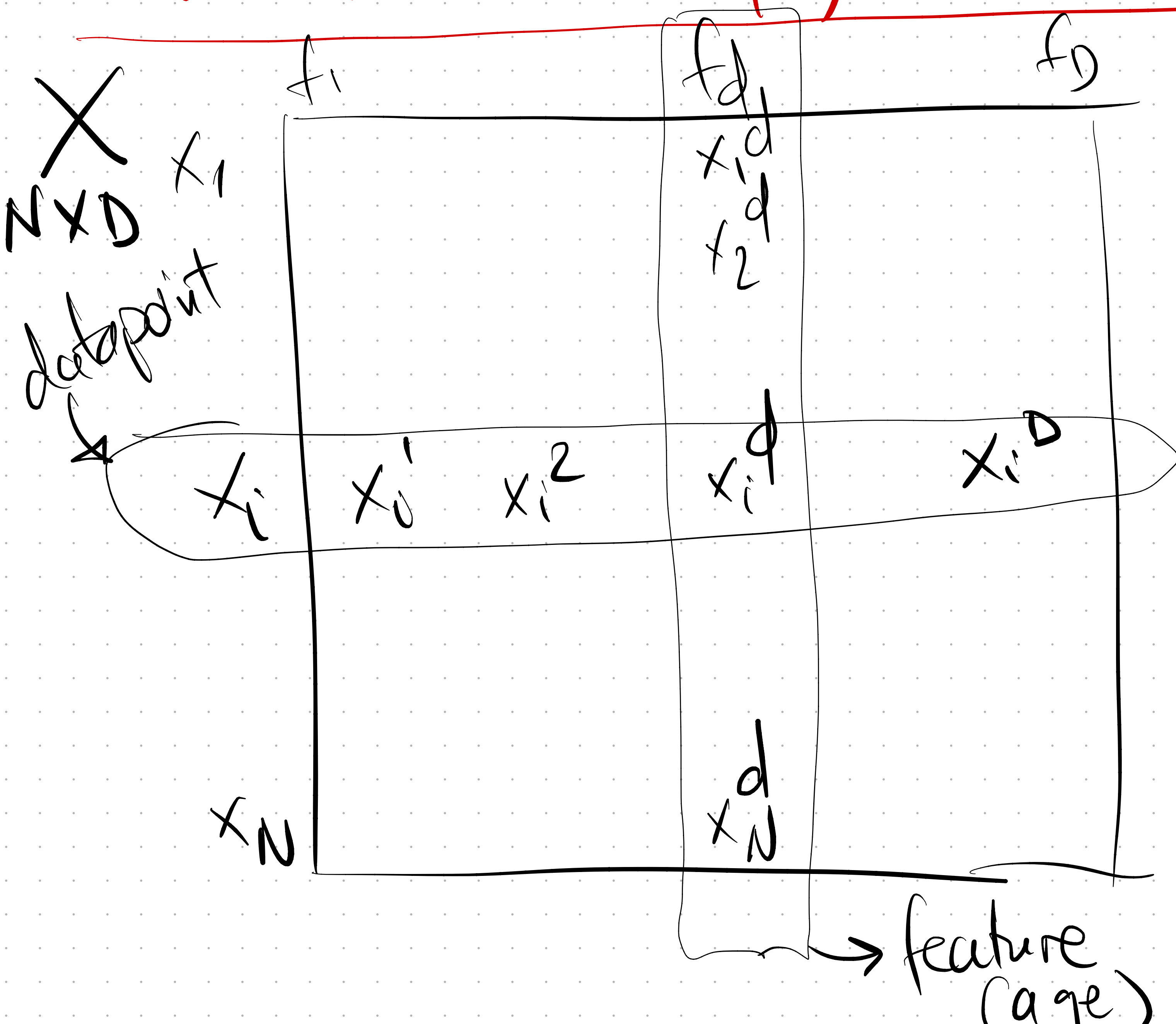
