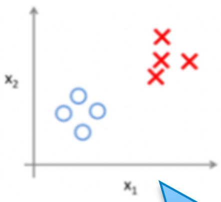
Imperial College
London

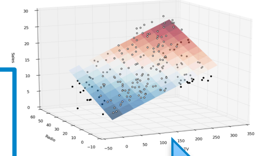
# Supervised vs Unsupervised Learning

**MACHINE LEARNING**

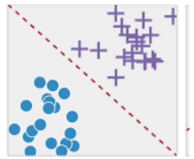


**SUPERVISED LEARNING**
**Training Set:**
**Data x with labels y**

Regression

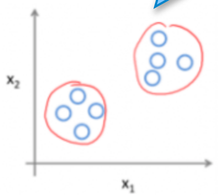


Classification



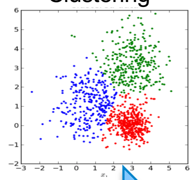
**Regression**
(y is real number/vector)

**Classification**
(y is class or integer number/vector)

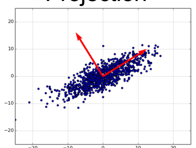


**UNSUPERVISED LEARNING**
**Training Set:**
**Just Data x**

Clustering



Projection

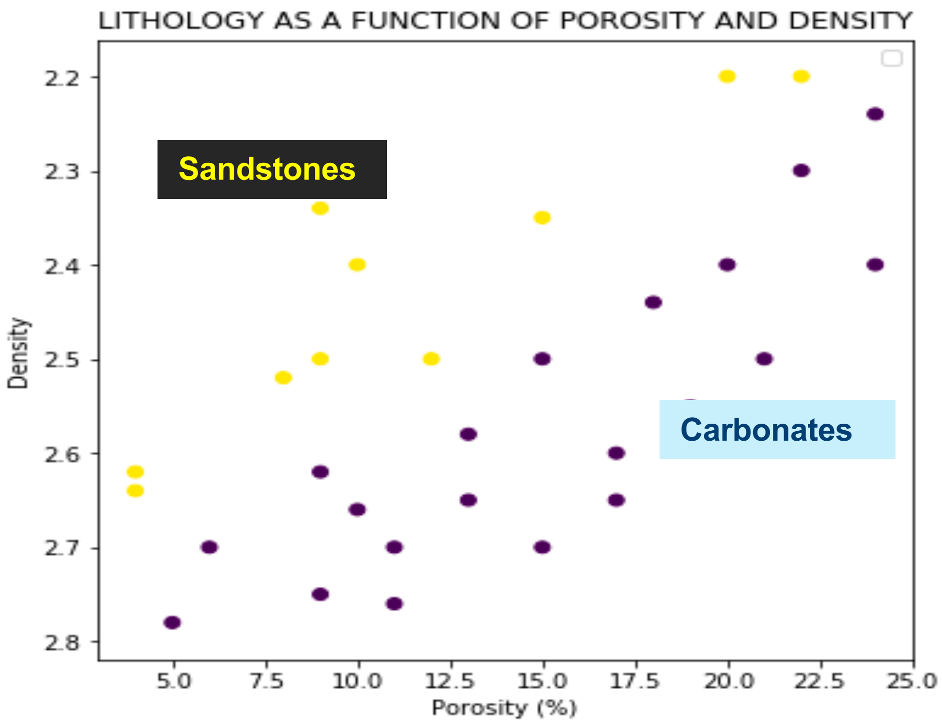


**Clustering**
(group x's in clusters)

**Dimensionality Reduction**
(compress x's)

Imperial College London

# Logistic Regression Problem




30 rock samples with Porosity, Density and Lithology as a label.

We wish to predict Lithology from Density and Porosity: this is a Supervised Classification problem .

