 9
Label: 2
Label: 1
Label: 3
Label: 1
Label: 4
Label: 3
Label: 5
Label: 3
Label: 6
Label: 1
Label: 7

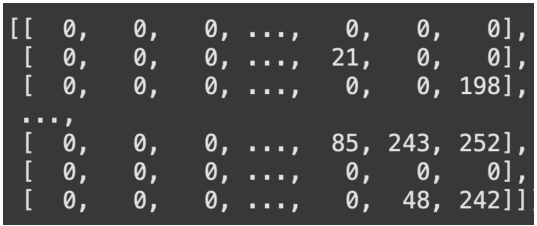
Image Data as a Random Vector

$X =$

x^1

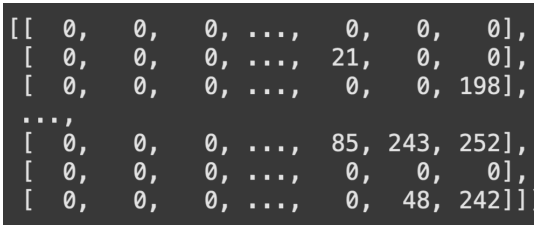
```
[[  0,   0,   0, ...,   0,   0,   0],
 [  0,   0,   0, ...,  21,   0,   0],
 [  0,   0,   0, ...,   0,   0, 198],
 ...,
 [  0,   0,   0, ...,  85, 243, 252],
 [  0,   0,   0, ...,   0,   0,   0],
 [  0,   0,   0, ...,   0,  48, 242]]
```

Image Data as a Random Vector



- Are all the features useful?

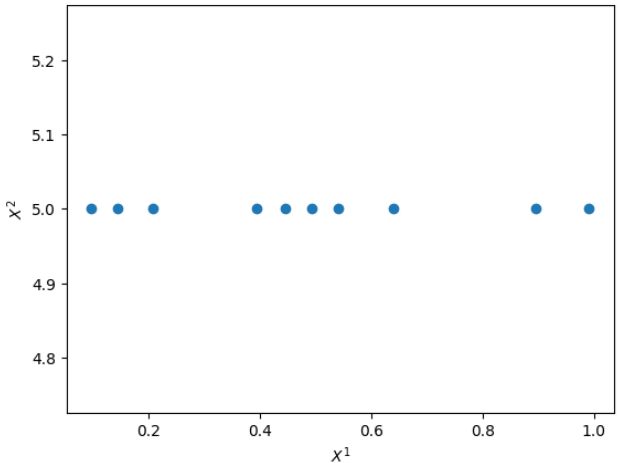
Image Data as a Random Vector



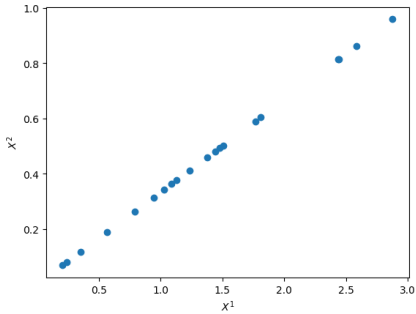
- Are all the features useful?
- How to capture important information?

Data

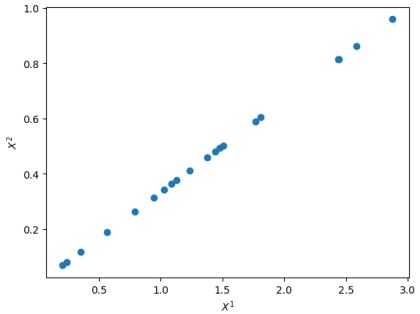
X^1	X^2
0.41709128	5.0
0.04221766	5.0
0.38424179	5.0
0.90469106	5.0
0.60924091	5.0
0.58330889	5.0
0.56814491	5.0
0.68974537	5.0
0.23745621	5.0
0.70578727	5.0

