Session 4 : جلسه 4 :

مدلهای ML

① Parametric m ← آسان

② Non_Parametric m ← آموزش [illegible]

مدل براساس داده‌های ورودی یاد میگیرد

مزیت: ساده، سریع، انعطاف‌پذیر [illegible]

عیب: برای داده‌های بزرگ [illegible] کند

K_NN

K nearest neighbors

① memory_based m

② Instance_based model

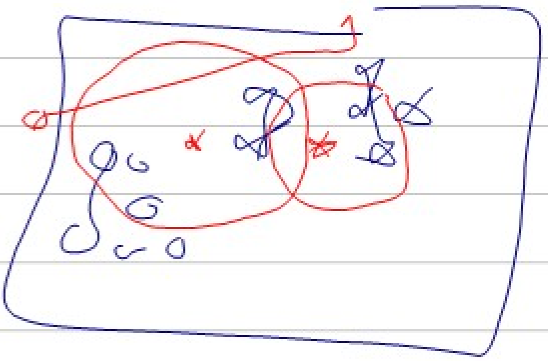
Class

reg

k-nn

مقایسه با اطلاعات داده

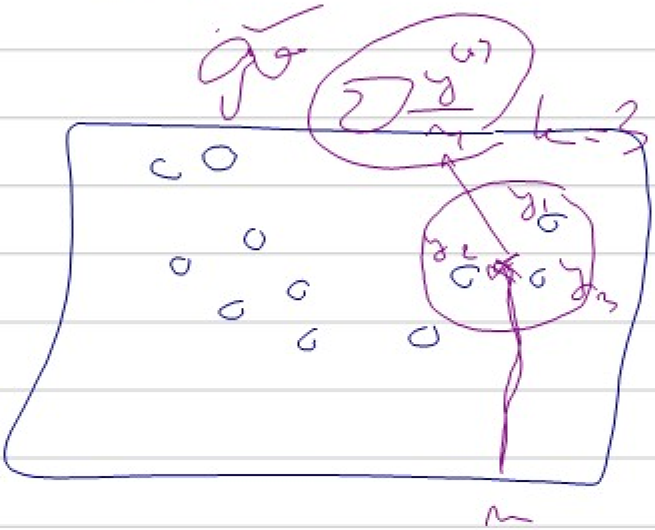
رای گیری



k چ انتخاب

k کوچک

خیلی کوچک

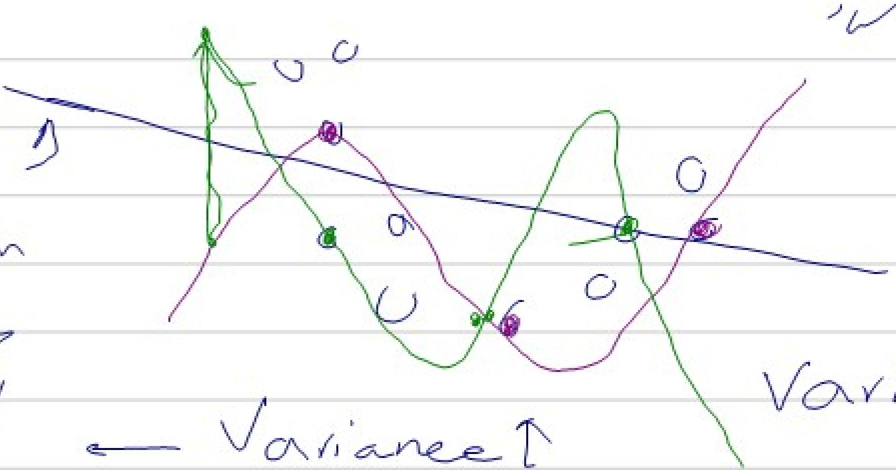


overfitt مدل ← k↓ V↑ non_liner

underfitt مدل ← k↑ مدل به مدل ~~خطی~~ liner نزدیکتر میشه

مدل ساده

که برای داده های train بیشتر خطاهای زیادی داره

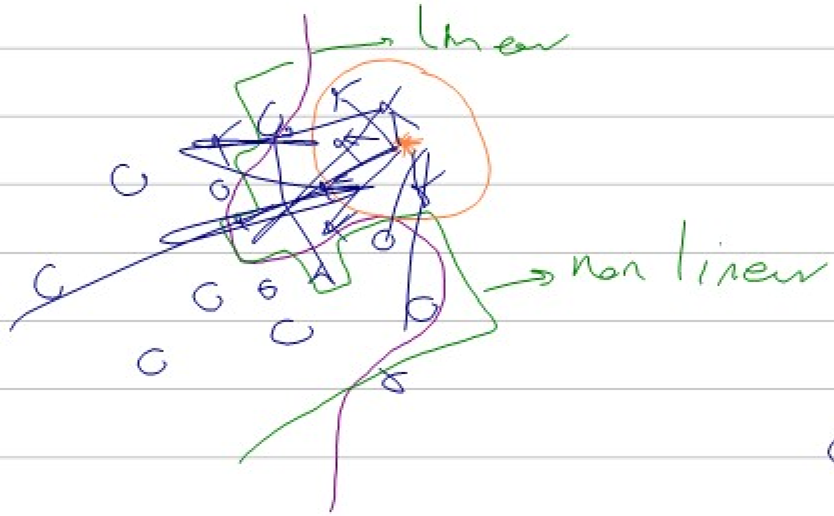


Variance ✓

bias ✓

overfit ← Variance↑

underfit ← bias↑



distance metric

فاصله اقلیدسی ①

$$d(x,x') = \sqrt{(x_1-x_1')^2+(x_2-x_2')^2+\dots(x_D-x_D')^2}$$

$$\underset{Ter}{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \qquad \underset{Tes}{x'} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_D' \end{bmatrix}$$