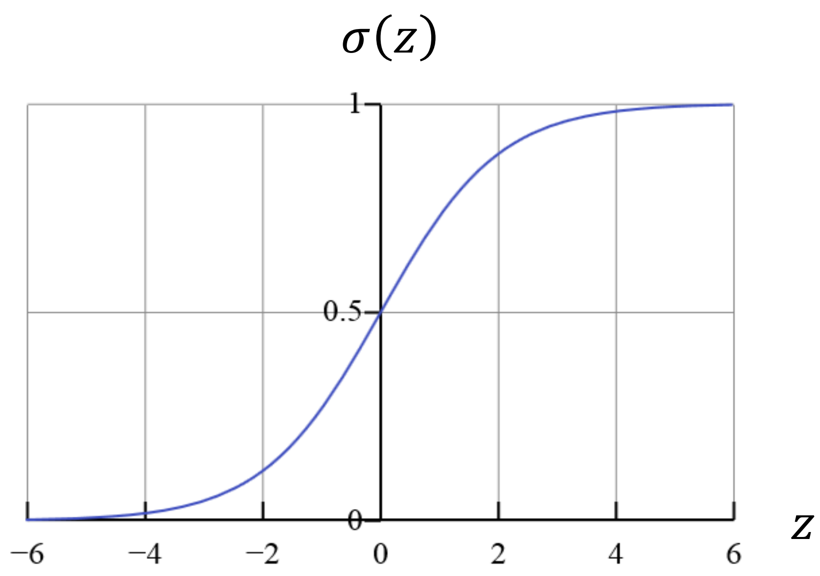
*Cannot use Linear Regression : need a transform from the domain of real values to the 0 or 1 indicator*

Imperial College London

# Sigmoid Function for transformation to [0,1] domain.

- It is also called the **Logistic** Function

- It takes any real value $z$ and transforms it into a value between 0 and 1




$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

It is easy to prove that

$$\sigma'(z) = \big(1 - \sigma(z)\big)\sigma(z)$$

# Interpreting the output of the Logistic function

- Say that $y$ is the outcome of a regression equation for an input $x$

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

- In vector form:

$$y = \theta^T x$$

- $y$ is a real number which can take any positive or negative value.

- If we apply the sigmoid (or logistic) function $\sigma(y)$ we obtain a value between 0 and 1, which **we interpret as the probability for the class to be 1**

$$h_\theta(x) = P(y = 1|\theta, x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Imperial College London

# Cost function for a Training whole Set: $\left(\left(x^{(i)}\right),\left(y^{(i)}\right)\right) i = 1$ ...m

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$ = **<u>Cross-Entropy</u>**

To minimize, just calculate the derivatives $\frac{\partial J(\theta)}{\partial \theta_j}$ for $\mathrm{j} = 1 \ldots . \mathrm{n}$ and apply Gradient Descent

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left[h_\theta\left(x^{(i)}\right) - y^{(i)}\right]x_j^{(i)}$$

*Exercise: prove this relationship!*

*In spite of its name, Logistic Regression is for <u>Classification</u> rather than Regression!*

Imperial College London

# Logistic Regression at Test time:

The vector of weights $\theta$ has been calculated at the Training stage.

Now, for any new point $x$ for which we only know the feature vector (and not the label)

$$x = (x_1, x_2, \dots, x_m)$$

We calculate

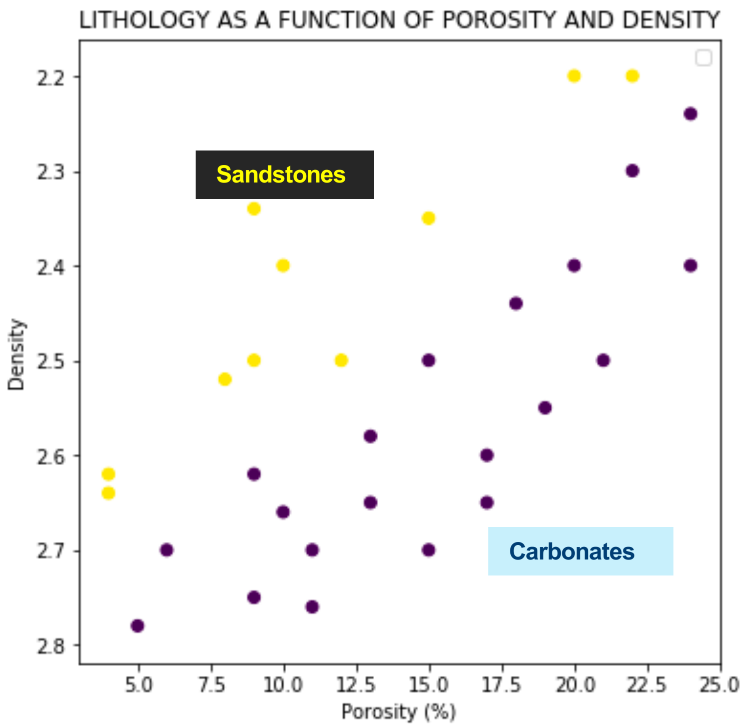
$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