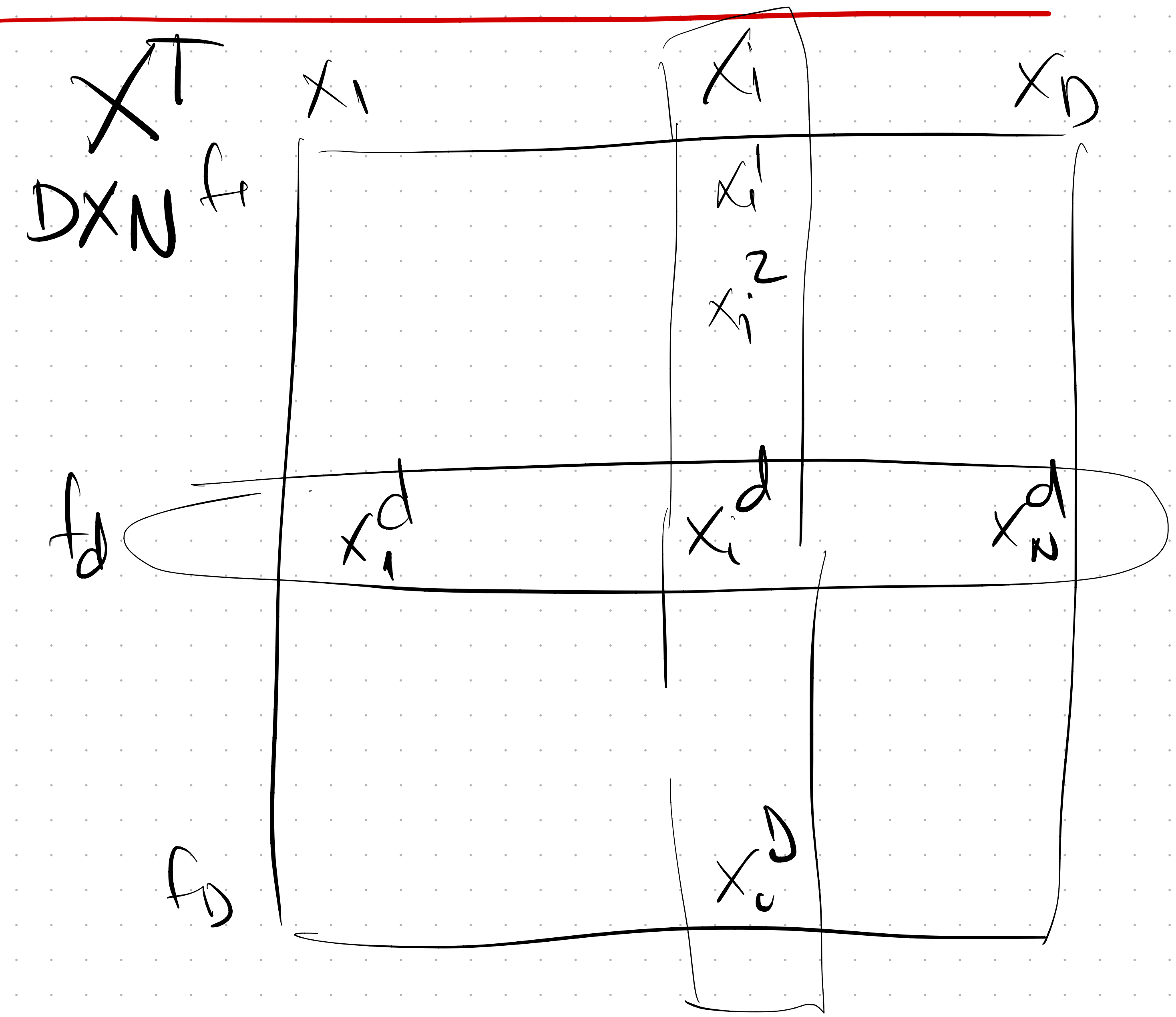
Lecture 5/16