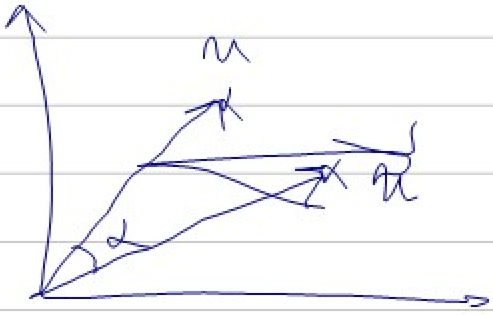
فاصله وزن دار ②

$$d_w(x,x') = \sqrt{\underset{.8}{\overset{\alpha}{\boxed{w_1}}}(x_1-x_1')^2+\underset{.3}{w_2}(x_2-x_2')^2+\dots+w_D(x_D-x_D')^2}$$

minkowski ③ distance

$$d(x,x') = \left(\sum_{i=1}^{n}(x_i-x_i^{*})^p\right)^{\frac{1}{p}}$$

Cosine distance ④



$$d(x,x') = 1 - \text{Cosine similarity}(x,x')$$

$$\text{Cosin}(x,x') = \frac{x \cdot x'}{\|x\| \, \|x'\|} \quad \frac{\sum x_i x'_i}{\sqrt{\sum_i x_i^2} \, \sqrt{\sum_i x'^2_i}}$$

---

classification معیارها

اولی

+ −

− + اشتباهی

: accuracy ①

100 سالم −N

100 بیمار +P

30 FP 70 TN 20 FN 80 TP

درست T

نادرست F

P } دارای بیماری

N }

real

	P	N
تشخیص +	TP	FP
−	FN	TN

$$acc = \frac{TP + TN}{\text{کل}}$$

+ سرطان ۱۰
- سالم ۹۹۰
} ۱۰۰۰ ش

$$acc = \frac{0+990}{1000} = 99\%$$

imbalanced data

$$\begin{array}{c|c|c} & 0 & 1 \\ \hline 0 & TP & FP \downarrow \\ \hline 1 & FN & \boxed{TN} \end{array}$$

