Recap - Logistic Regression

$$f_w(x) = \frac{1}{1+e^{-w^T x + w_0}} \Rightarrow P(Y=1|X)$$

$$\arg\min_w \sum_{i=1}^{n} \left[y_i \log(f_w(x_i)) + (1-y_i)\log(1-f_w(x_i)) \right]$$ [Cross entropy]

parameters: w (w_0)

Assumption: $y_i \in \{0, 1\}$

Binary prediction

- This is a convex optimization problem
- No closed form solution. Must use gradient descent.

Linear Regression: $w = (X^T X)^{-1} X^T y$

See Intro to ML textbook

$$w_0^{t+1} = w_0^t - \lambda \sum_{i=1}^{n} (f_w(x_i) - y_i)$$

$$w_j^{t+1} = w_j^t - \lambda \sum_{i=1}^{n} (f_w(x_i) - y_i) x_i^{(j)}$$

$w^{t+1} = w^t + \Delta w$

$\lambda \rightarrow$ learning rate

> 0

$t \Rightarrow$ training step iteration

$$f_w(x) = \frac{1}{1+e^{-w^T x}}$$

$$f_w(x) = w^T x + b$$

When to stop learning?

- Convergence in w^t

$$\frac{\|w^t - w^{t-1}\|_2^2}{\|w^t\|_2^2} < \varepsilon$$

$f_w(x) =$

- Convergence in cross entropy

$w = 0$

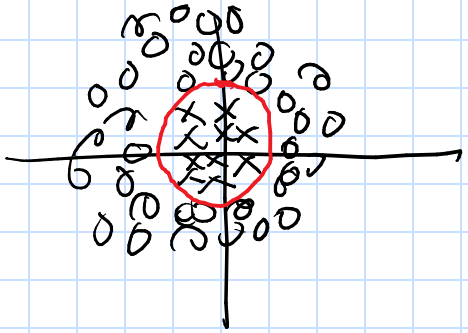
$w \sim N(0, \tau)$

Cover's Thm (1965)

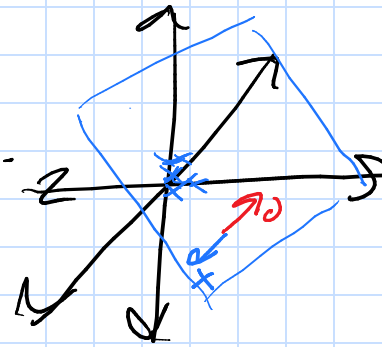
A complex pattern classification task, cast in a high dimensional space non-linearly, is more likely to be linearly separable than in the low-dimensional space.

* Provided the space is not densily populated

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



$$z = \Phi(x) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}$$



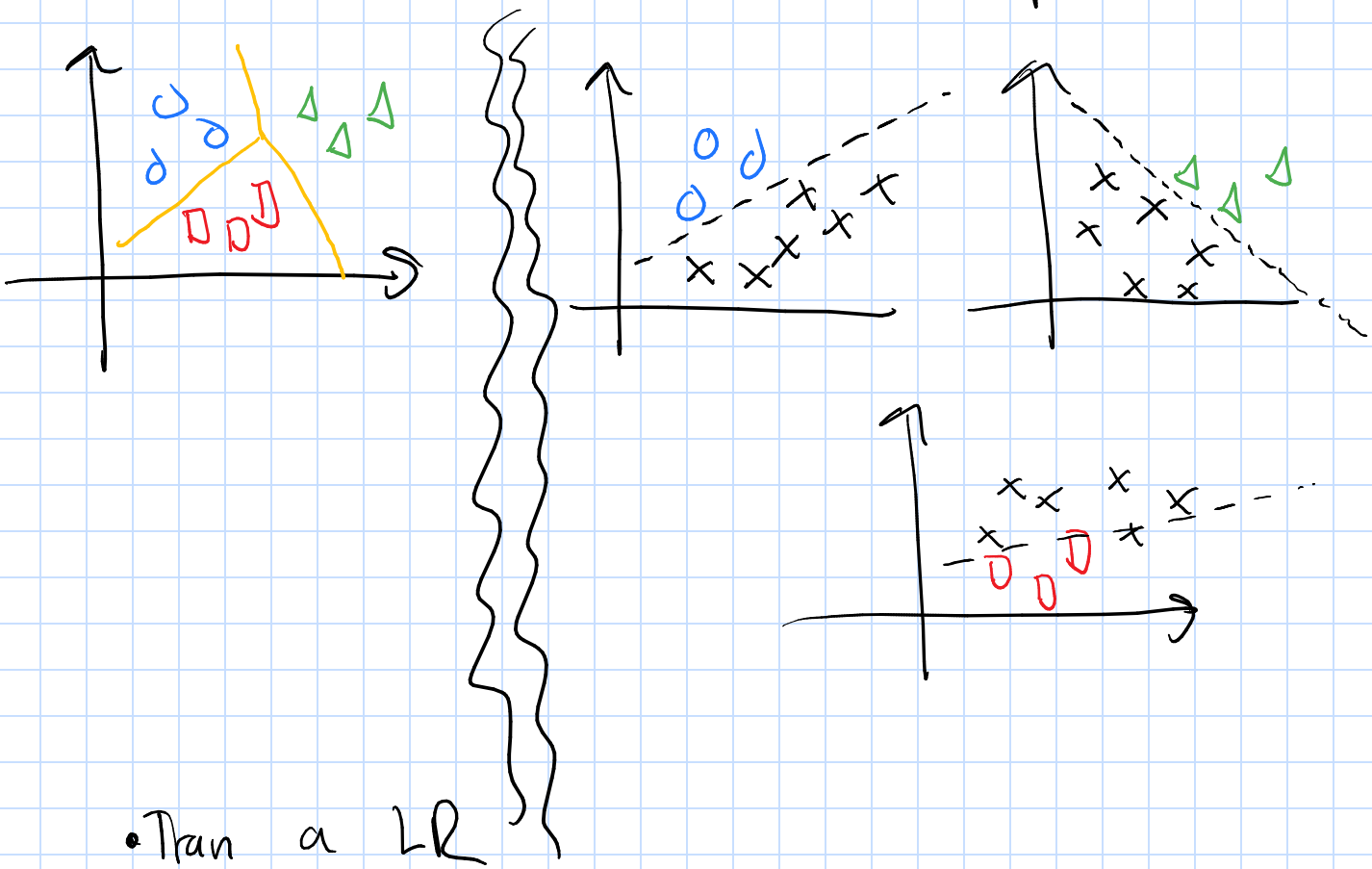
Sklearn → PolynomialFeatures

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \xrightarrow{\Phi} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1x_2 \\ x_1^2 \\ x_2^2 \\ x_1^2x_2 \\ x_2^2x_1 \\ \vdots \end{bmatrix}$$

Multiclass support

Assume that we have a multi-class dataset

- Split the data into two class problems



- Train a LR classifier on each binary fast

$$P(y = c \mid x; w_1, w_2, \ldots, w_C) = \frac{\exp(w_c^T x)}{\sum_{q=1}^{C} \exp(w_q^T x)}$$

C - classes

[soft max function]