- $X^T X$ matrix $\Rightarrow (X^T X + \lambda I)^{-1} \ldots$
- Linear Separability HW1 pb5 / vs Convex Hulls
- HW1 pb4: Entropy, Cond Entropy, Mutual Information





X [0-mean columns] $\Rightarrow$ X is $\boxed{\text{centered}}$

$$X = \begin{array}{ccc} x^1 & x^d & x^D \\ \hline & & \\ & & \\ \end{array}$$

$$\mu = \begin{pmatrix} \mu^1 & \mu^d & \mu^D \end{pmatrix}$$

D-dim

$\mu = \text{mean}(X)$

$\mu^d = \text{mean}(x\_column\ d)$

$$\xrightarrow{\text{CENTER}} \begin{bmatrix} x_1^1 - \mu^1 & x_1^2 - \mu^2 & \cdots & x_1^D - \mu^D \\ x_2^1 - \mu^1 & x_2^2 - \mu^2 & \cdots & x_2^D - \mu^D \\ \vdots & \vdots & & \vdots \\ x_N^1 - \mu^1 & x_N^2 - \mu^2 & \cdots & x_N^D - \mu^D \end{bmatrix}$$

$\mu_{(X\_CENTERED)} = 0$ on each col

X has 0-mean columns.

X centered. Look at $X^T X$ feat h

$$
X^T X = \begin{array}{c|c}
 & \text{feat } h \\
\hline
 & \begin{bmatrix}
\sum_{i=1}^{N} (x_i^1 \cdot x_i^1) & \cdots & \sum_{i=1}^{N} x_i^1 \cdot x_i^D \\
 & \ddots & \\
\text{feat } d \quad \cdots & \sum_{i=1}^{N} x_i^d \cdot x_i^h & \\
h \quad \cdots & \sum_{i=1}^{N} x_i^h \cdot x_i^h & \\
\sum_{i=1}^{N} x_i^D \cdot x_i^1 & \cdots & \sum_{i=1}^{N} x_i^D \cdot x_i^D
\end{bmatrix}_{D \times D}
\end{array}
=
\begin{array}{c}
 f \\
\begin{bmatrix}
\langle x^1 \cdot x^1 \rangle & \langle x^1 \cdot x^2 \rangle & \cdots & \langle x^1 \cdot x^D \rangle \\
d \quad \cdots & & \langle x^d \cdot x^f \rangle & \\
\langle x^D \cdot x^1 \rangle & & & \langle x^D x^D \rangle
\end{bmatrix}
\end{array}
$$

$\text{var}(x^1)$, $\text{var}(x^2)$, $\text{VAR}(x^h)$, $\text{VAR}(x^D)$

b in all (d, f) we have $x^d \cdot x^f =$
$\langle \text{feature\_d} \times \text{feature\_f} \rangle$
$= \text{Linear Correlation}\left(\begin{matrix} \text{feat } d \\ \text{feat } f \end{matrix} \right)$

if $\mu \neq 0$

$$\begin{array}{cc}
 & \begin{array}{cccc} \phantom{\sum_{i=1}^N (x_i^1-\mu^1)(x_i^1-\mu^1)} & d & h & \phantom{\sum_{i=1}^N (x_i^1-\mu^1)(x_i^D-\mu^D)} \end{array} \\
\begin{array}{c} \\ d \\ h \\ \\ \end{array} & \left[\begin{array}{cccc}
\sum_{i=1}^N (x_i^1-\mu^1)(x_i^1-\mu^1) & & & \sum_{i=1}^N (x_i^1-\mu^1)(x_i^D-\mu^D) \\
 & & \boxed{\sum_{i=1}^N (x_i^d-\mu^d)(x_i^h-\mu^h)} & \\
 & \boxed{\sum_{i=1}^N (x_i^h-\mu^h)(x_i^d-\mu^d)} & & \\
\sum_{i=1}^N (x_i^D-\mu^D)(x_i^1-\mu^1) & & & \sum_{i=1} (x_i^D-\mu^D)(x_i^D-\mu^D)
\end{array}\right]
\end{array}$$

$$= N \cdot \underset{D \times D}{\mathrm{COVAR}(X)}$$

where cell (d,h) is linear correl of feat d × feat_h if X centered ($\mu=0$) $\Longrightarrow$ $X^TX = \boxed{N} \cdot \mathrm{COVAR}(X)$

