Chapter - 10 Features

→ Topics already covered in class: categorical, Ordinal & Quantitative

→ Feature Transformation Techniques: Thresholding & Discretization.

Feature transformations that remove scale of quantitative features:

→ Discrete

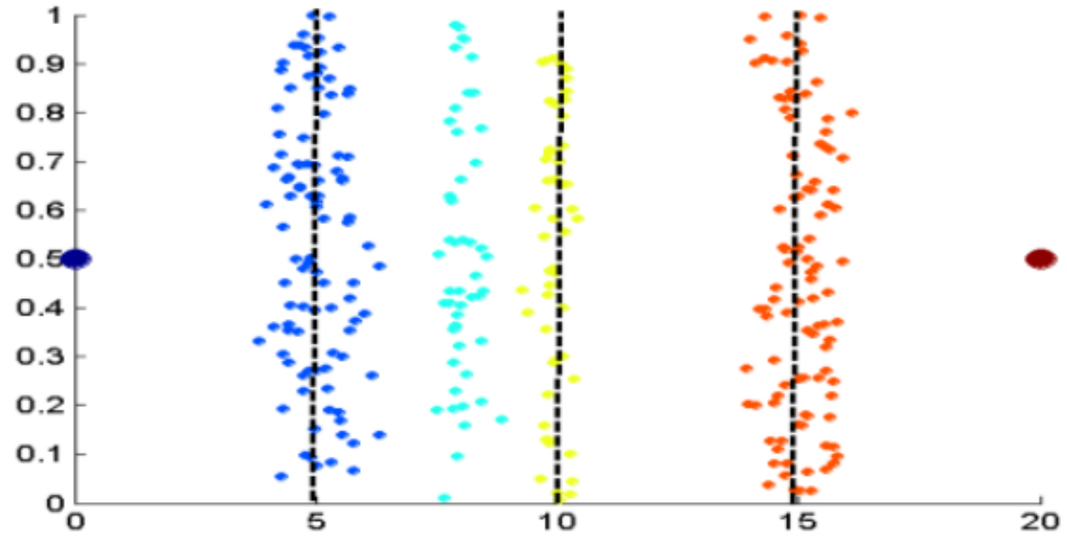
① Thresholding (choose a feature value as a split point)

② Discretization (transforms quantitative feature to ordinal where each ordinal value is referred to as bin)

< Quantitative → Discrete >
(continuous)

