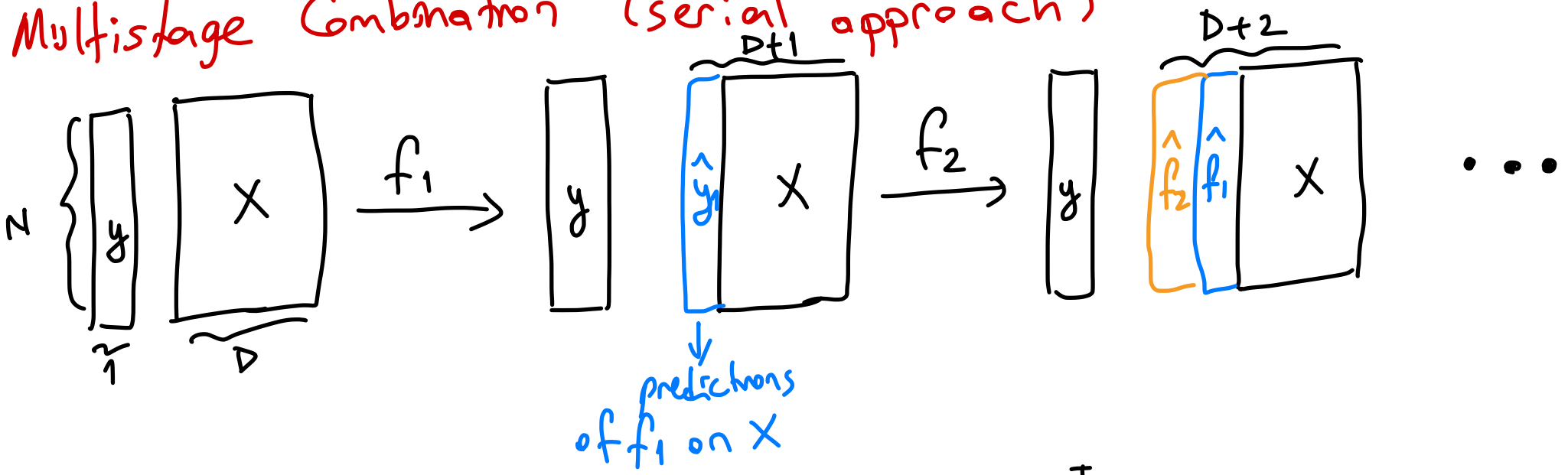
# Multistage Combination (serial approach)




$$x_{N+1} \rightarrow f_1(x_{N+1}) \rightarrow f_2([f_1(x_{N+1}), x_{N+1}^T]^T) \rightarrow \cdots$$

