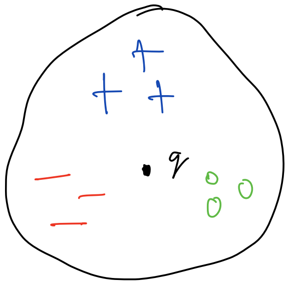
Agenda for today

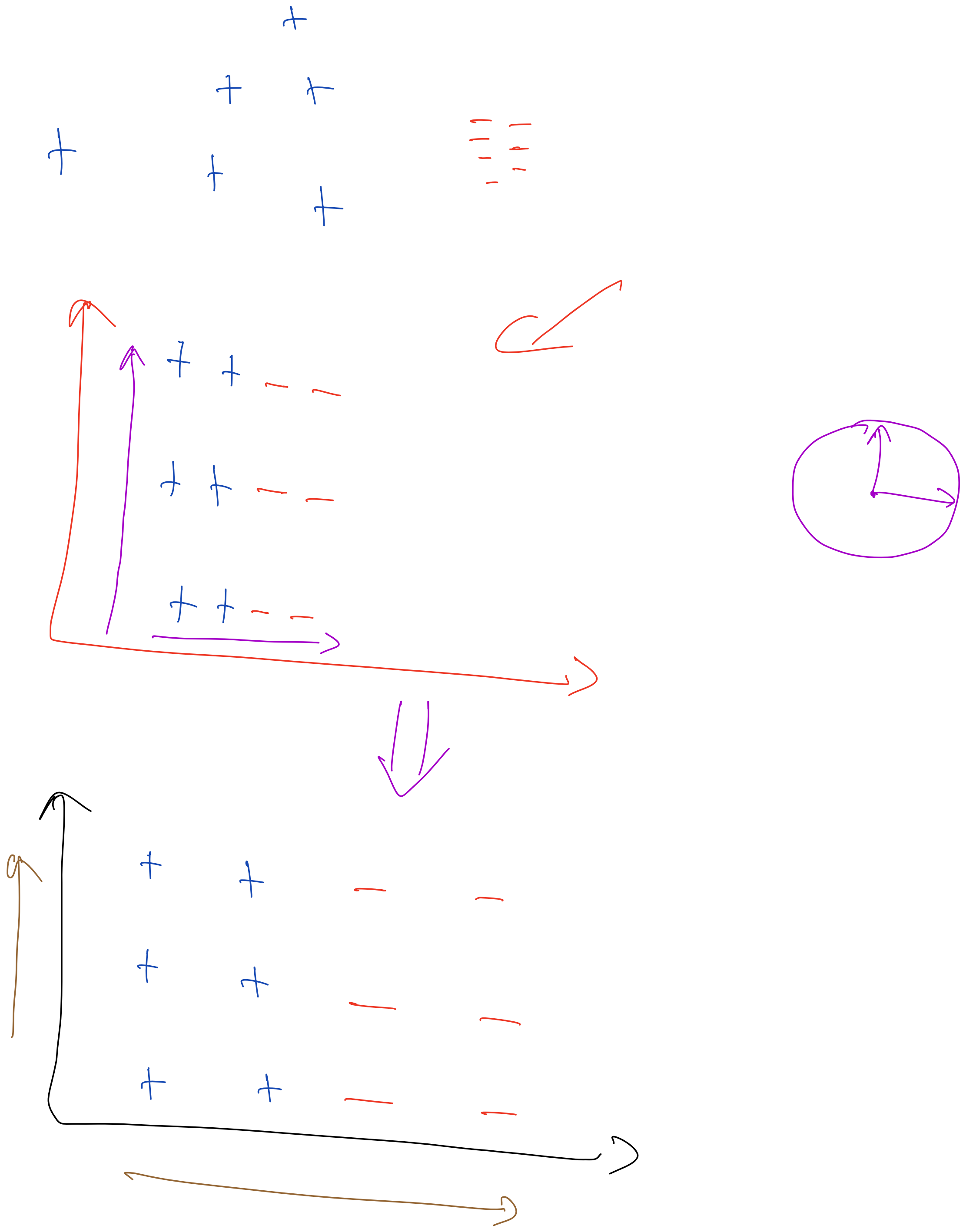
1) Blinkit Problem Statement
   ↳ issue with logistic regression
2) Geometric Intuition
3) kNN Algorithm
4) kNN scratch code
5) Assumptions of kNN
6) sklearn's kNN implementation
7) Bias- Variance trade-off
8) Train & Test time complexity
9) kNN for categorical data
10) LSH
11) kNN based imputation