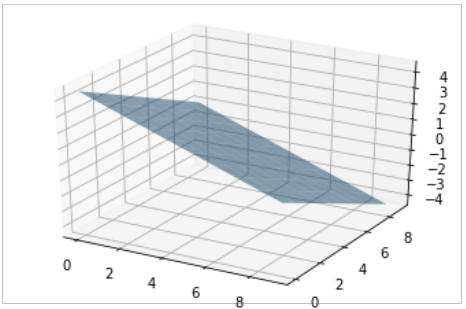


Data



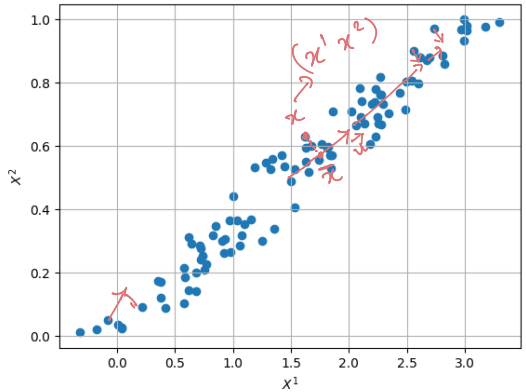
Data



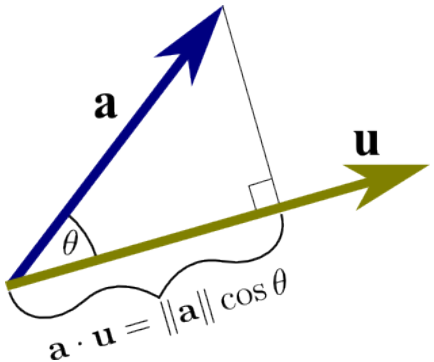


“Manifold”

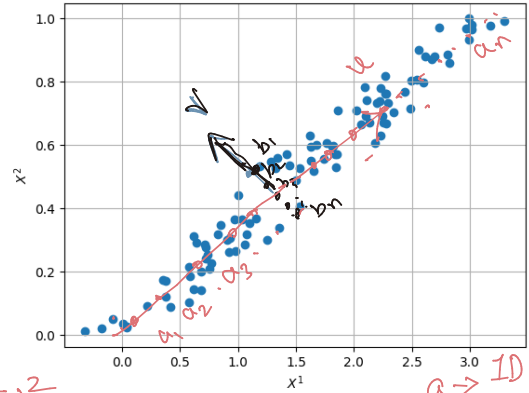
Data

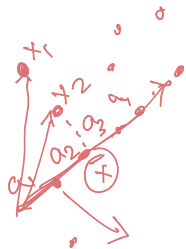


Projecting Data



Projecting Data





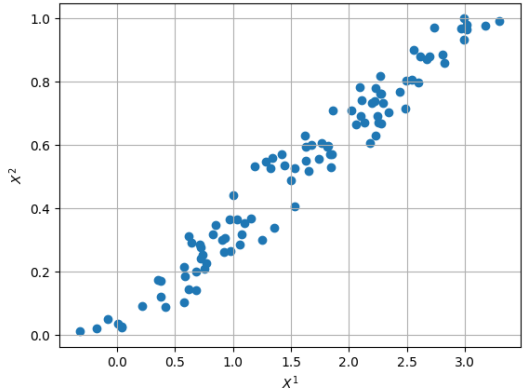
$$\frac{1}{N}\sum_{i=1}^{n}(a_i - \bar{a})^2$$

$$\frac{1}{N}\|X \cdot u - \bar{X} \cdot u\|^2$$

$a \rightarrow 1D \rightarrow [a_1 \; a_2 \; a_3 \; a_4 \; \cdots]$

$$2D \left\{ \begin{matrix} x_{11} & x_{21} & \cdots \\ x_{12} & x_{22} & \end{matrix} \right.$$