Unsupervised Discretization

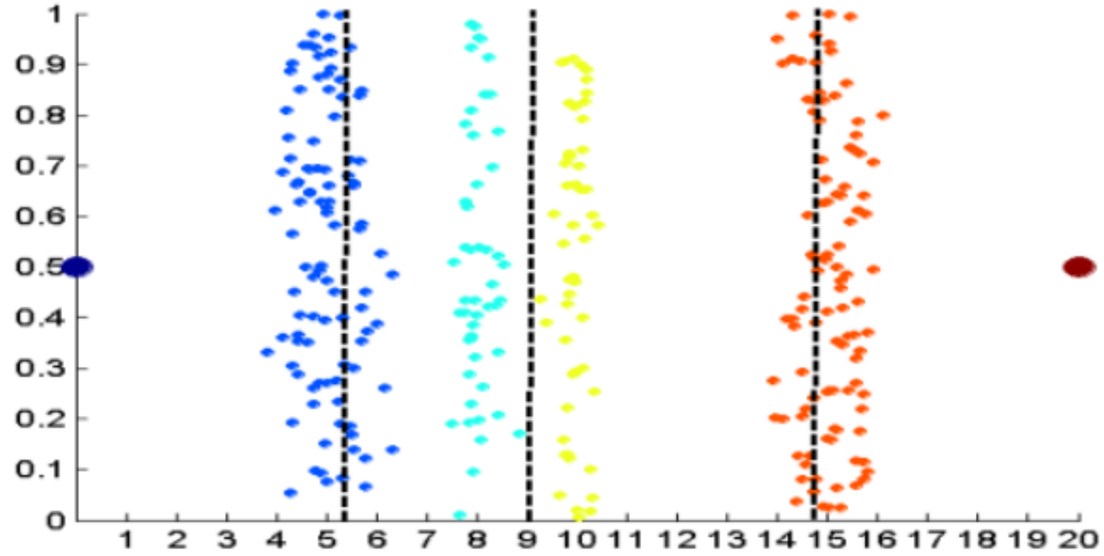
136



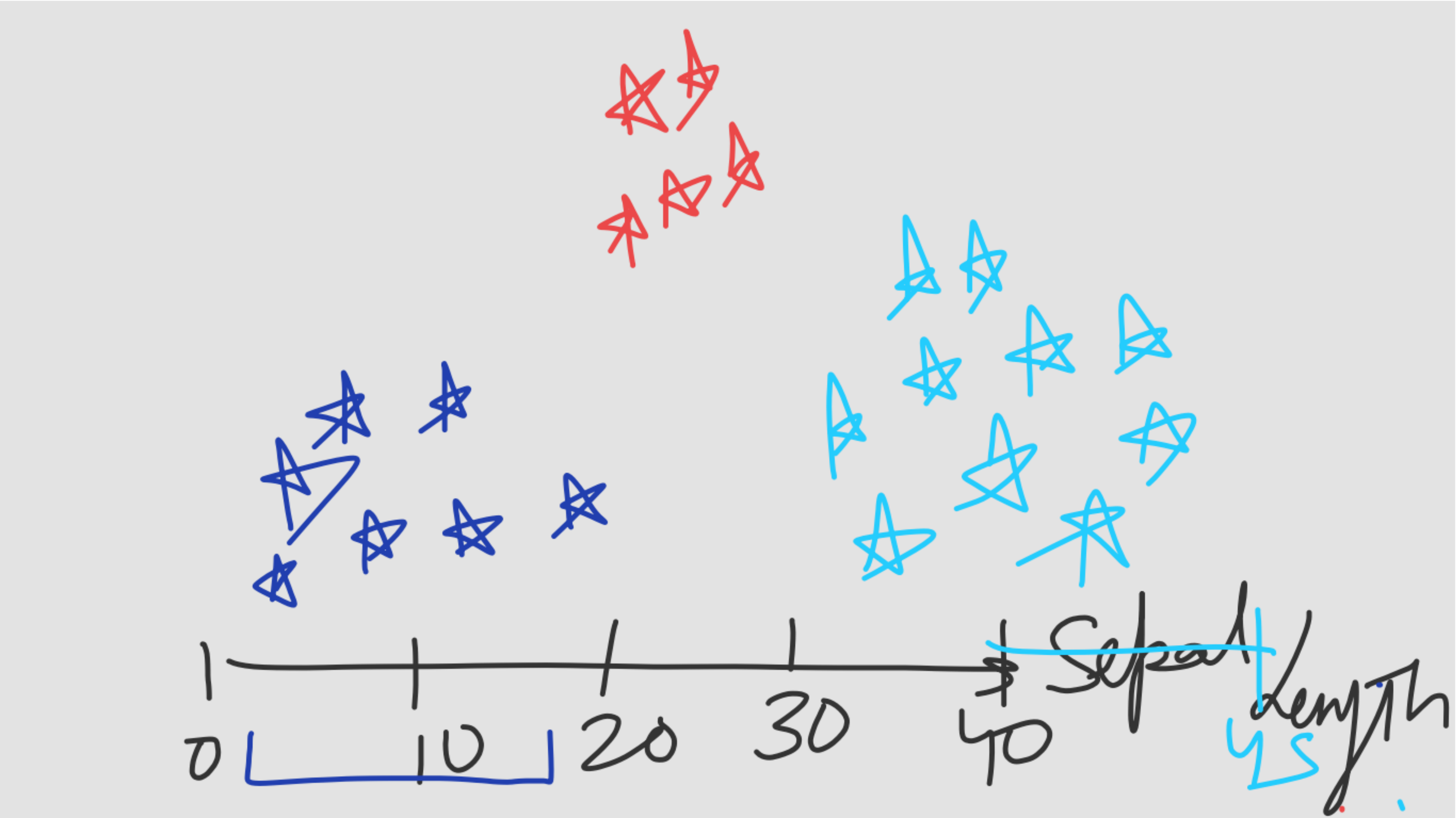
Equal interval width approach used to obtain 4 values.

Unsupervised Discretization

137



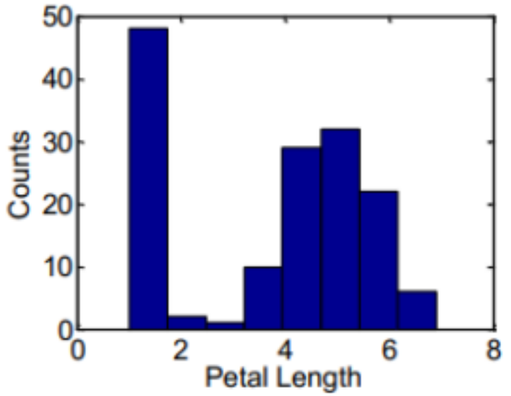
Equal frequency approach used to obtain 4 values.


1
0
10
20
30
40
Sepal
Length

Supervised Discretization

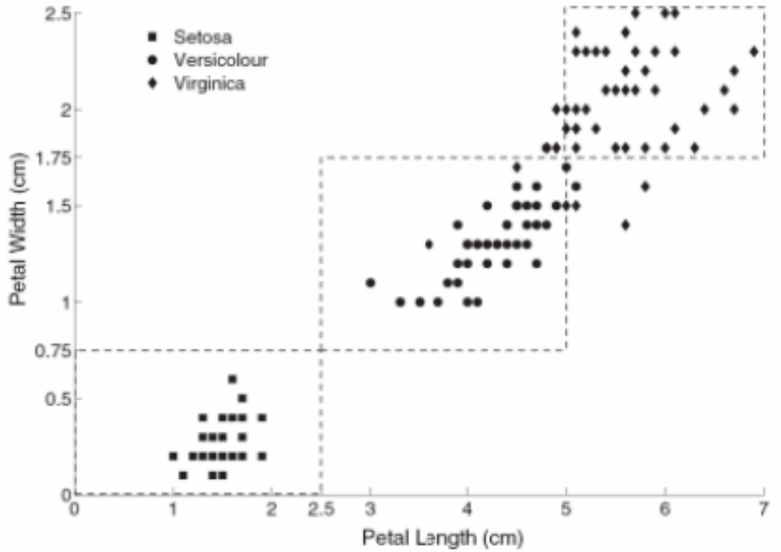
Discretization: Iris Example ...

- How can we tell what the best discretization is?
  - Unsupervised discretization: find breaks in the data values
    - Example: Petal Length