k = 9

+ve → 3
0 → 3
-ve → 3

randomly choose 1 of them



x_{q1} , x_{q2}

x_{i1} , x_{i2}

$(x_{q1}-x_{i1} , x_{q2}-x_{i2})^2$

$[(x_{q1}-x_{i1})^2 , (x_{q2}-x_{i2})^2]$

$$S = (x_{q_1} - \underset{1}{x_{i_1}})^2 + (x_{q_2} - x_{i_2})^2$$

$$ED = sqrt\_S = \sqrt{(x_{q_1} - x_{i_1})^2 + (x_{q_2} - x_{i_2})^2}$$

classes_ counts

idx	classes	counts
0	2	3
→ 1	$\boxed{1}$	$\boxed{4}$
2	3	

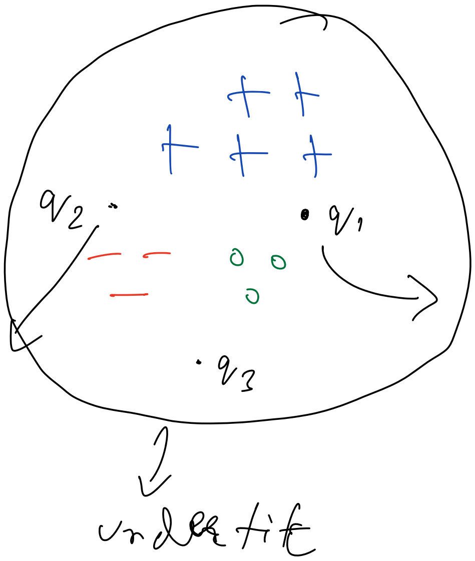
idx = 1 ↓ has max count

pred = classes [idx]

classes [1]

↓

value 1



underfit

$\boxed{K = 11}$

$q_1 = +$

$q_2 = +$

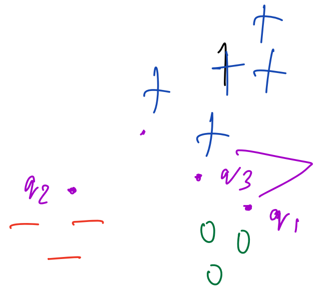
$q_3 = +$

+ve → 5

-ve → 3

0 → 3

11



K = 1 → overfitting

$q_1 = 0$

$q_2 = -ve$

$q_3 = +ve$

q_1 to q_3

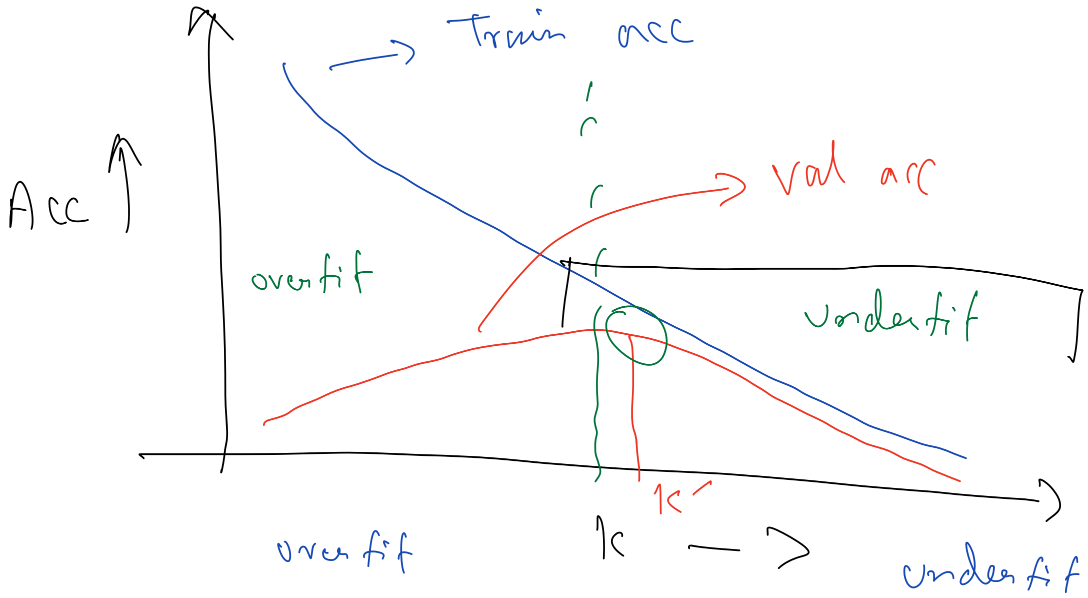
↓ ↓

0 +ve

slight change in query point → change in pred

k = large ⇒ underfit

k = small ⇒ overfit



$$\underline{E}D \longrightarrow \left[\sum_{i=1}^{d}\left(x_{q,i}-x_{j,i}\right)^{\boxed{2}}\right]^{1/\boxed{2}}$$

$$\begin{tikzcd}
ED \arrow[d] \\
(x_q, x_j) \arrow[d] \\
{[x_{q1},} \arrow[d, leftrightarrow] & x_{q2}, \arrow[d, leftrightarrow] & x_{q3}, \arrow[d, leftrightarrow] & \ldots & x_{qd}] \arrow[d, leftrightarrow] \\
{[x_{j1},} & x_{j2}, & & \ldots & x_{jd}]
\end{tikzcd}$$