Projecting Data



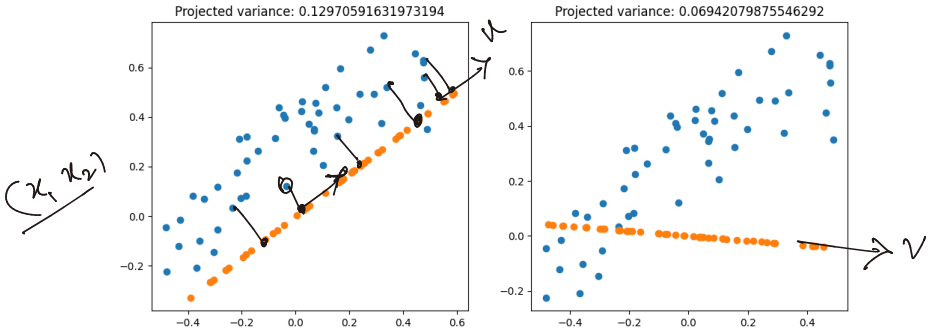
$$\|X \cdot v - \bar{X}.v\|^2 = \|(X - \bar{X}) \cdot v\|^2$$

Principal Component

Find the direction v which maximizes variance in the data?

Principal Component

Find the direction v which maximizes variance in the data?



Principal Components

Principal components are the directions which capture the maximum variance of the data

Principal Components

Principal components are the directions which capture the maximum variance of the data

- Data is d-dimensional

Principal Components

Principal components are the directions which capture the maximum variance of the data

- Data is d-dimensional
- Will need at most d directions to capture entire variance

Principal Components

Principal components are the directions which capture the maximum variance of the data

- Data is d-dimensional
- Will need at most d directions to capture entire variance
- Top k directions that preserve the maximum variance are the principal components

Covariance Matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{bmatrix}$$

Covariance Matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{bmatrix}$$

$$Cov(x_1, x_2) = \frac{1}{5} \sum_{i=1}^{5} (x_{i1} - \mu_1)(x_{i2} - \mu_2)$$

Covariance Matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{bmatrix}$$

$$Cov(x_1, x_2) = \frac{1}{5} \sum_{i=1}^{5} (x_{i1} - \mu_1)(x_{i2} - \mu_2)$$

$$\frac{1}{N}(X - \mu)^T (X - \mu) = \frac{1}{N} \bar{X}^T \bar{X}$$

$$= \frac{1}{N} \begin{bmatrix} \bar{x}_1^T \bar{x}_1 & \bar{x}_1^T \bar{x}_2 & \bar{x}_1^T \bar{x}_3 & \bar{x}_1^T \bar{x}_4 \\ & & \vdots & \\ & & \vdots & \\ \bar{x}_5^T \bar{x}_1 & \bar{x}_5^T \bar{x}_2 & \bar{x}_5^T \bar{x}_3 & \bar{x}_5^T \bar{x}_4 \end{bmatrix}$$

Covariance Matrix

Assume $\bar{X}$, $n \times d$ is mean normalized

- What can you say about $\bar{X}^T \bar{X}$?

Covariance Matrix

Assume $\bar{X}$, $n \times d$ is mean normalized

- What can you say about $\bar{X}^T \bar{X}$?
  - Square ($d \times d$)

Covariance Matrix

Assume $\bar{X}$, $n \times d$ is mean normalized

- What can you say about $\bar{X}^T \bar{X}$?
  - Square ($d \times d$)
  - Symmetric $(\bar{X}^T \bar{X})^T = \bar{X}^T (\bar{X}^T)^T = \bar{X}^T \bar{X}$

Covariance Matrix

Assume $\bar{X}$, $n \times d$ is mean normalized

- What can you say about $\bar{X}^T\bar{X}$?
  - Square ($d \times d$)
  - Symmetric $(\bar{X}^T\bar{X})^T = \bar{X}^T(\bar{X}^T)^T = \bar{X}^T\bar{X}$
  - $\bar{X}^T\bar{X}$ is positive semi-definite so non-negative eigen values