  - Supervised discretization: Use class labels to find breaks

Discretization: Iris Example



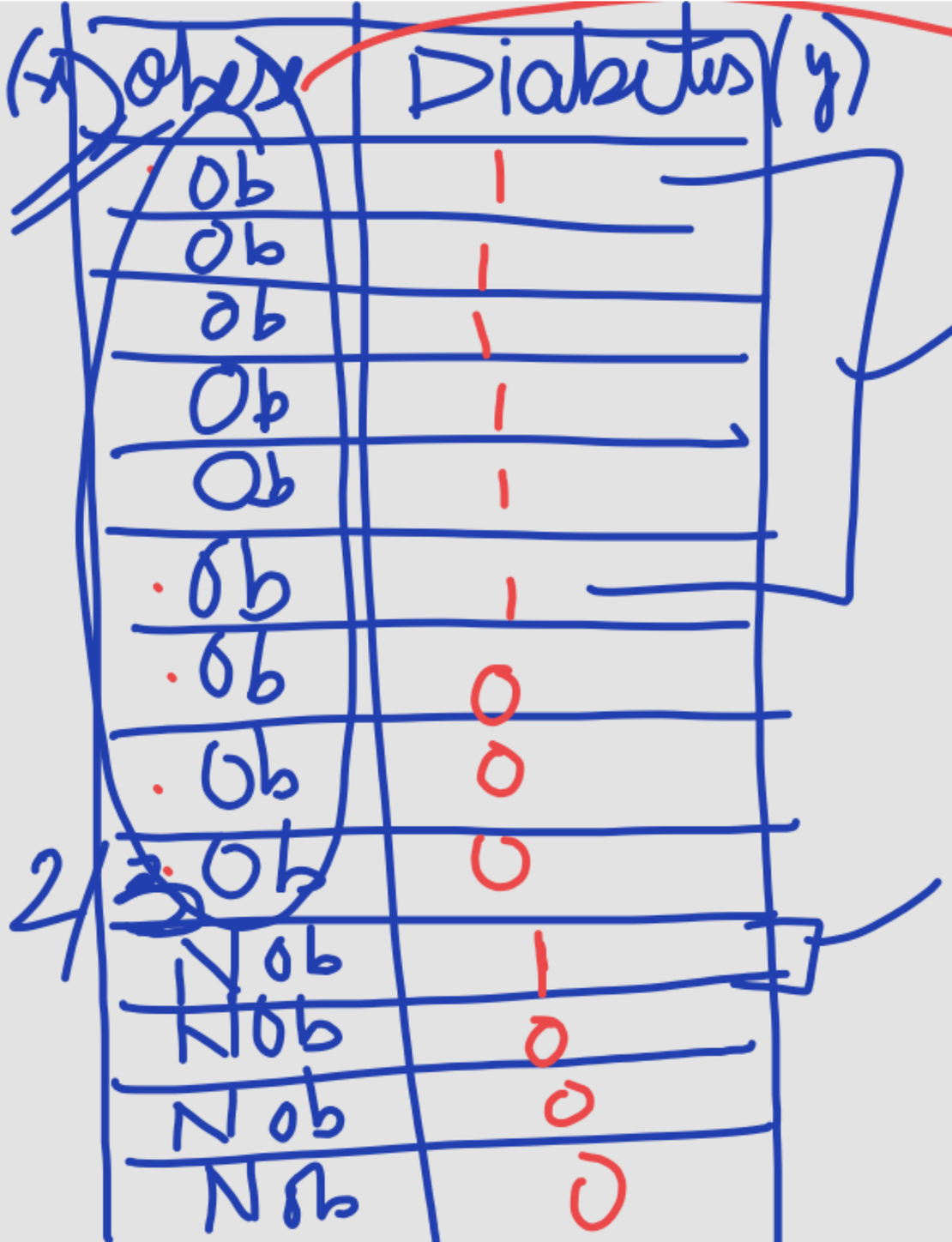
Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

Feature transformations that add scale to Ordinal or Categorical features

Quantitative
1. Unsupervised
   - Normalization → Ordinal
   - One-hot Encoding → Categorical
2. Supervised : Calibration

Categorical feature
→ Compute posterior prob. of class based on feature value
→ Adv. in Naive Bayes (No additional training only features calibrated)

Ordinal feature
↓
Discretize
↓
Apply Calibration similar to Categorical feature



Transformed feature
S.t. it encodes class info

$$P(x=\text{obese} \mid y=1)$$

$$P(y=1 \mid x=\text{obese}) = \frac{6}{9} \times 1 = .66$$

(x) Obese	Diabetes (y)
Ob	1
Ob	1
Ob	1
Ob	1
Ob	1
Ob	1
Ob	1
Ob	0
Ob	0
Ob	0
N ob	1
N ob	0
N ob	0
N ob	0

Calibrated feature

$$P\left(\frac{D}{Ob}\right) = \frac{6}{9} = 2/3$$

$$P\left(\frac{D}{N\,Ob}\right) = \frac{1}{4} = 1/4$$

Feature Construction

- Take Cartesian product of Original features
- ☹ Remove bias of Naive bayes that features have to be treated as independent
- ☺ Newly added feature may ↑ info gain (decision tree)

x_1 Shape $\in$ {Circle, Triangle, Square}

x_2 Color $\in$ {Red, Green, Blue}

Cartesian Product ?

x_1	x_2
Circle	Red

x_1, x_2

Red Circle

For quantitative features

→ Take arithmetic or polynomial combination

Eg: Marks1 | Marks2 .. | MarksN | —— Pass

→ Dividing mass by volume to get density.

Curse of Dimensionality

↳ As the no. of dimensions ↑ so, it becomes difficult to learn a model

- Classification → hinders models learning capability
- Clustering → distance not meaningful

→ ① Feature Selection

→ ② Feature Reduction

Feature Selection Approaches

Another way to reduce dimensionality of data- Use subset of features

Redundant features

- Duplicate much or all of the information contained in one or more other attributes
- Example: purchase price of a product and the amount of sales tax paid

