Applied Machine Learning

High-Dimensional Data

High-Dimensional Data

- High-dimensional datasets
- Mean
- Covariance

High-Dimensional Data

- Dataset:
- $$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$$
  - N items
  - d features: d-dimensional

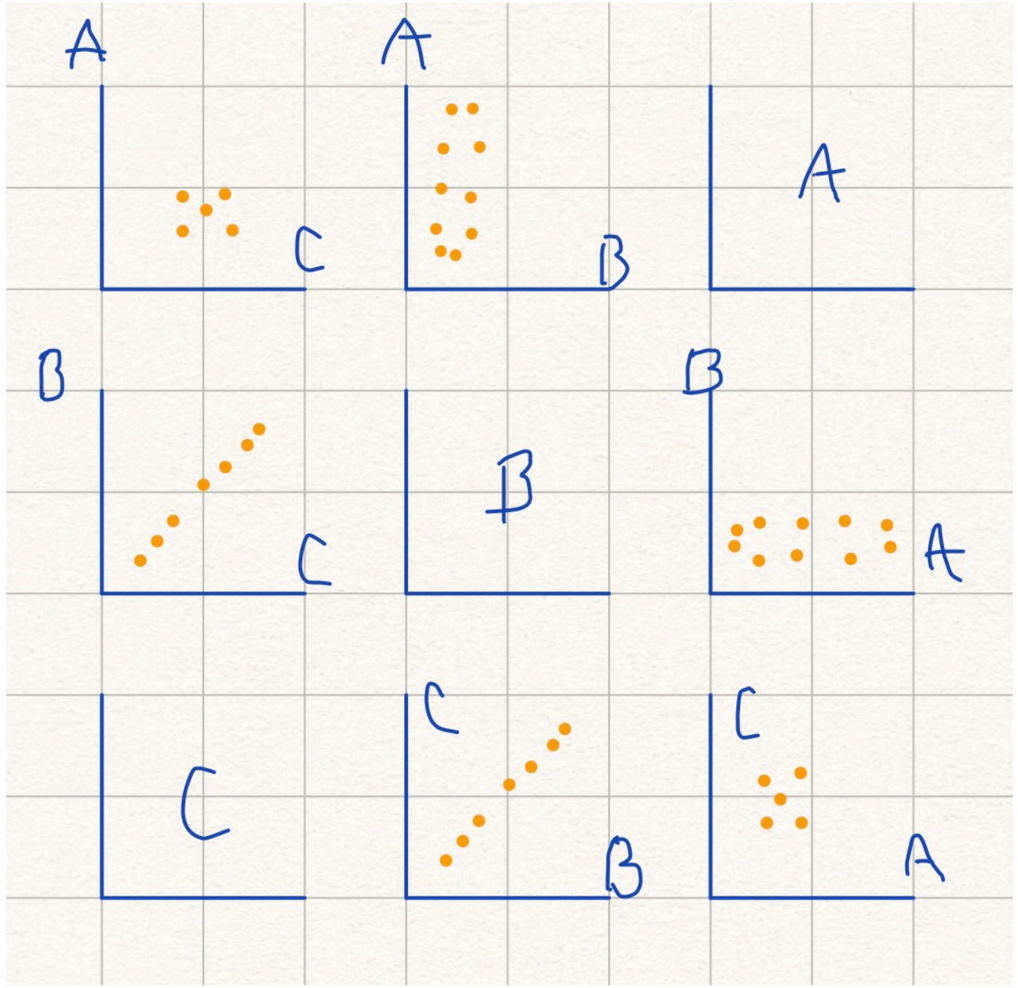
The Curse of Dimensionality

- Data in high-dimensions
  - as d increases, distances increase
  - mean and covariance
    - computational cost grows
    - need more data to get accurate estimations