Very nice-math properties. $\mathrm{COVAR} \simeq X^TX$:

- symmetric $X^TX_{i,j} = X^TX_{j,i}$ = lin-correlation(sim) feat-i, feat-j

- pos semidefinite : $\forall$ vector $Z = (z_1 z_2 \ldots z_D)$: $\underline{Z^T(X^TX)Z \geq 0}$
  D-dim

  $\underline{\text{proof}}$: $Z^T(X^TX)Z = Z^TX^T \cdot XZ = (XZ)^T \cdot XZ \geq 0$

$e_i^T \cdot e_j = 0$

- $\underline{\text{Spectral decomposition}}$. $\begin{cases} e_1, e_2, \ldots e_D \to \text{column eigenvectors} \\ \alpha_1 \;\; \alpha_2 \;\; \ldots \;\; \alpha_D \to \text{corresp eigenval} \end{cases}$

- Eigenvectors Orthonormal $\langle e_i^T e_i \rangle = 1$

$(X^TX) e_j = \alpha_j \cdot e_j \Longleftarrow$

$$X^TX = \underbrace{\begin{bmatrix} | & | & & | \\ e_1 & e_2 & \cdots & e_D \\ | & | & & | \end{bmatrix}}_{E} \cdot \underbrace{\begin{bmatrix} \alpha_1 & & & 0 \\ & \alpha_2 & & \\ & & \ddots & \\ 0 & & & \alpha_D \end{bmatrix}}_{\mathrm{diag}(\alpha)} \cdot \underbrace{\begin{bmatrix} — & e_1 & — \\ — & e_2 & — \\ & \vdots & \\ — & e_D & — \end{bmatrix}}_{E^T}$$

D columns, D rows

Kernel Matrix

$$\underset{N \times D}{X} \cdot \underset{D \times N}{X^T} = \underset{N \times N}{K}$$

2 datapoints

cell ij: $\boxed{K_{ij} = \langle x_i \cdot x_j \rangle = \text{linear similarity}(x_i, x_j)}$

$$\begin{array}{c|ccccc} & 1 & & j & & N \\ \hline 1 & x_1 \cdot x_1 & x_1 \cdot x_2 & & & x_1 \cdot x_N \\ & & & & & \\ i & x_i \cdot x_1 & \cdots & x_i \cdot x_j & \cdots & x_i \cdot x_N \\ & & & & & \\ N & x_N \cdot x_1 & & & & x_N \cdot x_N \end{array}$$