---

Let us say we have $L$ base learners

$f_j(x)$ $\quad f_1 \quad f_2 \quad \cdots \quad f_L$

$$\hat{y} = f(f_1, f_2, \cdots, f_L \mid \Phi)$$

↳ combination function

↳ combination parameters

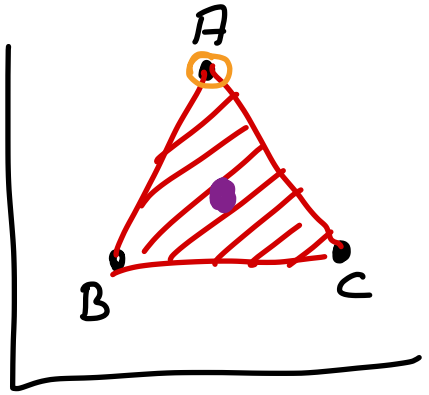
## VOTING

$$\hat{y}_i = \sum_{j=1}^{L} w_j f_j(x_i)$$

] linear opinion models, ensembles

Convex combination $\Rightarrow w_j \geq 0 \quad \forall j$

$$\sum_{j=1}^{L} w_j = 1$$

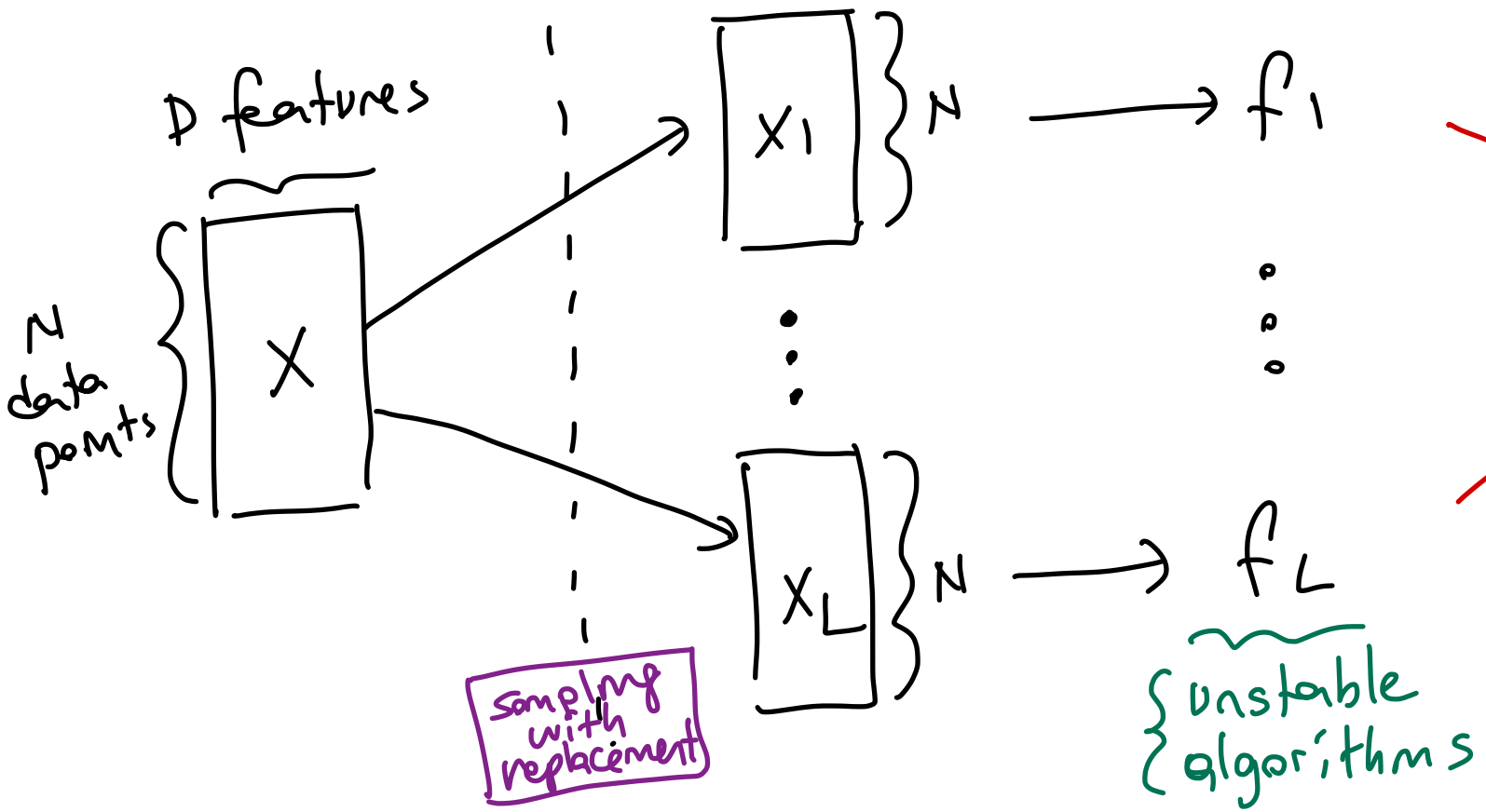



$w_A = 1$ $w_B = 0$ $w_C = 0$

$1/3$ $1/3$ $1/3$

