⊙ Supervised learning :-

→ Classification (labels)

→ pictures of cats & dogs ≡ labels {cat, dog}

→ image (100 x 100 pixels ~ 10^4 single precision numbers to denote the image)

⇒ D ≡ dimension of each pt. in my dataset.
= 10^4 (for 100x100 pixels)

⇒ X ≡ data itself, for eg, images

→ $X \in \mathbb{R}^D$

→ x^i ≡ data points
≡ for eg, for 1000 images $i = 1, 1000$

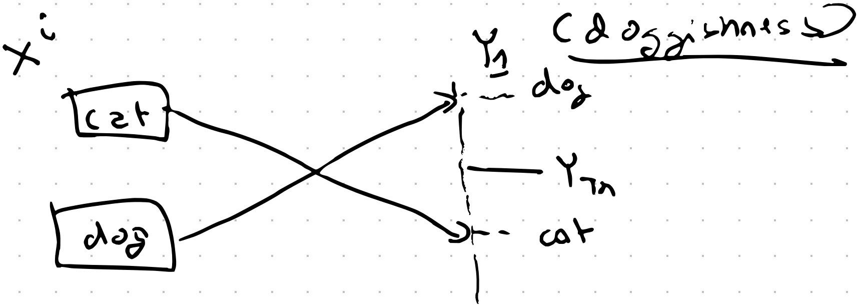
→ N ≡ no. of data points

→ in SL, we map each data pt. in a single variable to perform classification.

→ let's assume another parameter, $\boxed{Y \text{ (dogginess)}}$

→ each image (x^i) is mapped to a value of (Y), for cats, Y is low & for dogs, Y is high.

→ we put a threshold value of Y, Y_{TH}, s.that
$Y > Y_{TH} \equiv$ dog
$Y < Y_{TH} \equiv$ cat



⇒ $Y \equiv$ reduced representation of original data
$\equiv$ each data pt, x^i, is mapped to a single real no. y^i

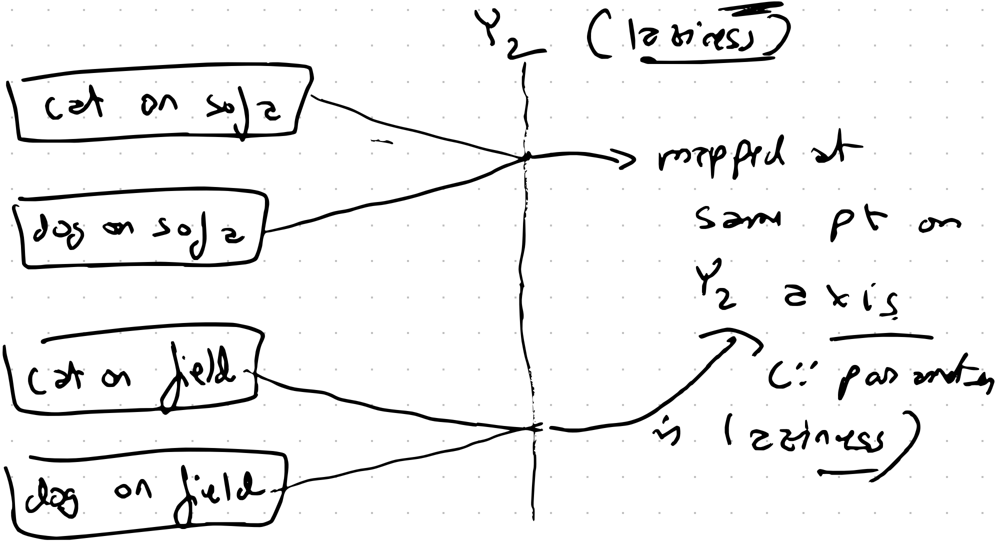
→ Here, y^i ≡ single $\mathbb{R}$ no.

→ In general $\boxed{Y^i = f_\pi(x^i)}$

NOT the label itself, but rather the output of the model

⇒ let's consider another parameter

Y_2 (laziness) ≡ whether animal was photographed on the sofa / on the grass



→ mapped at same pt on Y_2 axis

(∵ parameter is laziness)

⇒ TL;DR :- many different classification tasks that one can address.