$\|x_1\|^2$ $\|x_2\|^2$ $\|x_N\|^2$

$N \times N$

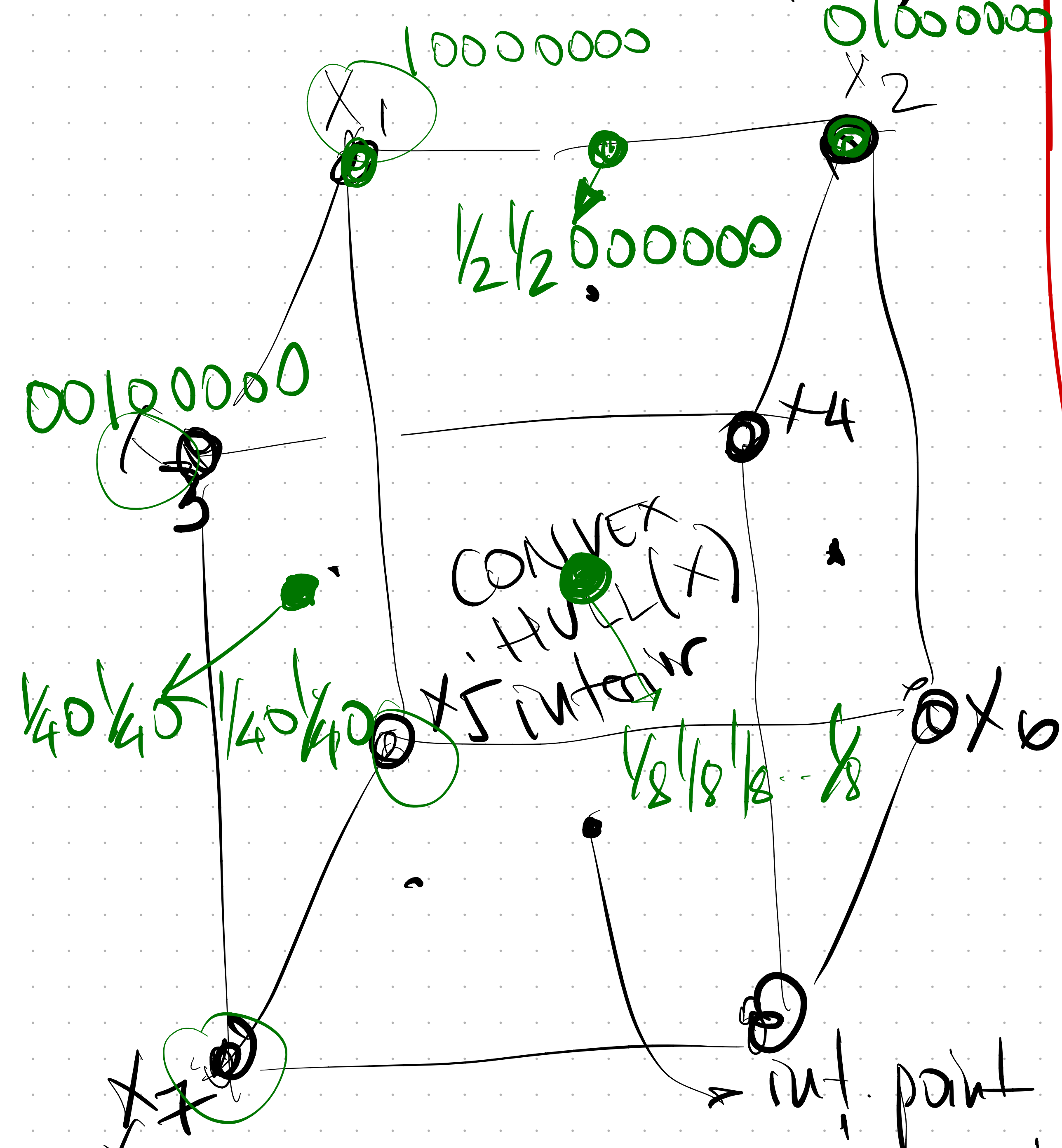
→ pairwise similarity