Linear combination $\Rightarrow w_j \in \mathbb{R} \quad \forall j$

# BAGGING (Bootstrap AGGregation)



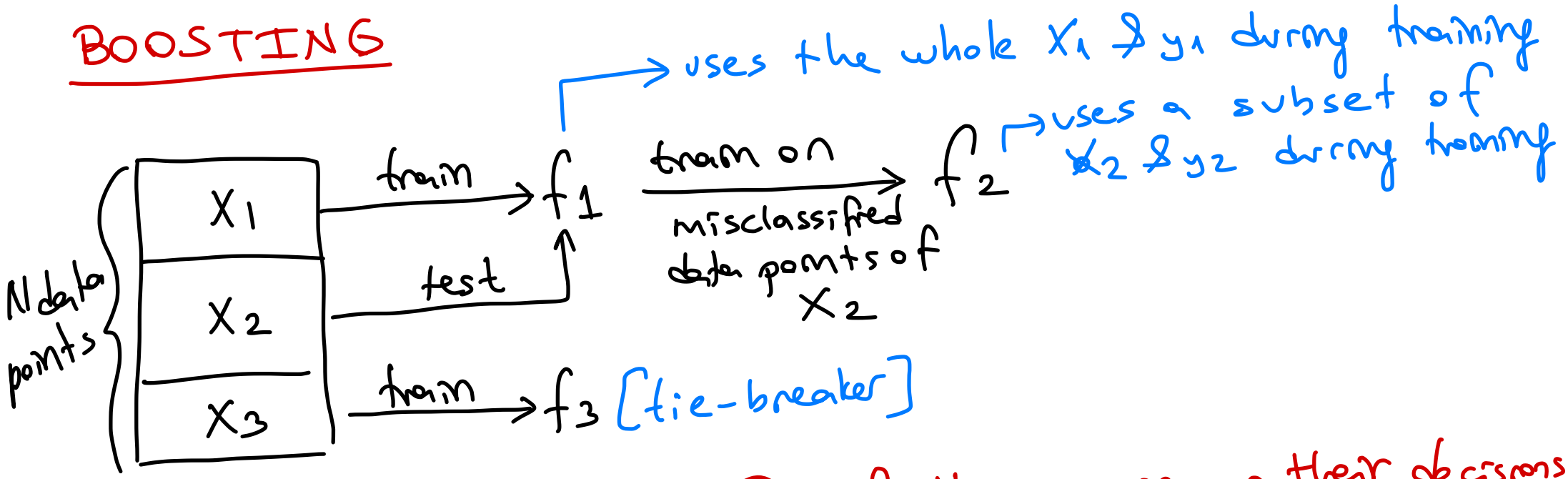

If N is very large, $N' < N$ data points should be picked so that training sets would become different enough.

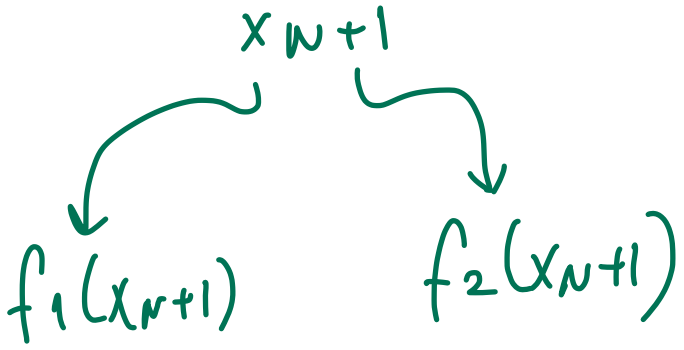
**Unstable Algorithm:** Highly affected by small changes in the training data set.

unstable ⇒ DT

stable ⇒ KNN

# BOOSTING




(1) If they agree on their decisions, no problem ⇒ use the predicted class label.