Minkowski Distance

$$= \left[\sum_{i=1}^{d}\left(x_{q,i}-x_{j,i}\right)^{p}\right]^{1/p}$$

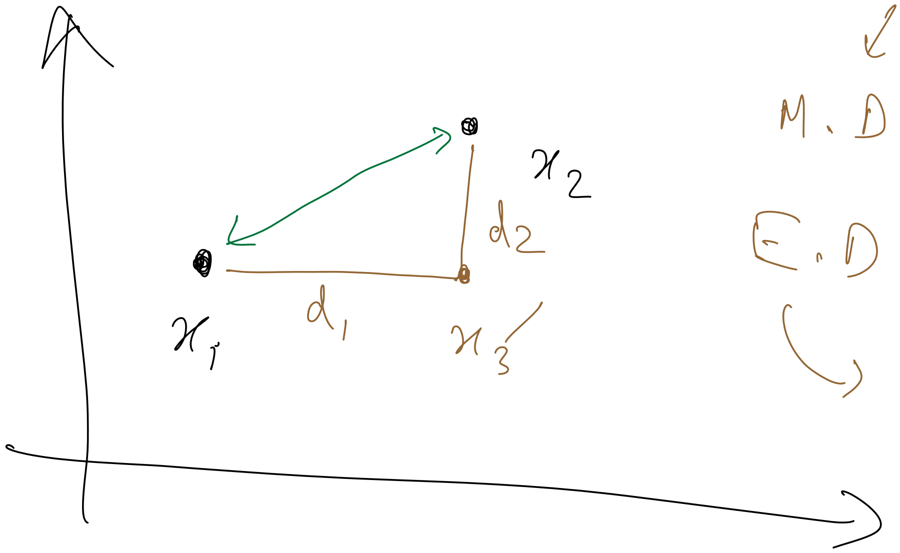
$l2$ norm

$\underline{p=2}$, $\Longrightarrow$ Euclidean Distance

$\underline{p=1}$, $\Longrightarrow$ Manhattan Distance

$$\xrightarrow[l1]{}$$

$$\left[\sum_{i=1}^{d} \left| x_{q,i} - x_{j,i} \right| \right]^{1}$$



L1 metric

$M.D = d_1 + d_2$

$E.D = \sqrt{d_1^2 + d_2^2}$

L2 metric

L1 norm $\Sigma |w_j|$

L2 norm Σw_j^2

N sample-points with d features

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix}$$

N rows, d columns

$N \times d$

Test - time

Dist Calculation $O(Nd)$ → for N data-points in the training set

Sort: $O(N \log N)$

choosing Top K: $O(k)$

histogram: (counts) $O(k)$

majority vote: $O(k)$

$$O(Nd + N \log N + \overline{k + k + k})$$

$$\boxed{k << N}$$

$d << N$?? X

↓

$$O(Nd + N \log N)$$

d is close to n

↳ is significant

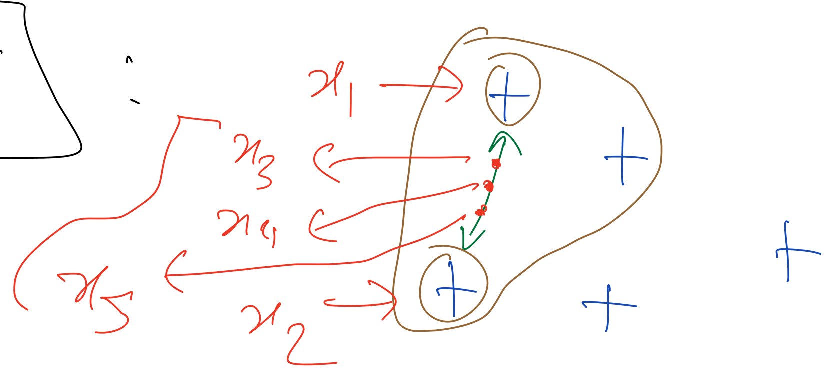
K is very small compared to N

Train time Complexity

$$O(1)$$

$\boxed{\text{SMOTE}}$:

$N = 5$
$K = 3$



$\boxed{K = 3}$

x_1 x_2 0.3

random value between 0 – 1

$$x_{new} = x_1 + 0.3 \times (x_2 - x_1)$$