- $$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$$

Representing High-Dimensional Data

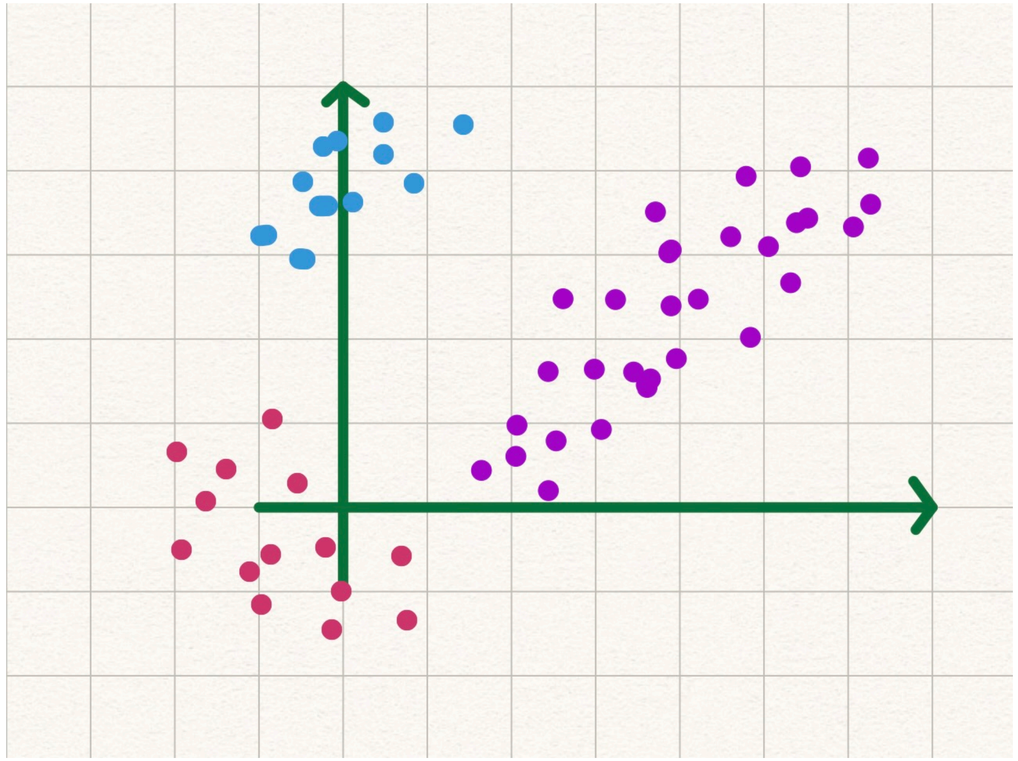
- Scatter plots
  - feature 1 vs feature 2
- Matrix of scatterplots



Representing High-Dimensional Data

- $$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$$

- Mean and Variance
  - $\text{mean}(\{\mathbf{x}\}) = \begin{bmatrix} [& \text{mean}(\mathbf{x})^{(1)} & \text{mean}(\mathbf{x})^{(2)} & \dots & \text{mean}(\mathbf{x})^{(d)} &] \end{bmatrix}$
- Covariance
  - $$\text{Covmat}\{\mathbf{x}\} = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{1,d} \\ \Sigma_{2,1} & \Sigma_{2,2} & \cdots & \Sigma_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{N,1} & \Sigma_{N,2} & \cdots & \Sigma_{N,d} \end{bmatrix}$$



Mean

- 1 dimension
  - $\text{mean}(\{x\}) = \frac{\sum_i x_i}{N}$
- d dimensions
  - $\text{mean}(\{\mathbf{x}\}) = \frac{\sum_i \mathbf{x}_i}{N}$
  - $\text{mean}(\{\mathbf{x} - \text{mean}(\{\mathbf{x}\})\}) = 0$
- $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$
- $\text{mean}(\{\mathbf{x}\}) = \begin{bmatrix} [& \text{mean}(\mathbf{x})^{(1)} & \text{mean}(\mathbf{x})^{(2)} & \dots & \text{mean}(\mathbf{x})^{(d)} &] \end{bmatrix}$