Irrelevant features

- Contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

Many techniques developed, especially for classification

Attr. Corr with Y

1

2

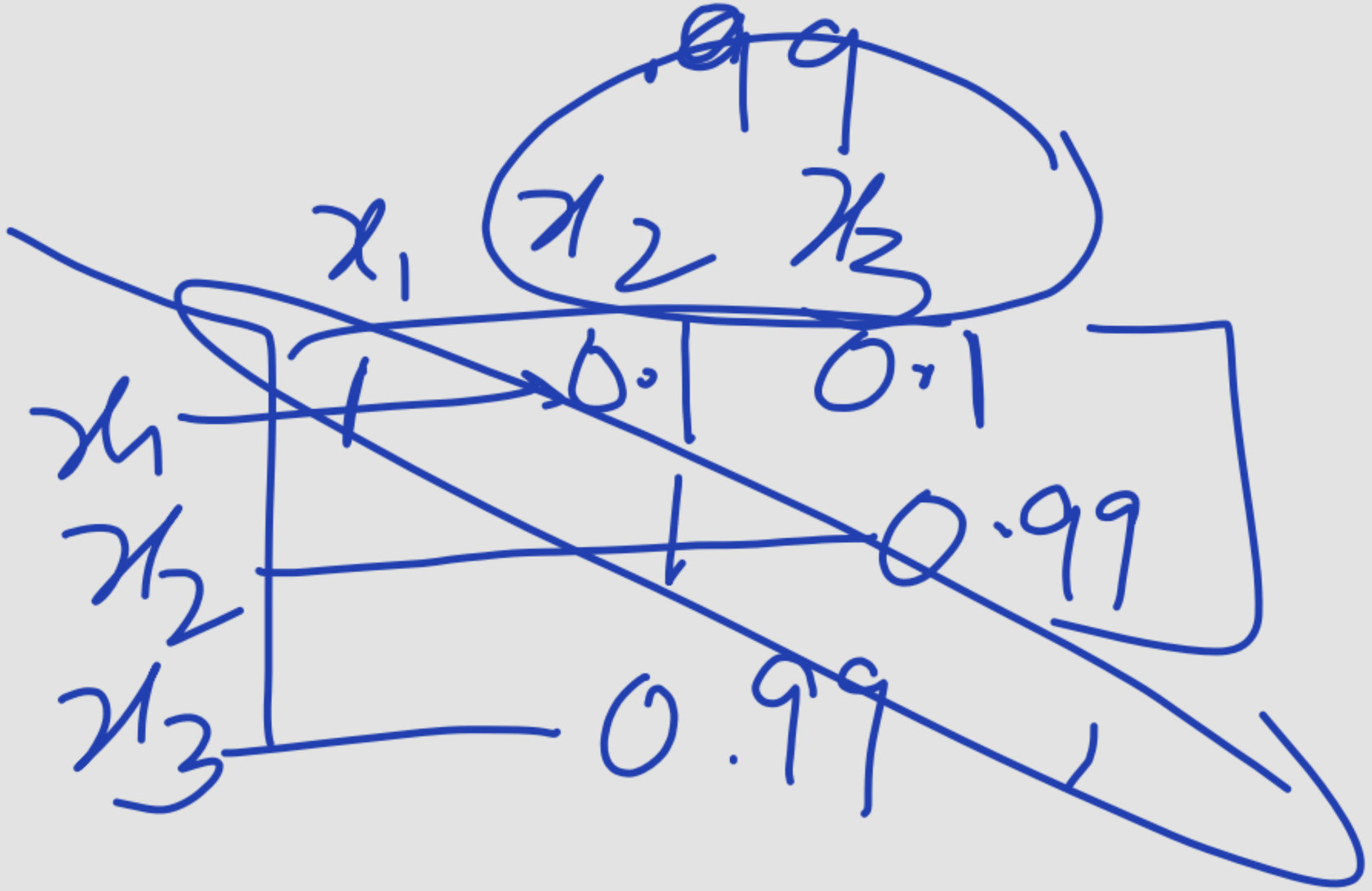
3

Embedded approaches Feature selection occurs naturally as part of the data mining algorithm. Specifically, during the operation of the data mining algorithm, the algorithm itself decides which attributes to use and which to ignore. Algorithms for building decision tree classifiers, which are discussed in Chapter 4, often operate in this manner.

Filter approaches Features are selected before the data mining algorithm is run, using some approach that is independent of the data mining task. For example, we might select sets of attributes whose pairwise correlation is as low as possible.

Wrapper approaches These methods use the target data mining algorithm as a black box to find the best subset of attributes, in a way similar to that of the ideal algorithm described above, but typically without enumerating all possible subsets.

① Pairwise correlation



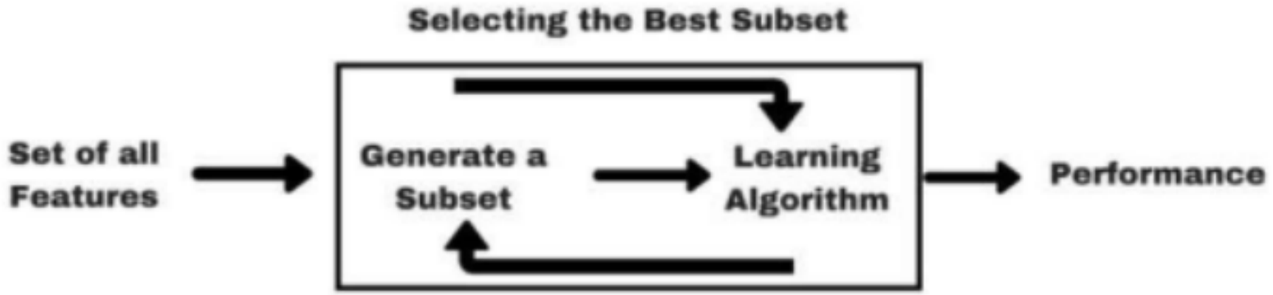
x_2	x_3
10	20
20	40
30	60
40	79

$y = 2x_2$

$y = x_3$

2. Filter Methods

3. Wrapper Methods



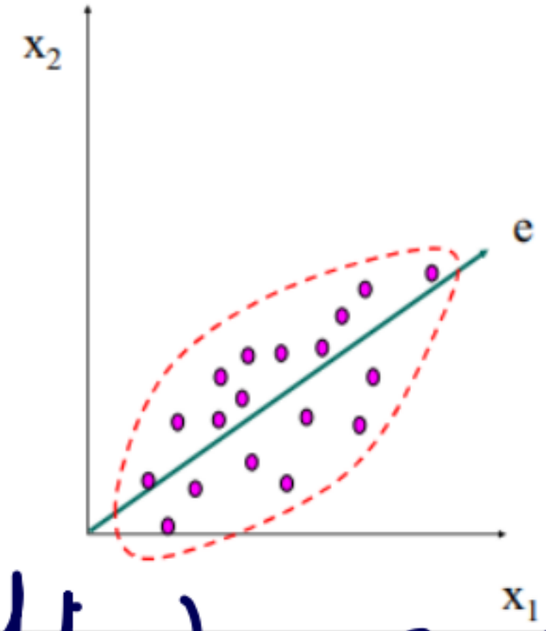
In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