of all datapoints

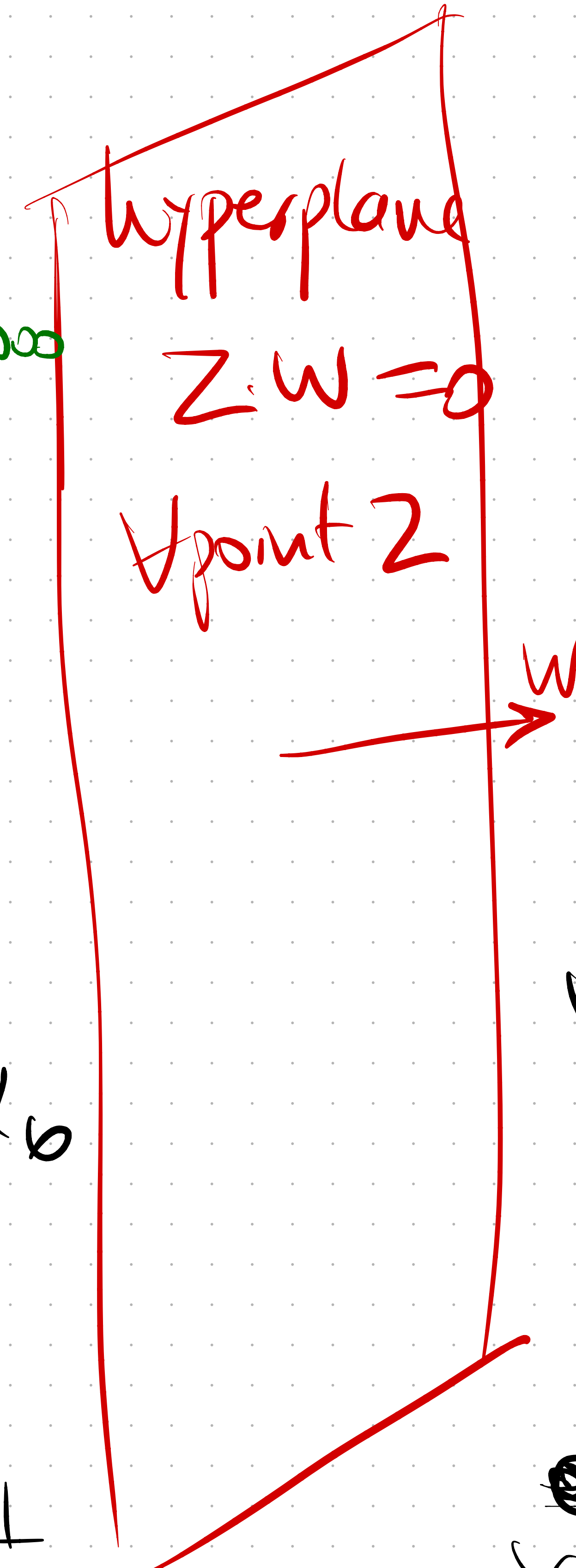
ALGEBRA: $X \cdot X^T$ has same property as $X^T \cdot X$ (matrix · matrix^T)

- symm
- pos def
- eigen spectral decomp

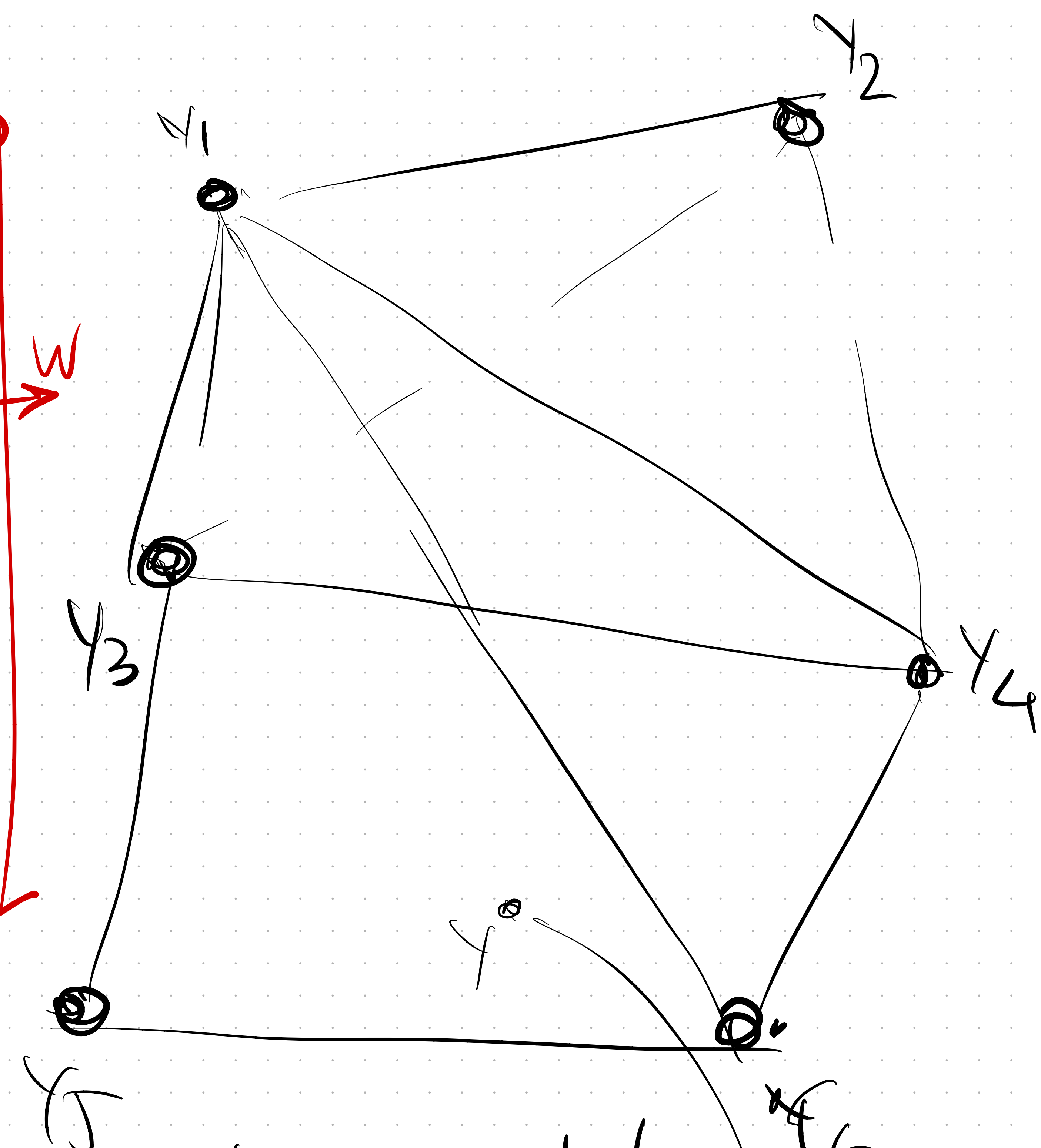
HW1 PB5:

X = set of vectors (3-D)





Y = set of vectors



$x = \sum_{i=1} \alpha_i x_i$

convex polyhedra

α-dist

$\sum \alpha = 1$

all α possib

$CH(X) = \{\forall \alpha \text{ weights } \sum \alpha_i x_i\}$

convex polyh

$CH(Y) = \{\forall \alpha \text{ weights } \sum \alpha_i Y_i$

A) either $CH(X)$, $CH(Y)$ donot intersect
or $\{x_i\}$ $\{y_j\}$ are linearly separable

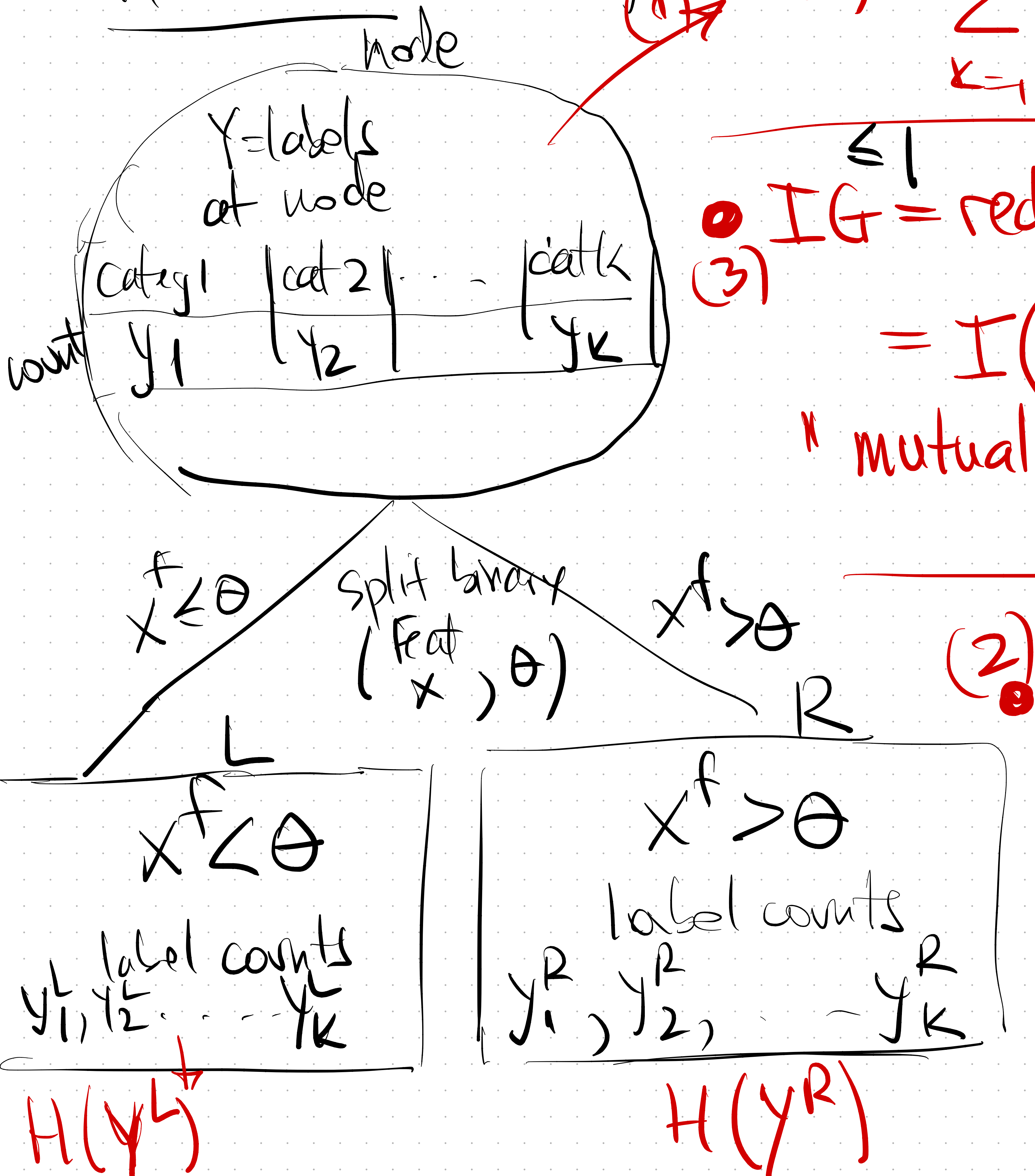
Hyperplane sep: one side X, the other side Y

cannot have both
- lin separation by hyp.
- $CH(X) \cap CH(Y) \neq 0$
  intersection.

B) one of the two must happen

if $CH(X) \cap CH(Y) = \emptyset$ $\Longrightarrow$ there is a separator hyperplane
do not intersect

HW1 PB4 K categ



(1) $H(Y) = \sum_{k=1}^{K} y_k \log \frac{1}{y_k}$ (y = normalized as distrib)

entropy before split

≤ 1

(3) IG = reduction in Entropy $= H(Y) - H(Y|X)$

$= I(X,Y) = \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)}$

"mutual information"

(2) entropy after split

$\#(x^f \le \theta) \cdot H(y^L) + \#(x^f > \theta) \cdot H(y^R)$

$\stackrel{?}{=} H(Y|X)$ cond entropy

labels | feat

$= \sum_{y,x} p(y,x) \cdot \log \frac{p(y,x)}{p(x)}$

joint

just x dist

- prove $I(X,Y) \stackrel{?}{=} H(Y) - H(Y|X)$

$$\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \stackrel{?}{=} \sum_{y} p(y) \cdot \log \frac{1}{p(y)} - \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)}$$

<u>hint</u>: $\sum_{x} \underbrace{p(x,y)}_{\text{joint}} = \underbrace{p(y)}_{\text{marginal}}$

- $I(X,Y) = H(Y) - H(Y|X)$

symmetric ?

$= H(X) - H(X|Y)$ - - - - $\Rightarrow$ pb 4

X = binary split (2 branches)