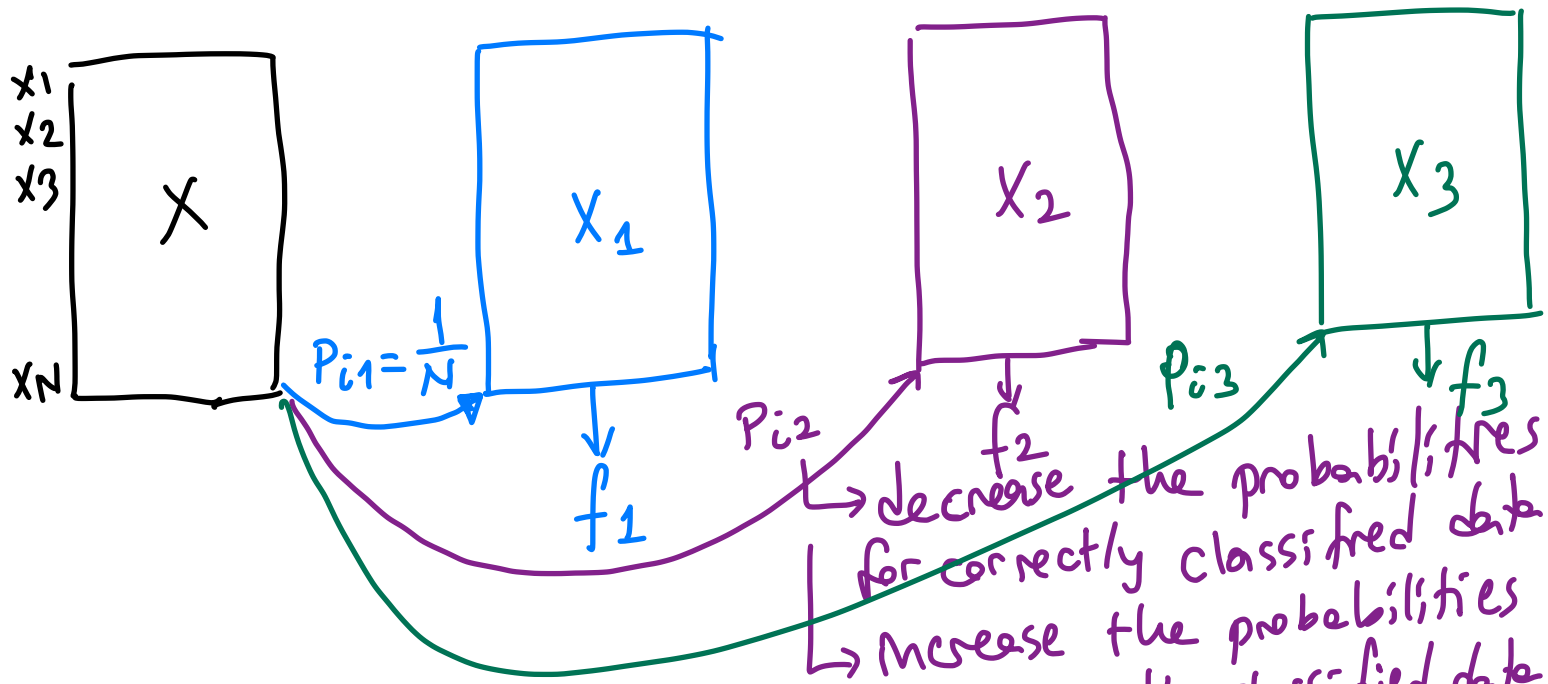
(2) If they do not agree, use $f_3(x_{N+1})$ as the predicted class label

**AdaBoost**: Modify the probabilities of drawing instances as a function of the error.

$P_{ij}$ = the probability that the data point $x_i$ is selected (used in training) by classifier $f_j$.

$$w_j = \log\left[\frac{1}{\beta_j}\right]$$

$$\beta_j = \frac{\epsilon_j}{1-\epsilon_j}$$




- decrease the probabilities for correctly classified data points
- increase the probabilities for incorrectly classified data points

• • •

50% accurate

$\beta_j = \frac{0.5}{1-0.5}$

$\log(1/\beta_j) = 0$

99% accurate

$\beta_j = \frac{0.01}{0.99}$

$\log(1/\beta_j) = \log(99)$

$$x_{N+1} \Rightarrow f(x_{N+1}) = w_1 f_1(x_{N+1}) + w_2 f_2(x_{N+1}) + \cdots + w_L f_L(x_{N+1})$$

based on their error rate

80% accurate

$\hookrightarrow \log(4)$

# Mixture of Experts (MoE)

Voting $\Rightarrow \hat{y}_{N+1} = \sum_{j=1}^{L} w_j f_j(x_{N+1})$ → constant over the input space