Feature Reduction Approach: PCA

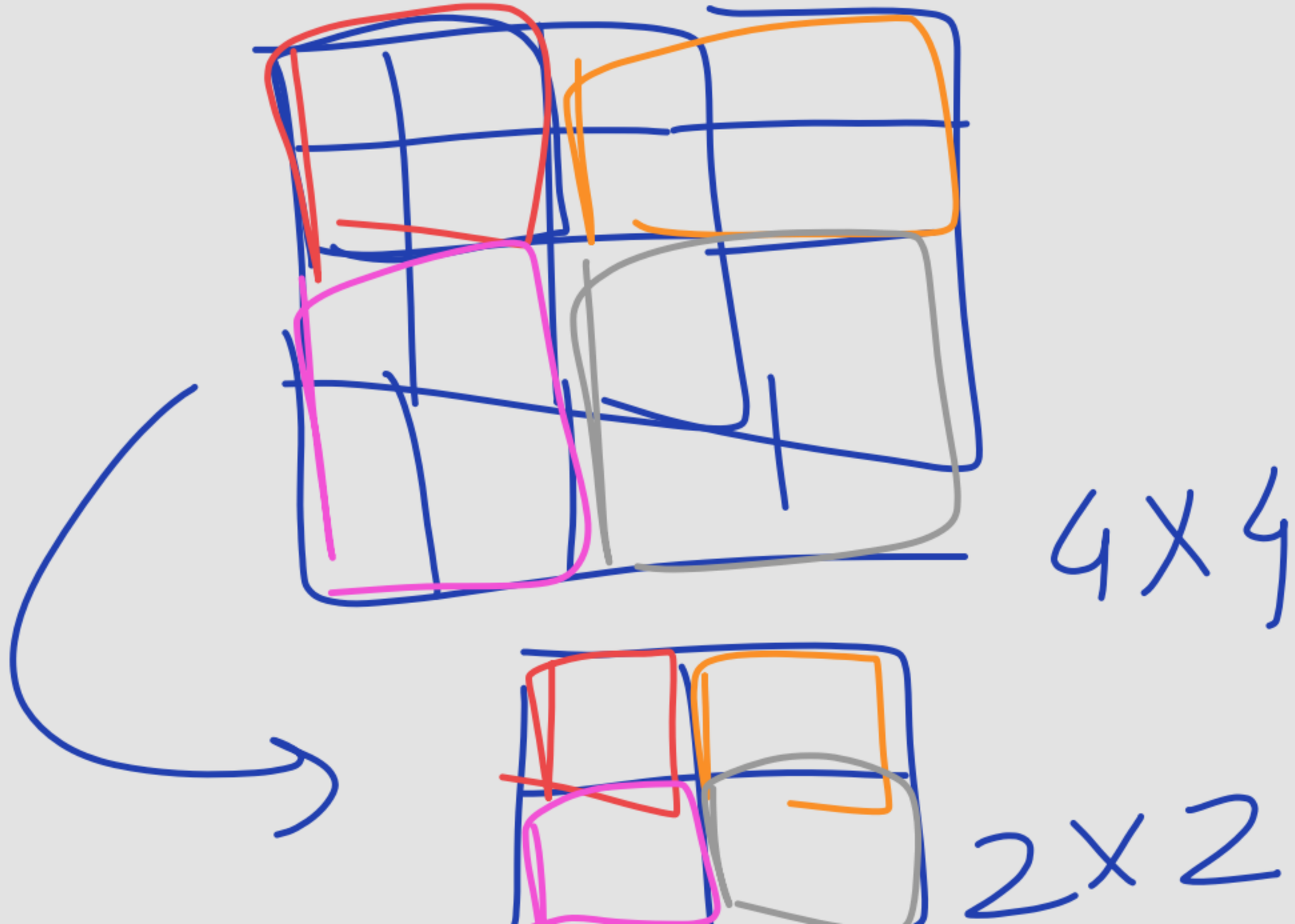
Dimensionality Reduction: PCA

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
- Techniques
  - Principal Components Analysis (PCA)
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

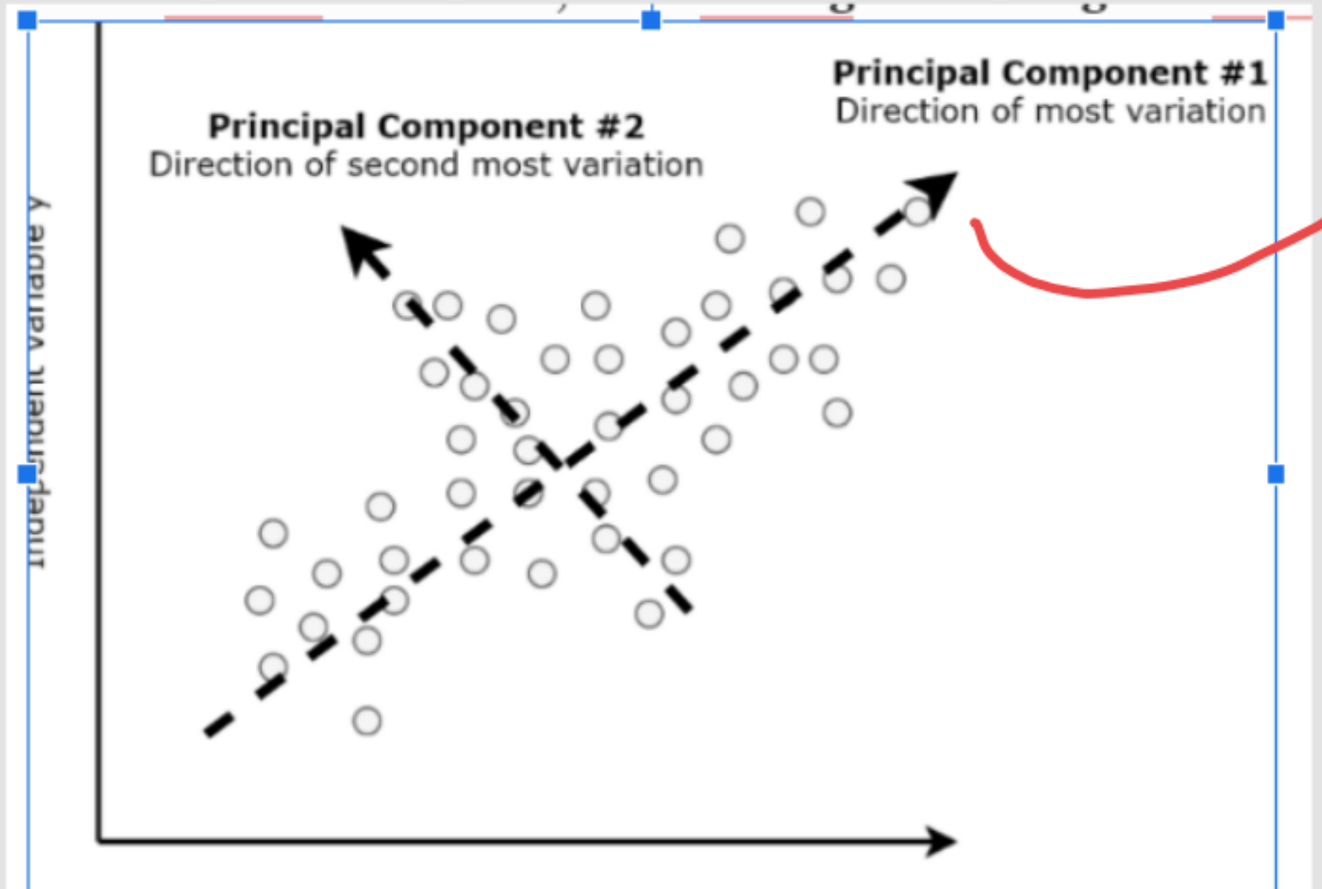
- Goal is to find a projection that captures the largest amount of variation in data