$\xrightarrow{} \frac{TP}{FR+FP}$ Sensitivity (Recall) ②

$\xrightarrow{} \frac{TN}{FN+TN}$ Precision ③

$$\frac{TN}{FP+TN}$$

$$\frac{TP}{TP+fN} = \frac{0}{0+990} = 0 \times$$

Recall

$$F1\_Score = \frac{2PR}{P+R} = \frac{2 \times 1}{\frac{1}{P}+\frac{1}{R}}$$

F1_Score ④

مدل

$$\text{پیش‌بینی}\begin{cases} T \\ F \end{cases}$$

$$\begin{rcases} P \\ N \end{rcases}\ \text{مثبت / منفی}$$

<u>TP</u>

<u>TN</u>

		واقعی P	N
پیش‌بینی	P	TP	FP
	N	FN	TN

<u>FP</u>

<u>FN</u>

class
Reg ← Ensemble learning

✗ strong learning

[illegible] [illegible] ضعیف ← weak learning

n

$N_1 \quad N_2 \quad N_3 \quad N_n$

[illegible] ←

[illegible] ←

Parallel ①
Seri

$\to N_1$ [illegible]
$\to N_2$ [illegible]
$\to N_3$ [illegible]

آموزش کل
end_to_end training

$$\overset{P_1}{N_1} \to \overset{P_2}{N_2} \to \cdots \to N_n$$

سری ②

Random Forest Parallel ← bagging ②

کاهش var موازی

سریال ← Boosting ①

کاهش Bias متوالی

✓ Adaboost

Xgboost ✓

---

bagging چیست

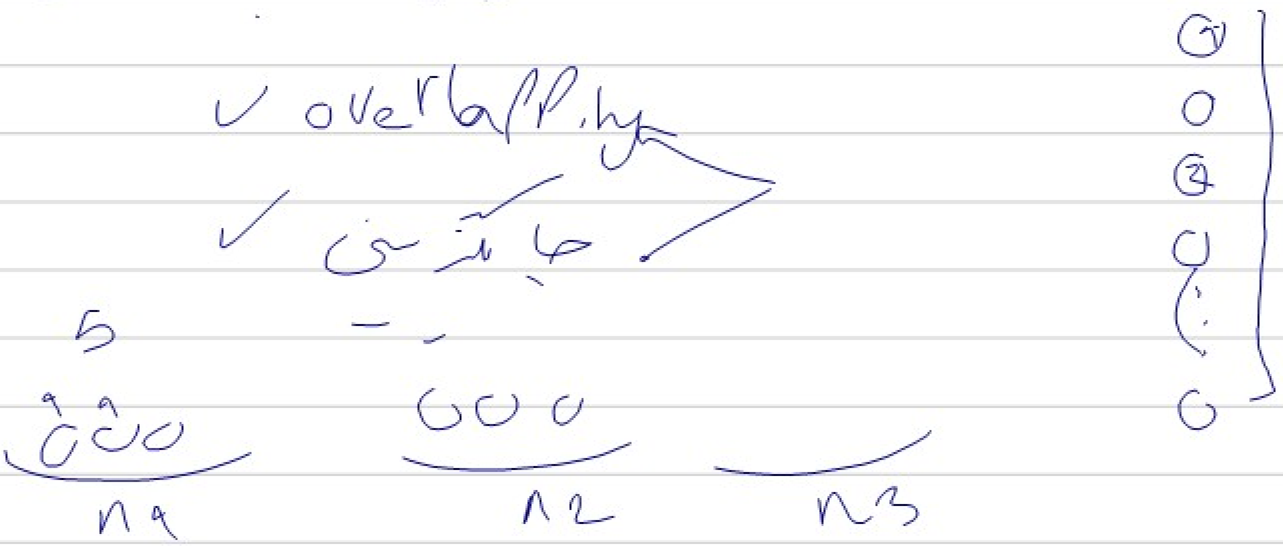
≡ Bootstrap aggregating

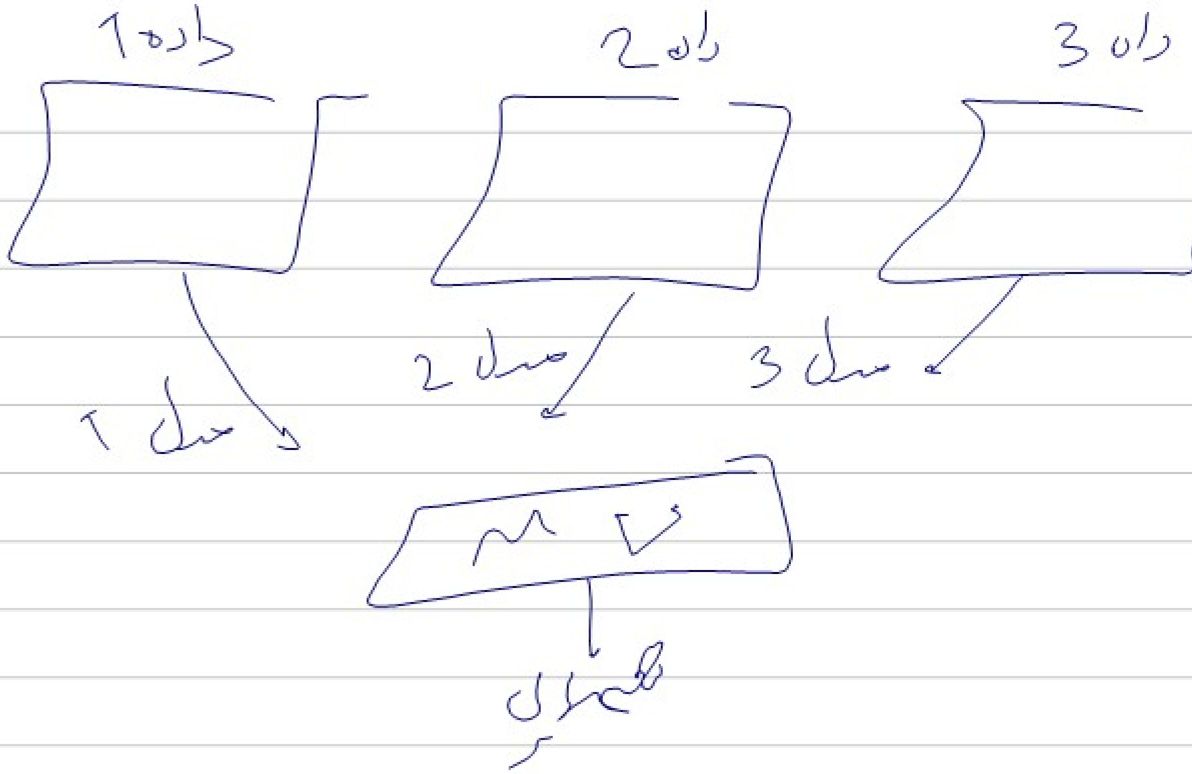
Bootstrapping چیست

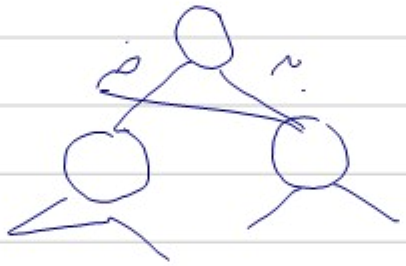
از داده‌ها n تا نمونه برداریم

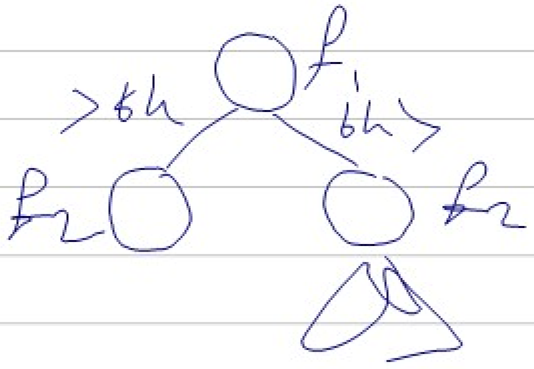
✓ overlapping

✓ جانشانی









Decision Tree مدل

رگرسیون

classification

↑

Information_gain

شاخص

[illegible]