→ another parameter, $Y_3 \equiv$ orientation of the image

$\equiv$ front-view/side-view

⇒ [Intrinsic Dimension] of the dataset ↳

→ no. of independent classification tasks that one can meaningfully perform on a dataset. (informal defn)

(i) independent ≡ if all dogs are on grass & all cats are on the sofa, then classification by laziness is same as " " species. So we only count variables with low mutual information.

(i) meaningfully ≡ very low sample size, that data which is under-represented, cannot be used to train the model meaningfully.

→ ID is related to semantic complexity of the input dataset.

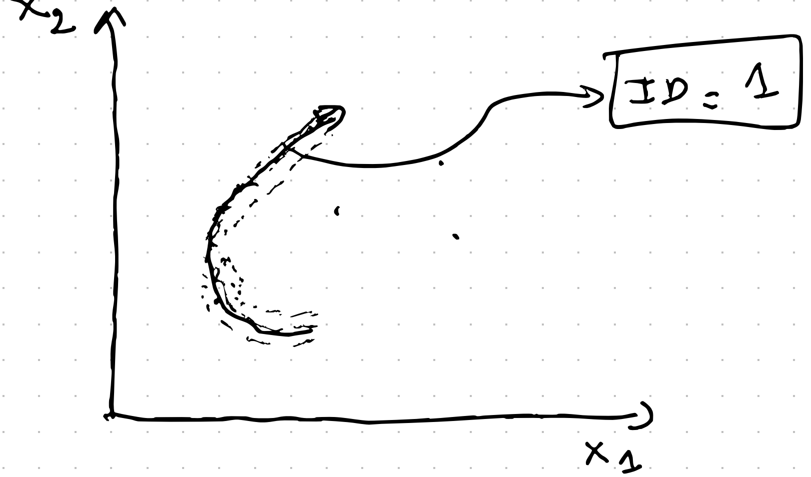
→ From a very qualitative perspective, if ID is large, we can gather many diff. kinds of information from the dataset.

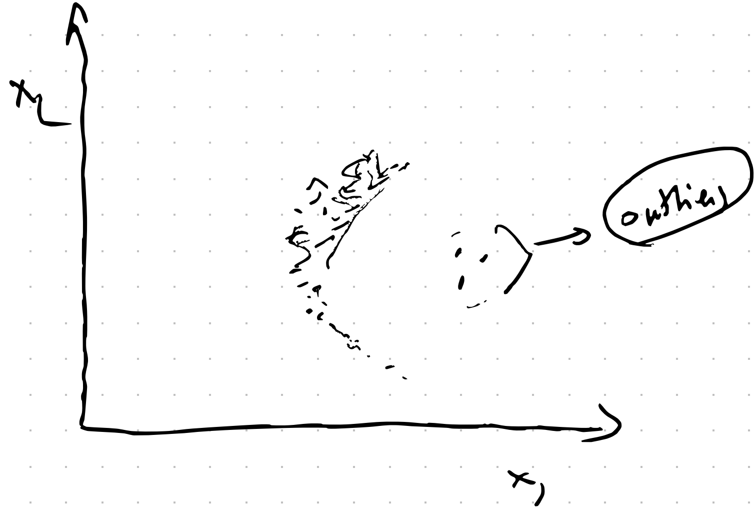
→ How many independent directions do I need to describe my dataset?

⇒ entry to USL from SL, ∵ to answer this question, labels aren't necessary.

→ $x^i \in \mathbb{R}^D$

(2 co-ords: x_1, x_2; $D=2$)





→ ID = 2 here ∵ there are more data pts. along one Axis.

→ So, in practicality ID depends on the SCALE of the system.

⇒ $\boxed{ID \equiv d}$ → min. no. of indep. variables,

⇒ outliers are ignored. In above image, The 3 outliers DO NOT make $d = 2$ ∵ there are too few to train the model.