→ May be applied on 256x256 image to yield 64x64 image.


4x4
2x2

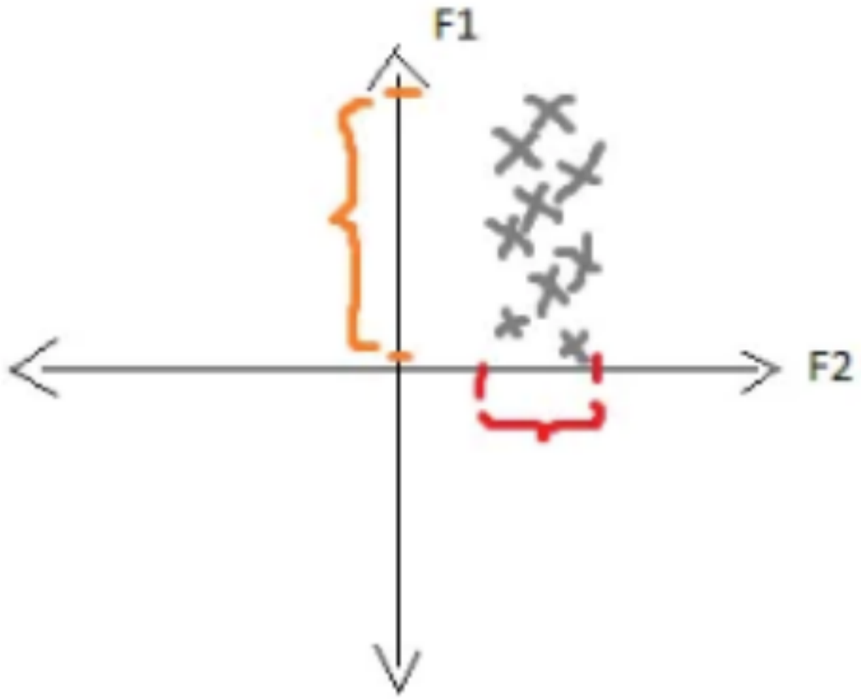
1. Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.
2. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.
3. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



Variation(spread) of values along this line should be maximal.

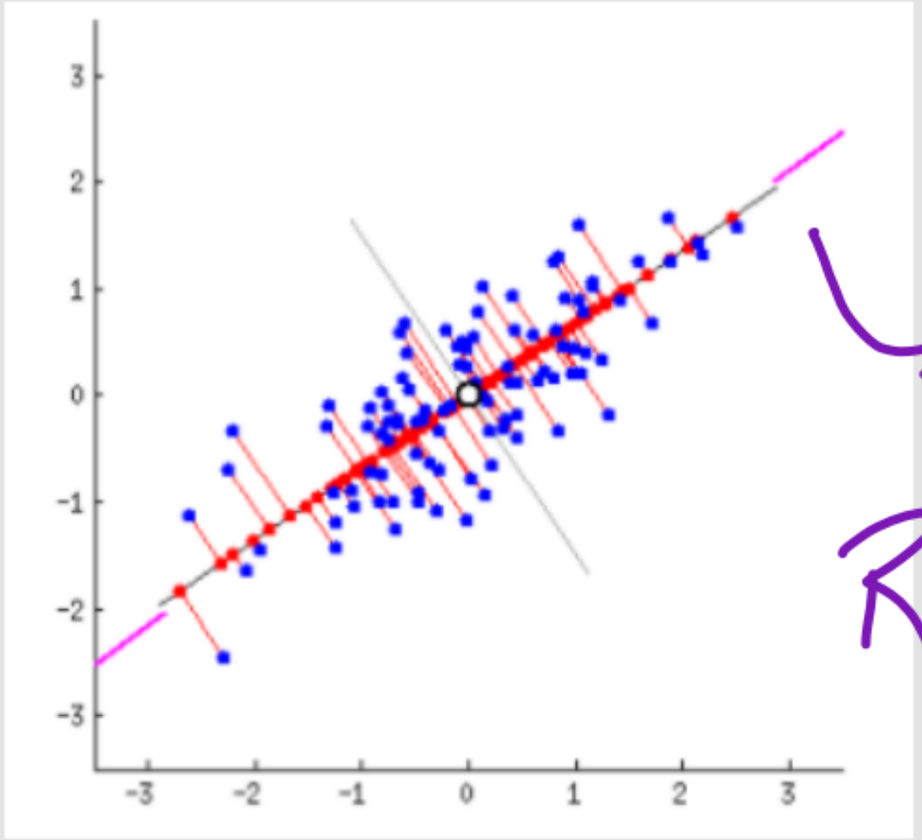
Or

Principal components represent the directions of the data that explain a maximal amount of variance



In above image ‘x’ represents the data points plotted in 2D plot. F1, F2 are the features in dataset(X). Orange selection shows the spread of data points on F1. Red selection shows the spread of data points on F2.