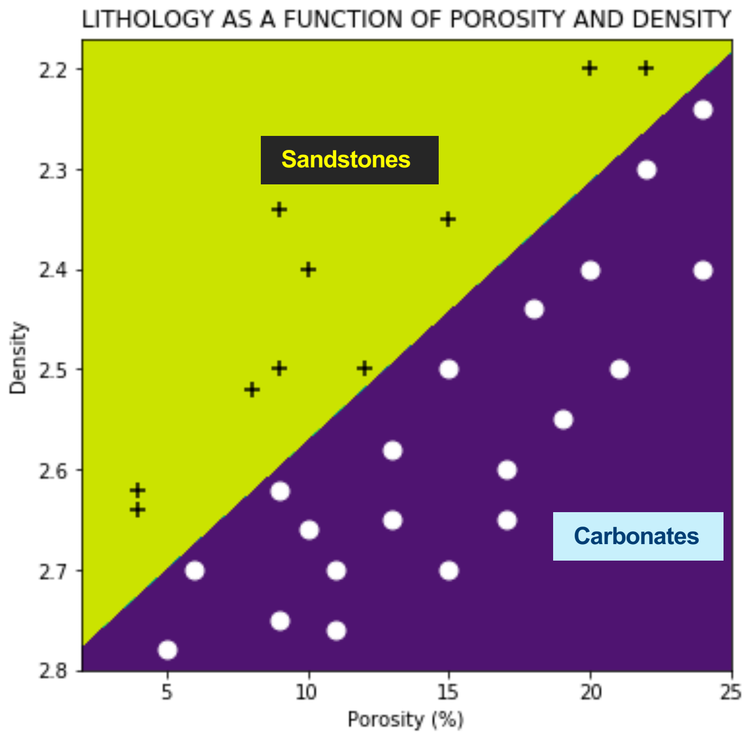
And the predicted class for $x$ is:

$$Class(x) = 0 \; if \; h_\theta(x) < 0.5 \qquad Class(x) = 1 \; if \; h_\theta(x) > 0.5$$