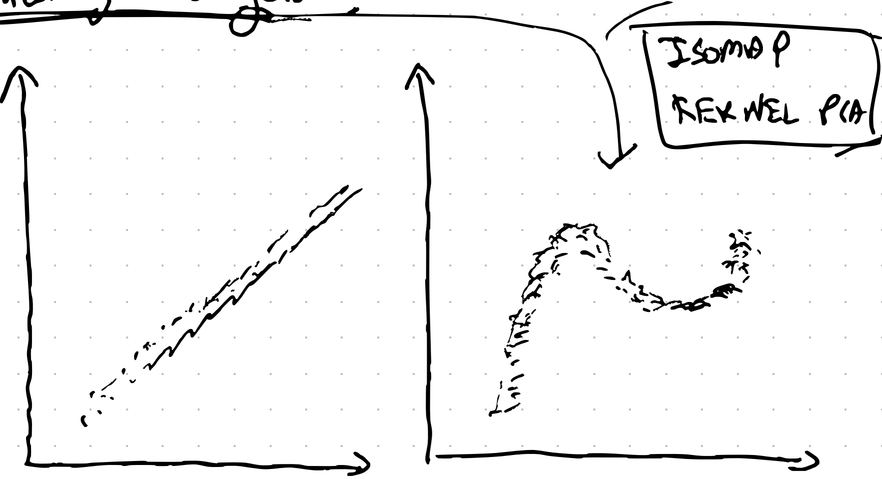
□ Task 1 :- estimating the ID :-

→ one can meaningfully talk about ID if APPROXIMATELY, the ID is scale invariant.

□ Task 2 :- Finding explicitly a set of 'd' coordinates describing my dataset!

$$x^i \rightarrow f(x^i) = Y^i, \quad Y^i \in \mathbb{R}^d, \quad x^i \in \mathbb{R}^D$$

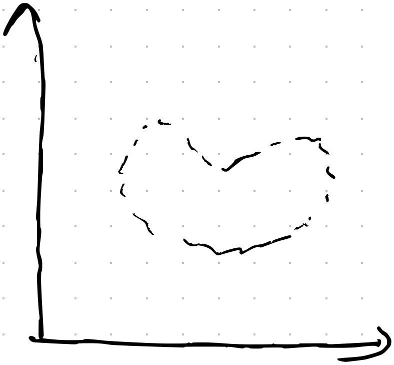
⇒ The manifold containing the data is called "embedding manifold"



PCA

topologically equivalent to a hyperplane.

→ ISOMAP & Kernel-PCA help resolve the embedding manifold problems ∴ in this figure, $D = 1$, but it will be shown as $D = 2$ due to projection, but ∵ it is on a single manifold, $D = 1$ actually.

⇒ 

Kernel PCA & N.N. (Auto-encoders)

→ task ② is impossible in this specific case.

→ mapping a cylinder on 2-D plane opens up the shape, & def^n of distance changes, for eg, taking PBC into account.

⇒ Task ② is possible if data manifold is on a hyperplane OR isomorphic to a hyperplane.

(i) <u>Corrected task ②</u> :-

Finding explicitly a set of $(\tilde{d})$ coordinates describing any data set, with

$\tilde{d} \geq d$.

→ i.e., the total no. of dimensions will be > than the ID.

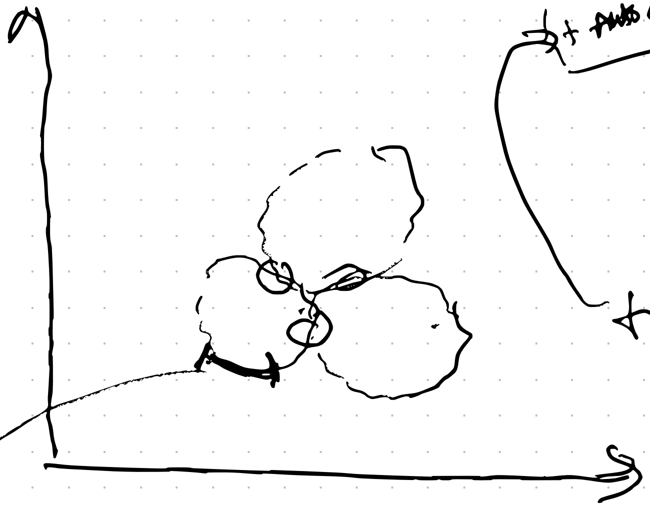
This is the corrected statement of task ② & can be applied to any manifold, regardless of manifold being equivalent / isomorphic to a hyperplane