Covariance

- Tendency for a pair of elements in two sets $\{x\}$, $\{y\}$ to be above or below their means
- $\text{cov}(\{x\}, \{y\}) = \frac{\sum_i [x_i - \text{mean}(\{x\})][y_i - \text{mean}(\{y\})]}{N}$
  - Positive: pairs of elements $\{x\}$,$\{y\}$ tend to lie at the same side of their mean
  - Negative: pairs of elements in $\{x\}$, $\{y\}$ tend to be on the opposite side of their mean
- $\{x\}$, $\{y\}$: pairs of features in the dataset
- $$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$$

Covariance Matrix Σ

- $\text{cov}(\{x\}, \{y\}) = \dfrac{\sum_i [x_i - \text{mean}(\{x\})][y_i - \text{mean}(\{y\})]}{N}$
- $\text{Covmat}\{\mathbf{x}\} = \dfrac{\sum_i [\mathbf{x}_i - mean(\{\mathbf{x}\})][\mathbf{x}_i - mean(\{\mathbf{x}\})]^\top}{N}$
  - $[\mathbf{x}_i - \text{mean}(\{\mathbf{x}\})]_{d\times 1}$
  - $\{\mathbf{x}\}, \{\mathbf{y}\}$: pairs of features $\{\mathbf{x}^{(i)}\},\{\mathbf{x}^{(j)}\}$
  - $\Sigma_{i,j} = \text{cov}(\{\mathbf{x}^{(i)}\}, \{\mathbf{x}^{(j)}\})$
- $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} [& x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} &] \\ [& x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} &] \\ & \vdots & \vdots & \ddots & \vdots & \\ [& x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(d)} &] \end{bmatrix}$
- $\text{mean}(\{\mathbf{x}\}) = \begin{bmatrix} [& \text{mean}(\mathbf{x})^{(1)} & \text{mean}(\mathbf{x})^{(2)} & \dots & \text{mean}(\mathbf{x})^{(d)} &] \end{bmatrix}$
- $\text{Covmat}\{\mathbf{x}\} = \Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{1,d} \\ \Sigma_{2,1} & \Sigma_{2,2} & \cdots & \Sigma_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{d,1} & \Sigma_{d,2} & \cdots & \Sigma_{d,d} \end{bmatrix}_{d\times d}$