We can observe that spread/variance of data points on F1 is more than



Minimal Reconstruction Error!

Summarizing the PCA approach

Listed below are the 6 general steps for performing a principal component analysis, which we will investigate in the following sections.

1. Take the whole dataset consisting of d-dimensional samples ignoring the class labels
2. Compute the d-dimensional mean vector (i.e., the means for every dimension of the whole dataset)
3. Compute the scatter matrix (alternatively, the covariance matrix) of the whole data set
4. Compute eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) of covariance matrix
5. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector)
6. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the mathematical equation: $\boldsymbol{y} = \boldsymbol{W}^T \times \boldsymbol{x}$ (where $\boldsymbol{x}$ is a $d \times 1$-dimensional vector representing one sample, and $\boldsymbol{y}$ is the transformed $k \times 1$-dimensional sample in the new subspace.)

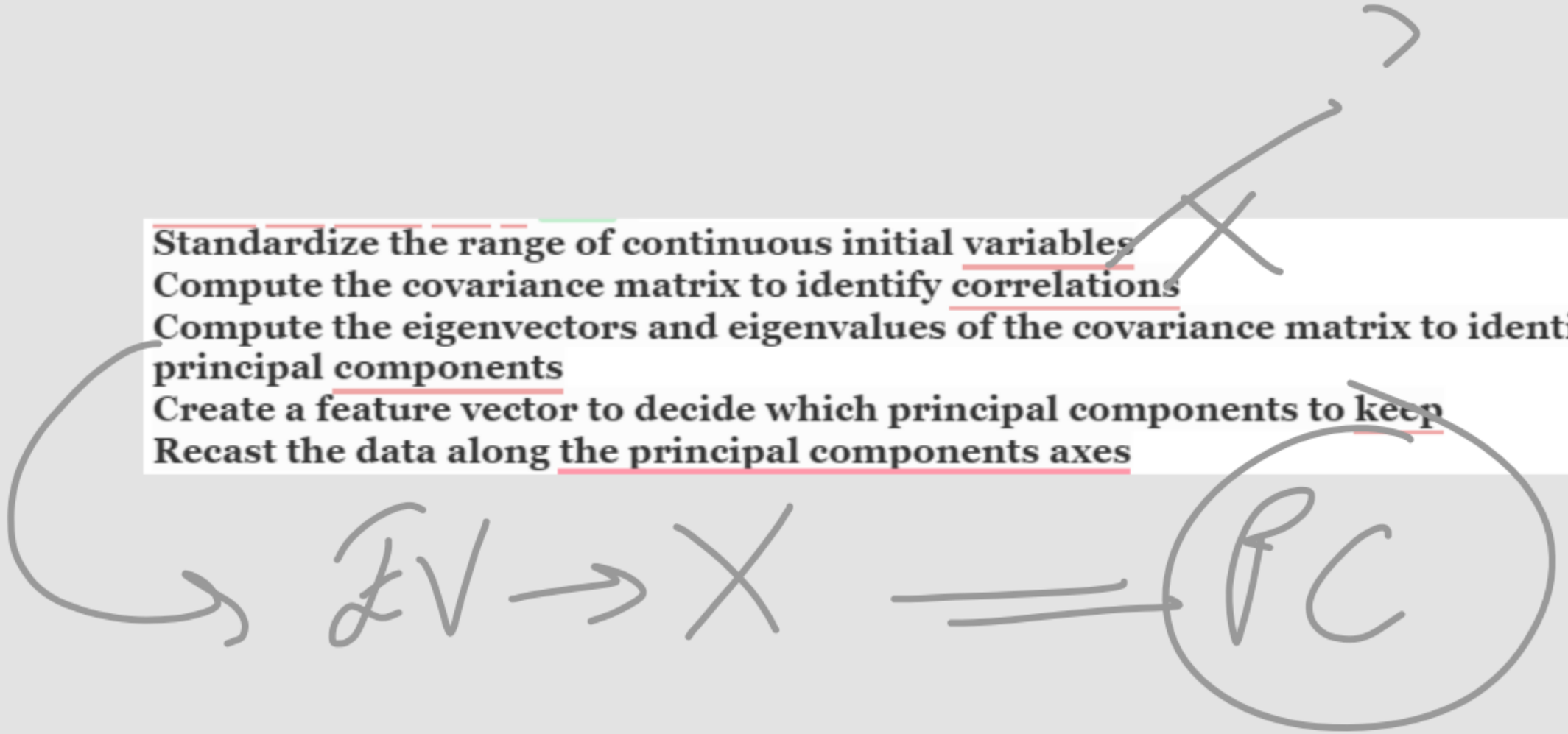
gives variance or direction of spread

length or magnitude of eigen vector

Just go thrayh blue text

PCA we are trying to find the axes (EIGEN VECTORS) with maximum variances where the data is most spread

Standardize the range of continuous initial variables
Compute the covariance matrix to identify correlations
Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
Create a feature vector to decide which principal components to keep
Recast the data along the principal components axes



Kind Of Features

Qualitative

Categorical or Nominal (Discrete)
- → no Scale or Ordering
- → Eg: Eye Color or Gender, Marital Status
- → Measure of Central Tendency: Mode
- → Can only be compared for equality $=, \neq$

Ordinal (Discrete)
- → Features with Ordering, but no Scale.
- → Ordinal features have Ordered Set of discrete values as
- → Eg: Grade or Hall No. (good, better, best)
- → Measure of Central Tendency: Mode, Median
- → Measure of Dispersion: Quantiles (M, Med, Me)
- → Discrete Values that provide enough info. to Order Objects ($>, <, =$)

Quantitative (Continuous)
- → Features with a numerical Scale
- Eg: Age, Price
- → Measure of CT: Mean
- → Measure of DT: range, IQ range, variance, SD, ...
- → Continuous attr on which Arithmetic operators ($+, -, *, /$) can be applied

Feature Transformation
→ Improving utility of a feature by removing, changing, or adding info

Category 1: Converting feature of one type to feature of another type
Category 2: Changing Scale of Quantitative features (or adding a scale to Ordinal, categorical and boolean features)

Application: Linear Regression
Type of features req: Numerical / Quantitative

① I/P: Quantitative ↓
② I/P f: Ordinal ↓ → Change to numerical values such that Order is preserved

- Normalization (adapting the scale in unsupervised manner)
- Calibration
  - ↳ Supervised feature transformation
  - ↳ adding a meaningful scale carrying class info to arbitrary features

Good	0
Bad	1
Worst	2

③ I/P f: Categorical → Binarization
↳ One Hot Encoding
Map a feature to # of features = # of Unique/Distinct values (i.e. one for each value of categorical feature)

∴ they have common value for the feature ✗

(*) Variance for Marks: 50, 60, 70

$\frac{1}{3}\frac{1}{3}\left((10)^2+(0)^2+(10)^2\right)$

$\sqrt{\frac{200}{3}} = \sqrt{66.66}$ (10)

Deviation from mean

Do you Know?? There are 3 Type of Statistics?