⇓

Dimensional reduction of dataset <u>without</u> <u>information loss</u>

<u>OR</u> <u>practically</u> with minimum information loss

□ Real world data :- (it's like foam)



[+ autoencoders]

Kernel PCA (50-50 performance)

+ Neural networks (best performance)

→ 2D representation of very high-D space.

→ explicit " is only found locally.

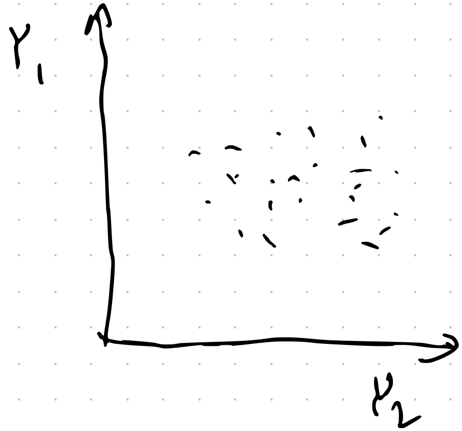
→ on a global level, we run into too many problems with topology of the dataset.

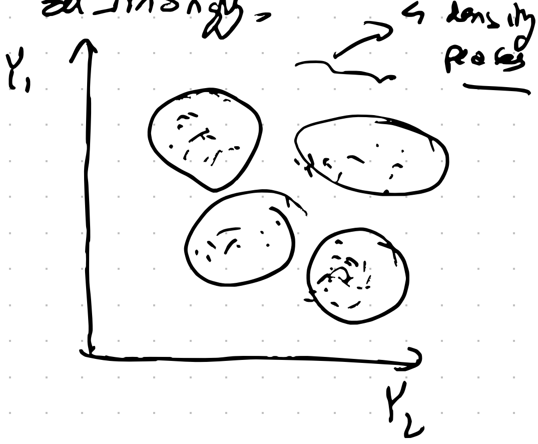
→ to find meaningful low-D representation of data, | $f(x^i)$ = highly complex & nonlinear.

→ The flavors of NNs we specifically use are auto-encoders

→ Going back to e.g. of cats & dogs

$Y_1 \equiv$ doggishness
$Y_2 \equiv$ laziness
⇒ these 2 should be correlated only very mildly, not at all strongly.





↓

both of these have $d = 2$, but which is more likely in real-world scenario?

→ RHS is more likely ∵ we have a gap (i.e, very few data pts) where we can't determine if its a dog/cat.

LHS is evenly distributed data which is very unlikely IRL.

→ How do we distinguish btw LHS & RHS?

(·) Task 3 :- Estimating the probability density.

⇓

→ trivial in 2D

→ impossible in $\mathbb{R}^D$ space due to computational issues

→ ∴ density can be meaningfully defined & computed ONLY on the embedding manifold.

$p(Y)$ ≡ prob. density as a fn^2 of Y, :- otherwise it can't be estimated numerically.

→ we will find density of the embedding manifold, without explicitly finding coordinates Y.

Task 5 :- Clustering / Recovering clusters in the data.

clustering :- groups of data points that are similar / close to each other, i.e., of relatively high density, but far from other such groups.

⇒ Implicit assumption for all these tasks :-

data pts are harvested from the same probability density.

→ A major exception to this assumption is :-

→ time-series analysis

~~for eg~~, MD trajectories

→ USL techniques for time-series analysis / time-ordered data points

⇓

→ Markov-State Modelling

→ Time-lagged independent PCA

(generalization of PCA)

{ useful for determining autocorrelation across time }