Covariance Matrix

Assume $\bar{X}$, $n \times d$ is mean normalized

- What can you say about $\bar{X}^T\bar{X}$?
  - Square ($d \times d$)
  - Symmetric $(\bar{X}^T\bar{X})^T = \bar{X}^T(\bar{X}^T)^T = \bar{X}^T\bar{X}$
  - $\bar{X}^T\bar{X}$ is positive semi-definite so non-negative eigen values
  - Since symmetric one can chose eigen vectors to be orthonormal

$$V = \begin{bmatrix} | & | & & | \\ v_1 & v_2 & \dots & v_d \\ | & | & & | \end{bmatrix}$$

Principal Component Analysis

The direction which preserves the most variance is,

- Eigen vector of covariance matrix with largest eigen value

$$\bar{X}^T \bar{X} v_1 = \lambda_1 v_1$$

Principal Component Analysis

The direction which preserves the most variance is,

- Eigen vector of covariance matrix with largest eigen value

$$\bar{X}^T \bar{X} v_1 = \lambda_1 v_1$$

- Variance preserved is equal to the eigen value

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_d$$

$$v_1 \quad v_2 \quad v_3 \quad \cdots \quad v_d$$

Principal Component Analysis

The direction which preserves the most variance is,

- Eigen vector of covariance matrix with largest eigen value

$$\bar{X}^T \bar{X} v_1 = \lambda_1 v_1$$

- Variance preserved is equal to the eigen value
- $d \times d$ matrix can have atmost d eigen vectors

Principal Component Analysis

The direction which preserves the most variance is,

- Eigen vector of covariance matrix with largest eigen value

$$\bar{X}^T \bar{X} v_1 = \lambda_1 v_1$$

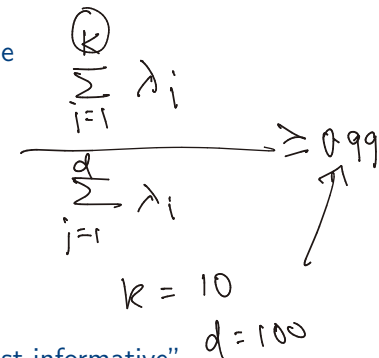
- Variance preserved is equal to the eigen value
- $d \times d$ matrix can have atmost d eigen vectors
- Total variance is $\sum_{i=1}^{d} \lambda_i$

Principal Component Analysis

The direction which preserves the most variance is,

- Eigen vector of covariance matrix with largest eigen value

$$\bar{X}^T \bar{X} v_1 = \lambda_1 v_1$$



- Variance preserved is equal to the eigen value
- $d \times d$ matrix can have atmost d eigen vectors
- Total variance is $\sum_{i=1}^{d} \lambda_i$
- Project data onto the top k eigen vectors yields the "most informative" k-dimensional data

PCA Algorithm

- Data Compression
- Data Visualization
- Data Denoising

PCA Algorithm

Algorithm Principal Component Analysis (PCA)

Require: Data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d features, number of components k

Ensure: Top k principal components

PCA Algorithm

Algorithm Principal Component Analysis (PCA)

Require: Data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d features, number of components k

Ensure: Top k principal components

1: **Center the data:**

$$\text{Let } \mu = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \overline{X} = X - \mu$$

PCA Algorithm

Algorithm Principal Component Analysis (PCA)

Require: Data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d features, number of components k

Ensure: Top k principal components

1: **Center the data:**

$$\text{Let } \mu = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \tilde{X} = X - \mu$$