MoE $\Rightarrow \hat{y}_{N+1} = \sum_{j=1}^{L} w_j(x_{N+1}) f_j(x_{N+1})$

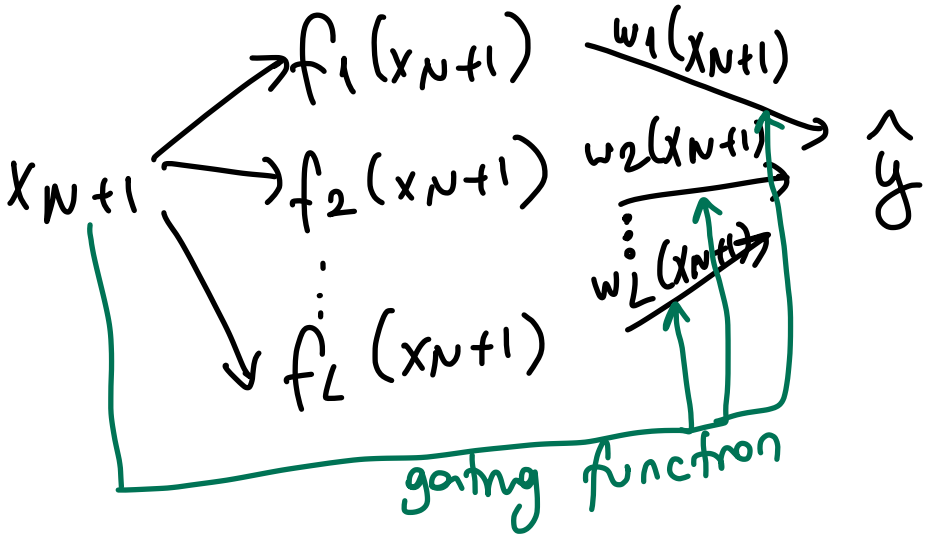
↳ $w_j$'s will assigned by the gating function

## Cooperative

- $w_1, w_2, \ldots, w_L$ are assumed to be independent.

⇓ sigmoid ⇓

$$w_j(x) = \frac{1}{1+\exp\left[-v_j^T x - v_{j0}\right]}$$

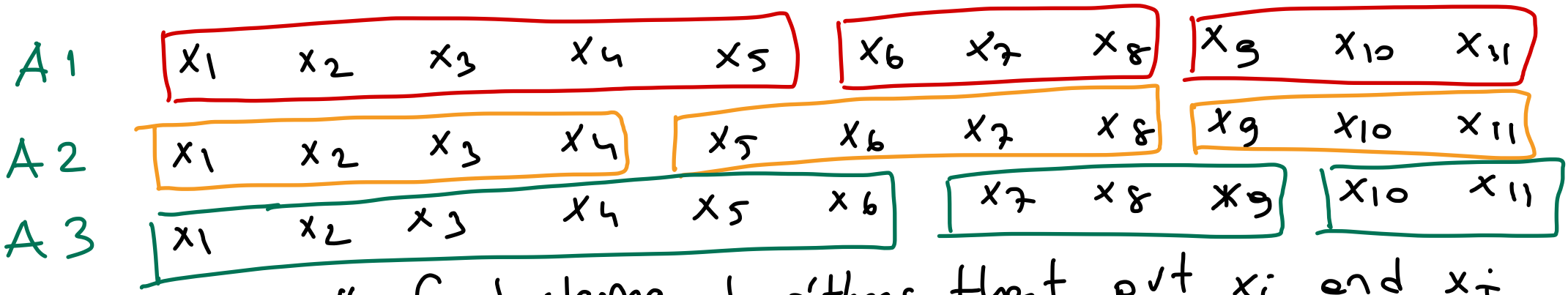



## Competitive

- $w_1, w_2, \ldots, w_L$ are producing sparse weights mostly zero
- one or some of them are nonzero

⇓ softmax ⇓

$$w_j(x) = \frac{\exp(v_j^T x + v_{j0})}{\sum_{k=1}^{L} \exp(v_k^T x + v_{k0})}$$




$C_{ij}$ = # of clustering algorithms that put $x_i$ and $x_j$ into the same cluster.

$C =$

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ | $x_9$ | $x_{10}$ | $x_{11}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_2$ | 3 | 3 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 3 | 3 | 3 | 3 | 2 | | | | | | |
| $x_4$ | 3 | 3 | 3 | 3 | 2 | | | | | | |
| $x_5$ | 2 | 2 | 2 | 2 | 3 | | | | | | |
| $x_6$ | | | | | | | | | | | |
| $x_7$ | | | | | | | | | | | |
| $x_8$ | | | | | | | | | | | |
| $x_9$ | | | | | | | | | | | |
| $x_{10}$ | | | | | | | | | | | |
| $x_{11}$ | | | | | | | | | | 3 | |

$11 \times 11$

⇒ clustering on the C matrix would give you A. ✱

Consensus clustering