[illegible]

MSE

reg

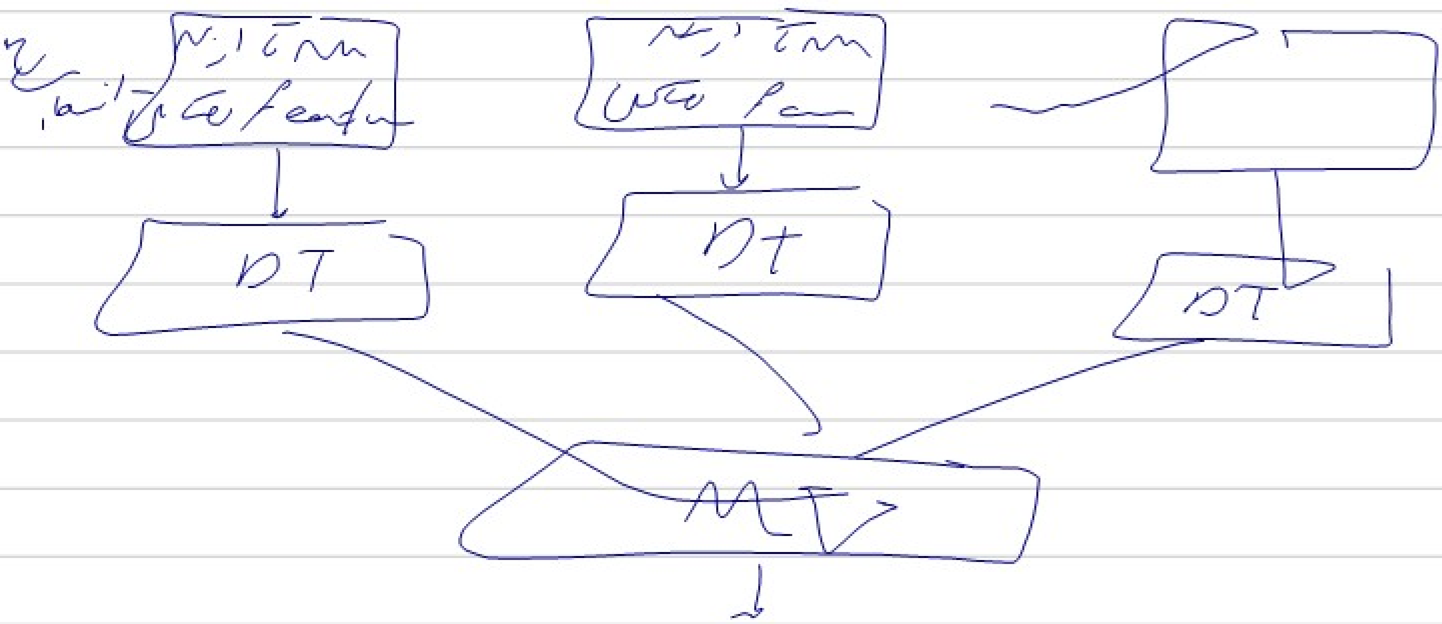
*

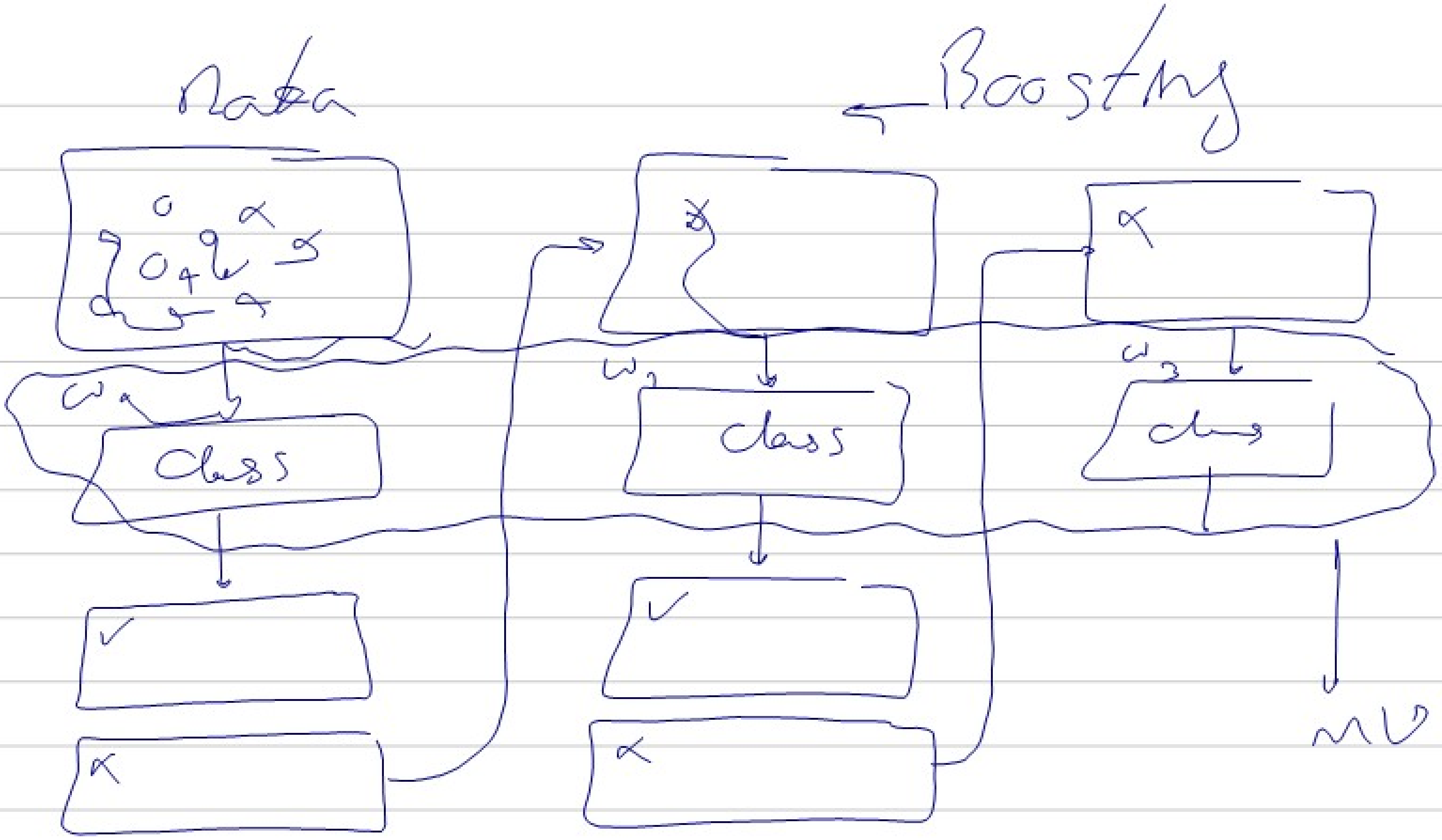
Random Forest

bagging [illegible]

DT → W

[illegible} N → P





Adaboost

xgboost

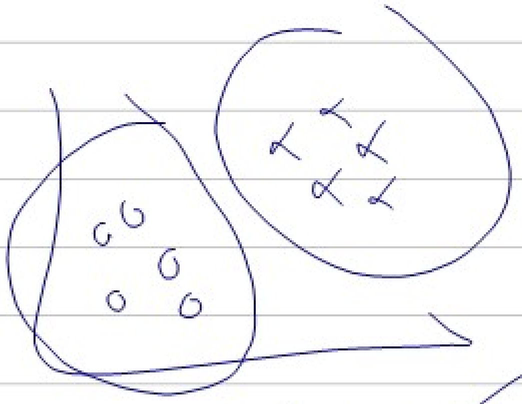
$$H_m(x) = \underline{\alpha_1 H_1(x)} + \underline{\alpha_2 H_2(x)} + \dots + \underline{\alpha_m H_m(x)}$$