2: **Compute the covariance matrix:**

$$C = \frac{1}{n} \tilde{X}^\top \tilde{X}$$

Algorithm Principal Component Analysis (PCA) - Continued

3: **Compute the eigenvectors and eigenvalues of** C:

$$Cv_i = \lambda_i v_i, \quad i = 1, \ldots, d$$

Algorithm Principal Component Analysis (PCA) - Continued

3: **Compute the eigenvectors and eigenvalues of** C:

$$Cv_i = \lambda_i v_i, \quad i = 1, \ldots, d$$

4: **Sort the eigenvectors by decreasing eigenvalues:**

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

Algorithm Principal Component Analysis (PCA) - Continued

3: **Compute the eigenvectors and eigenvalues of** C:

$$Cv_i = \lambda_i v_i, \quad i = 1, \ldots, d$$

4: **Sort the eigenvectors by decreasing eigenvalues:**

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$$

5: **Select the top** k **eigenvectors:**

$$W = [v_1, v_2, \ldots, v_k]$$

Algorithm Principal Component Analysis (PCA) - Continued

6: **Project the data onto the top k components:**

$$Z = \tilde{X}W$$

PCA Algorithm

Algorithm Principal Component Analysis (PCA) - Continued

6: **Project the data onto the top k components:**

$$Z = \tilde{X}W$$

$n \times d$ $\quad$ $d \times k$ $\quad$ $n \times k$

$k \leq\!< d$

7: **return** Projected data Z and components W

PCA Algorithm

Algorithm Principal Component Analysis (PCA) - Continued

6: **Project the data onto the top k components:**

$$Z = \tilde{X}W$$

7: **return** Projected data Z and components W

Viz

Implementation of PCA

- Notebook
- Blog

PCA - Failure

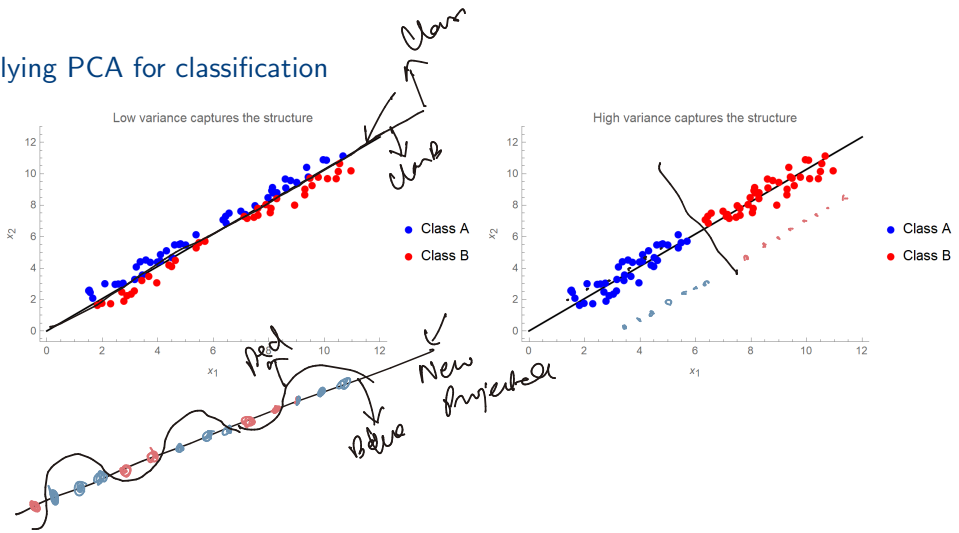
PCA may not always work!