① Statistics of Central Tendency (Representative Point / Central Value / Middle Value) that attempts to describe the whole dataset

1.1 Mean or average Value
1.2 Median : Middle Value
1.3 Mode : Most Freq. Value (Categorical) & Most freq. grade

* Median is preferred to mean for Skewed distribution (Long Right or Long left tail)

② Statistics of Dispersion (disperse / Spread)

2.1 Variance (Av. Squared deviation from mean)
2.2 Std Dev (Sq. root of Var.)
(Both Same, SD measured on Same Scale as feature i.e. Kg as Compare to Kg)
2.3 Range : Diff between max. and min. value
2.4 Percentile : p-th percentile is the value s.t. p% of the instances fall below it.

If multiple of 25 it is called quartile | If p is multiple of 10, it is called decile

③ Shape Statistics → Skewness (m_3/σ^3) : $\frac{1}{n-1}\sum_i (x_i-\mu)^3$: +Value : Distribution is Right Skewed, Right tail is longer than left tail; -ve : left Skewed Distribution

↳ Kurtosis $(m_4/\sigma^4) = \frac{1}{n-1}\sum_i (x_i-\mu)^4$

+ve Value : Distribution is Sharply peaked than normal distr.

Q1 Diff b/w Arithmetic Mean : $\frac{a+b}{2}$

Harmonic Mean : $(1/a)+(1/b)/2 = \frac{2ab}{a+b}$

Geometric Mean : $\sqrt{ab}$

extent to which distribution is spread / streched

extent of deviation or spread around mean or Central point

★ or Temp → Cold / Hot; Fever Y/N; Temp.

② To Discrete

2.1 To Categ. Application : Classification / Decision Tree

Say Type of feature req. : Categorical

Or We want Target Value to be a Category

Feature : Attend | T1 IA | T2 IA | Past CGPA

Say CGPA grade: A, A+, A, B+, B → Class; Map CGPA Regi.

① I/P : Categorical ✓

② I/P : Ordinal : → UnOrdering (Categories Discarding Ordering) → Converts Ordinal feature to

③ I/P : Quantitative

Eg: Mapping [4, 5, 6] : [Short, Medium, Tall] Height

Techniques to Convert Ordinal / Quantitative feature to Categorical

① Thresholding (Single Threshold) → Choose a feature value u as split point (Boolean Category)

$f : x_1 \rightarrow \{Class1, Class2\}$

Petal length: Setosa, Versi, Virginica

or Supervised Threshold

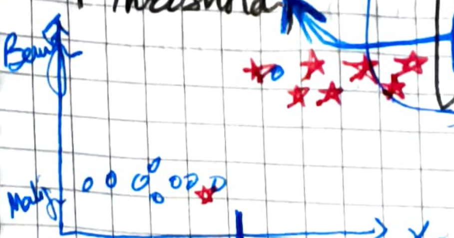
Unsupervised Thresholding: for any training instance i, if $f(x_1^{(i)}) > u$ Class0 else Class1

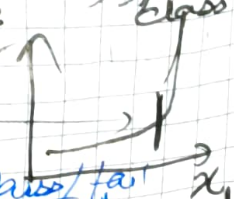
→ Split point could be mean

[or check the impact on Target Class

→ Most Sensible thresholds are mean & median

Knee point





Eg. Marks → pass/fail, → grade

Ht Short / Medium / Tall

Tumor Size, Choose this a Split point

~~2~~ To Ordinal Grade CGPA → Grade (multiple thresholds)

(2) Discretization
↳ Transforms Quantitative feature to Ordinal feature

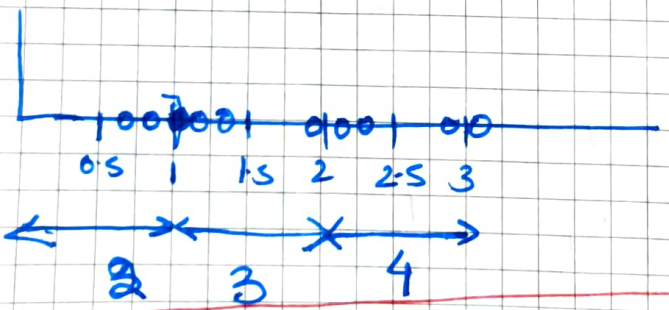
→ Each Ordinal Value is referred to as bin and corresponds to an interval of the Original quantitative feature

Unsupervised Discretization

2.1 Equal-frequency Discretization
Prereq → Need to know no of bins beforehand
→ Chose the bins so that each bin has approx. equal no of instances
Eg: [4 bins: Bin boundaries will be quantiles (divides with 10 bin)]

2.2 Equal Width Discretization
→ Chose bin boundaries so that each interval has the same width.
↳ Can be determined by dividing the feature range by no of bins.

Supervised Discretization
(Top Down & Bottom up)
Divisive
↳ Progressively Split Bins
(Clustering)
Agglomerative
↳ Starts with each instance as a bin & Successively merge bins



0	S
0.1	S
0.2	S
0.3	V
1.3	V
1.4	V
1.5	V
2.4	Vi
3	Vi

★ Feature Transformations that remove Scale of quantitative features — ① Thresholding
② Discretization

★ Supervised feature transformation adding a meaningful Scale carrying class info to arbitrary features
① Feature Calibration