$$\text{Prediction} = \operatorname{sign}(H_m(x))$$

unsupervised

گروه‌ها :

clustering ①



کاهش ابعاد

Dimentionality Red ②

تعداد فیچرها

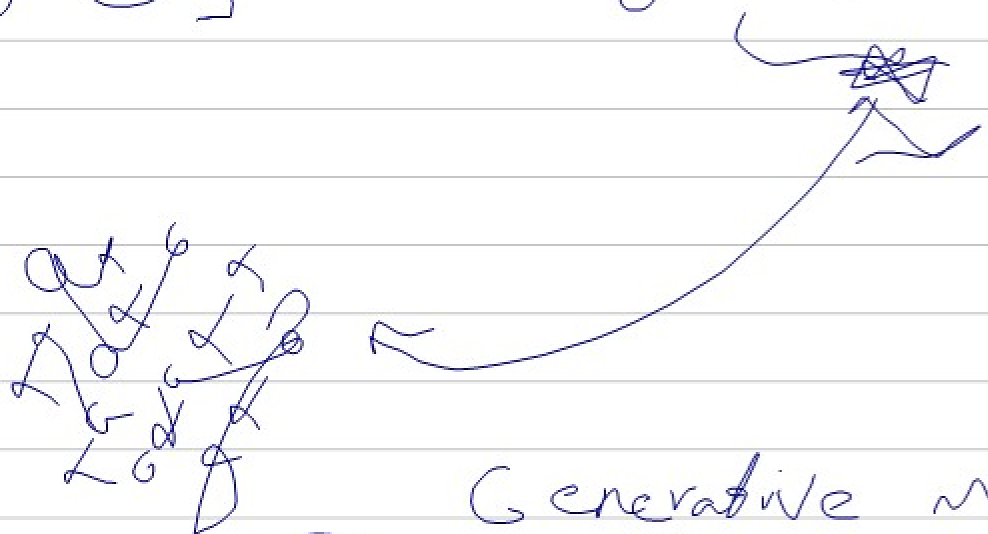
RGB

30,000

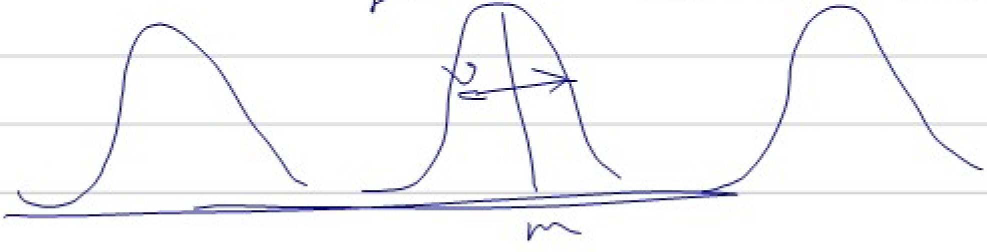
100 x 100 x 3

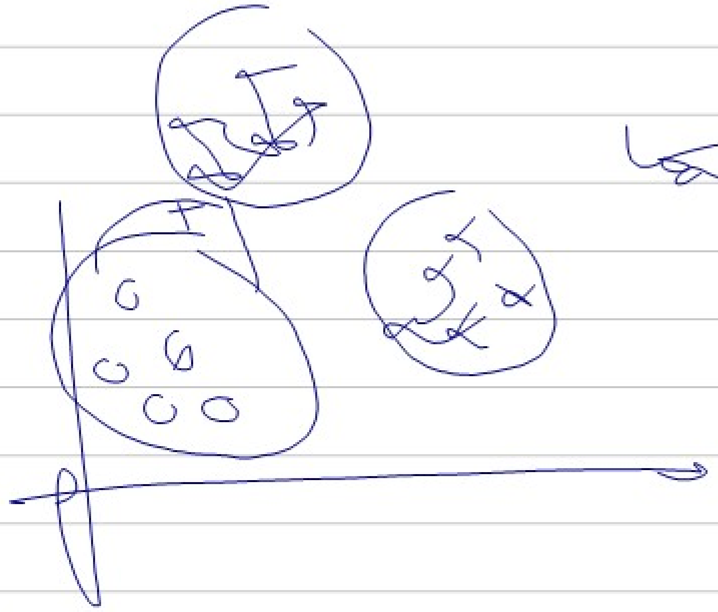
کشف ناهنجاری

Anomaly Det ③



Generative m ④





، clustering