PCA

Applying PCA for classification

PCA

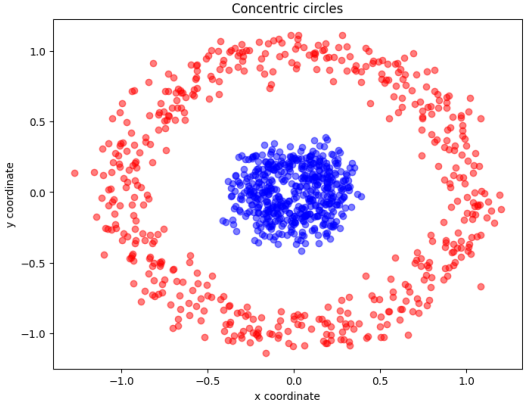
Applying PCA for classification



PCA

Applying PCA for Concentric Circles

Applying PCA for Concentric Circles

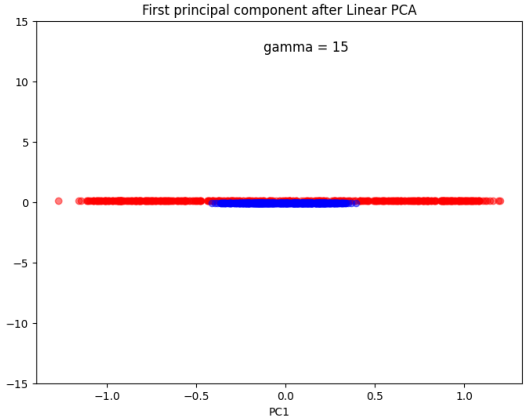


PCA

Applying PCA for Concentric Circles

PCA

Applying PCA for Concentric Circles



Mean

Given n samples $x_1, x_2, \ldots, x_n$

Mean

Given n samples $x_1, x_2, \ldots, x_n$

Mean

$\frac{\sum x_i}{n}$

Given n samples $x_1, x_2, \ldots, x_n$

- What is the mean?

Mean

Given n samples $x_1, x_2, \ldots, x_n$

- What is the mean?
- What if the samples are Bernoulli draws from $p = 0.5$?

$\frac{\sum x_i}{n}$ 0.5 p=

Mean

Given n samples $x_1, x_2, \ldots, x_n$

- What is the mean?
- What if the samples are Bernoulli draws from $p = 0.5$?
- What if the samples are from exponential distribution $\lambda = 5$?

$$\left(\mu = \frac{1}{5} - \frac{\sum x_i}{n}\right)$$

$n \to$

Sample Mean

Sample Mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Sample Mean

Sample Mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

True Mean: μ

Viz

Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$.

Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$. Define the sample mean as:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

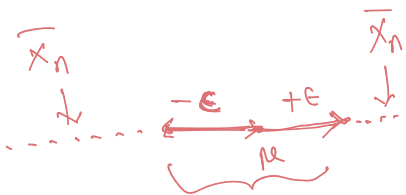
Weak Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean $\mu = \mathbb{E}[X_i]$ and finite variance $\sigma^2 = \mathrm{Var}(X_i)$. Define the sample mean as:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then, for any $\varepsilon > 0$:

$$\lim_{n \to \infty} \mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) = 0$$



Weak Law of Large Numbers

$$E[\bar{X}_n] = \mu$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n}$$

Proof?

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2}$$

$$\leq \frac{\sigma^2}{n\epsilon^2}$$

$$\downarrow$$

$$0$$

$p = 0.5$

$(H\ H\ H\ H\ H\ H \cdots H)$ $n \to \infty$

$(0.5)^{100}$ $1 \to$ $= \bar{X}_n = 1$ 0.5

Weak Law of Large Numbers

Proof?

What is the probability that you get all heads when you toss a fair coin?

Weak Law of Large Numbers

I estimate the sum of n random real numbers by rounding each to the nearest integer, and adding the resulting integers. What is the probability that the total error is at most $\pm\sqrt{n}$?

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, as $n \to \infty$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, as $n \to \infty$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

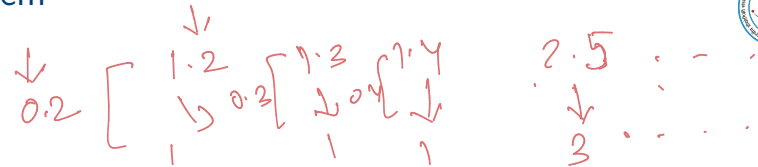
Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables with expected value $\mathbb{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, as $n \to \infty$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Viz

Central Limit Theorem



I estimate the sum of n random real numbers by rounding each to the nearest integer, and adding the resulting integers. What is the probability that the total error is at most $\pm\sqrt{n}$?