Covariance Matrix Σ

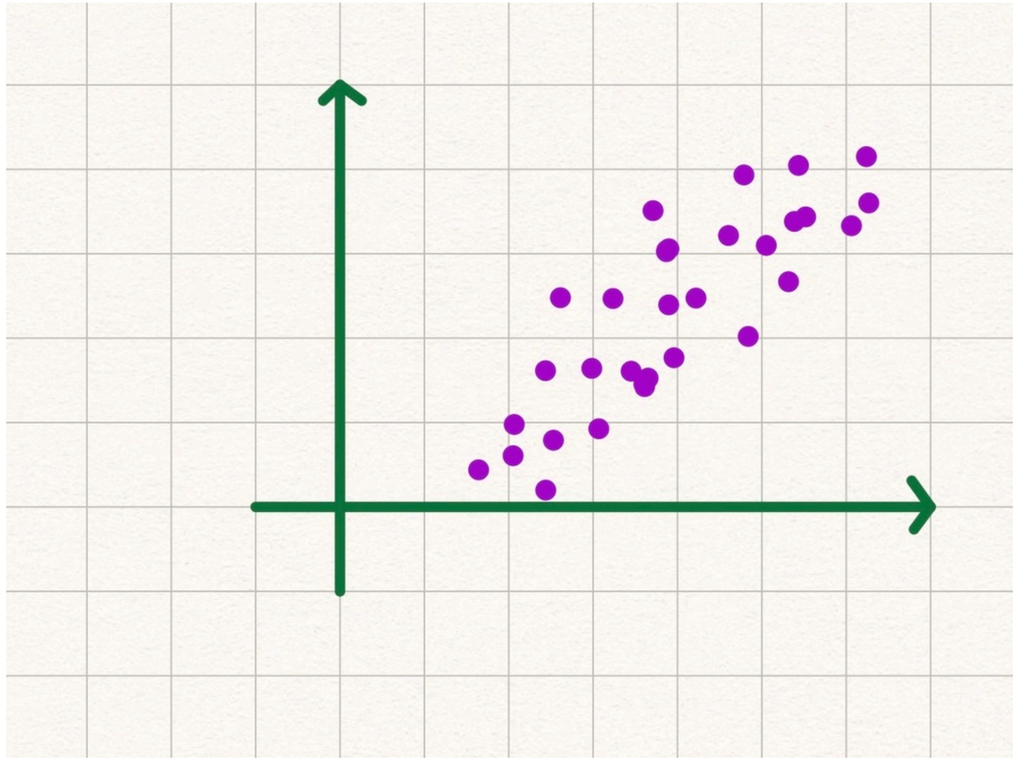
- Covmat$\{\mathbf{x}\} = \frac{\sum_i (\mathbf{x}_i - mean(\{\mathbf{x}\}))(\mathbf{x}_i - mean(\{\mathbf{x}\}))^\top}{N}$
- Positive semidefinite
  - Symmetric $\Sigma = \Sigma^\top$, $\Sigma_{i,j} = \Sigma_{j,i}$
  - all eigenvalues are non-negative
  - Positive definite
    - No vector $\mathbf{a}$ makes $\mathbf{a}^\top[\mathbf{x}_i - mean(\{\mathbf{x}\})] = \mathbf{0}$ for all i
      - any vector $\|\mathbf{v}\| > 0$, will make $\mathbf{v}^\top\Sigma\mathbf{v} > 0$
- Covmat$\{\mathbf{x}\} = \Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \cdots & \Sigma_{1,d} \\ \Sigma_{2,1} & \Sigma_{2,2} & \cdots & \Sigma_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{d,1} & \Sigma_{d,2} & \cdots & \Sigma_{d,d} \end{bmatrix}_{d \times d}$

Covariance Matrix Σ

- $\text{Covmat}\{\mathbf{x}\} = \frac{\sum_i (\mathbf{x}_i - mean(\{\mathbf{x}\}))(\mathbf{x}_i - mean(\{\mathbf{x}\}))^\top}{N}$
- Standard deviation and Variance
  - $\text{std}(x)^2 = \sigma_x^2 = \text{var}(\{x\}) = \text{cov}(\{x\}, \{x\})$
  - $\text{cov}(\{x\}, \{y\}) = \frac{\sum_i [x_i - \text{mean}(\{x\})][y_i - \text{mean}(\{y\})]}{N}$
  - $\text{Covmat}\{\mathbf{x}\} = \Sigma = \begin{bmatrix} \sigma_1^2 & \Sigma_{1,2} & \cdots & \Sigma_{1,d} \\ \Sigma_{2,1} & \sigma_2^2 & \cdots & \Sigma_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{d,1} & \Sigma_{d,2} & \cdots & \sigma_d^2 \end{bmatrix}_{d \times d}$
- Correlation
  - $\text{corr}(\{(x, y)\}) = \frac{\text{COV}(\{x\}, \{y\})}{\sqrt{\text{COV}(\{x\}, \{x\})}\sqrt{\text{COV}(\{y\}, \{y\})}}$
  - $= \frac{\text{COV}(\{x\}, \{y\})}{\sigma_x \sigma_y}$

Mean and Covariance in High-Dimensional Data

- Blobs
  - Mean
  - Covariance



High-Dimensional Data

- High-dimensional datasets
- Mean
- Covariance

Applied Machine Learning

High-Dimensional Data