پیدا کردن دسته‌ها

① کیفیت شباهت سنجی

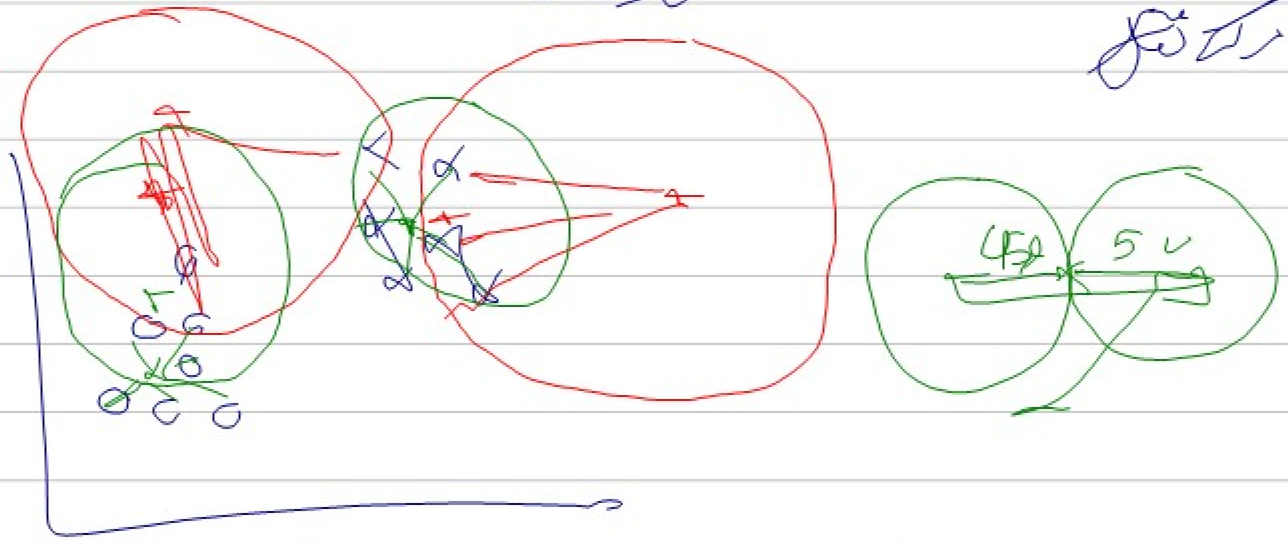
② تعیین تعداد دسته‌ها

---

k-means

تعداد دسته‌ها

مرکز ثقل

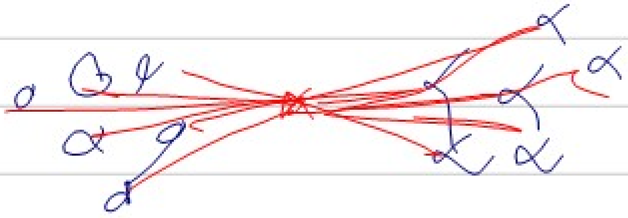


WCSS

within-cluster sum of square

$$WCSS = \sum_{\text{خوشه } 1}^{k} \sum_{x \in \text{خوشه } i} \|x - \mu_i\|^2$$

با نشان دادن WCSS ی مختلف بهترین k را بدست آورد



$WCSS_1 > WCSS_2$