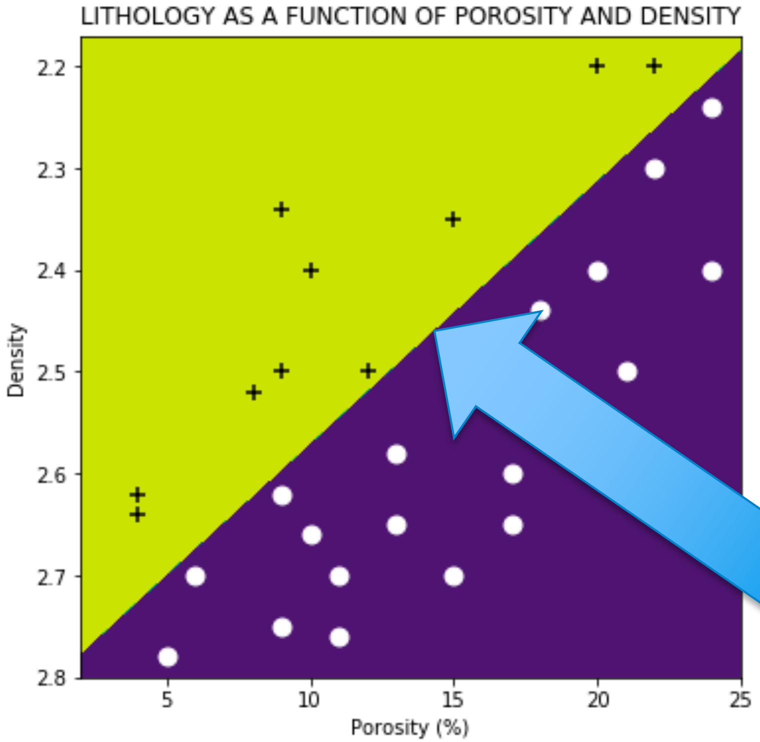
Imperial College
London

# Example of Binary Logistic Regression Result

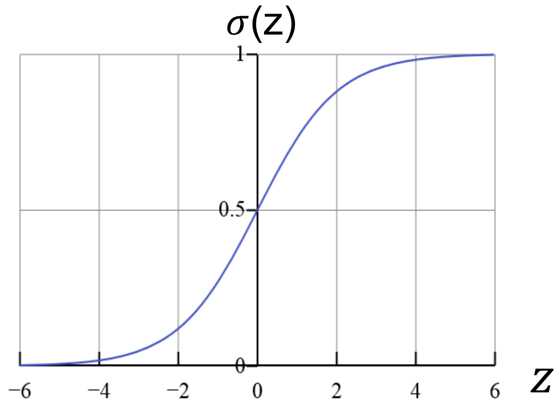


# Definition of the Decision Boundary




$$P(y = 1|\theta, x) = \sigma(\theta^T x)$$




Sigmoid function $\sigma(z)$ is >0.5 if $\theta^T x$>0

**Hence the Decision Boundary in 2D is the line of equation $\theta^T x$=0**

# Softmax Regression on MNIST

**The Results:**

*On the 60000 Training Images*

Mean Accuracy: 0.94

Misclassified Images: 3893 (6.5%)

On the 10000 Test Images

Mean Accuracy: 0.92

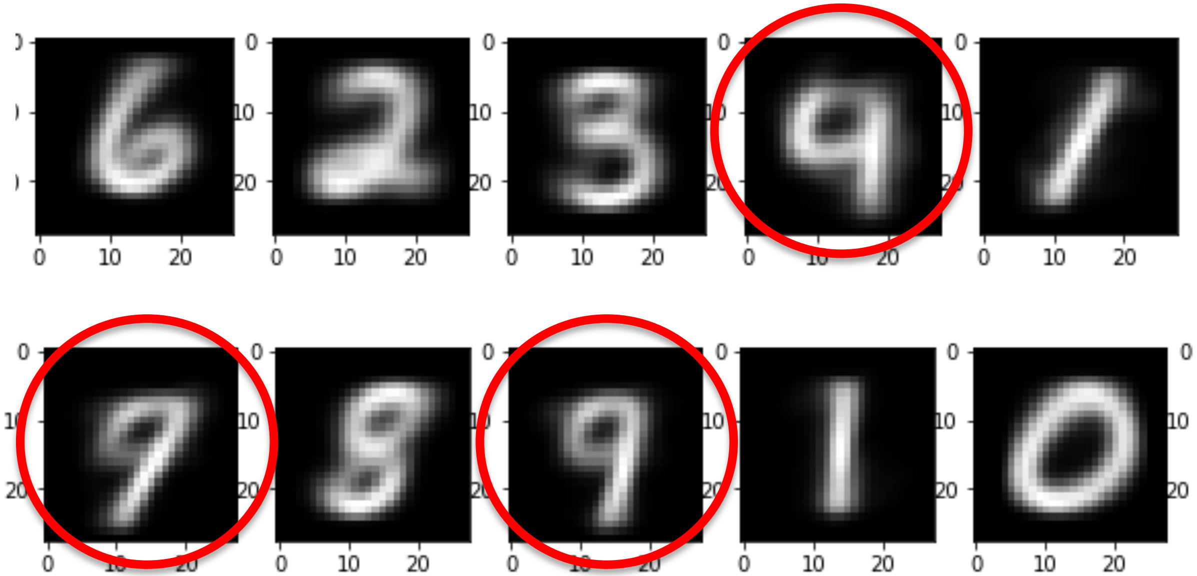
Misclassified Images: 817 (8.2%)

Imperial College London

# Applying K-Means to the MNIST Example

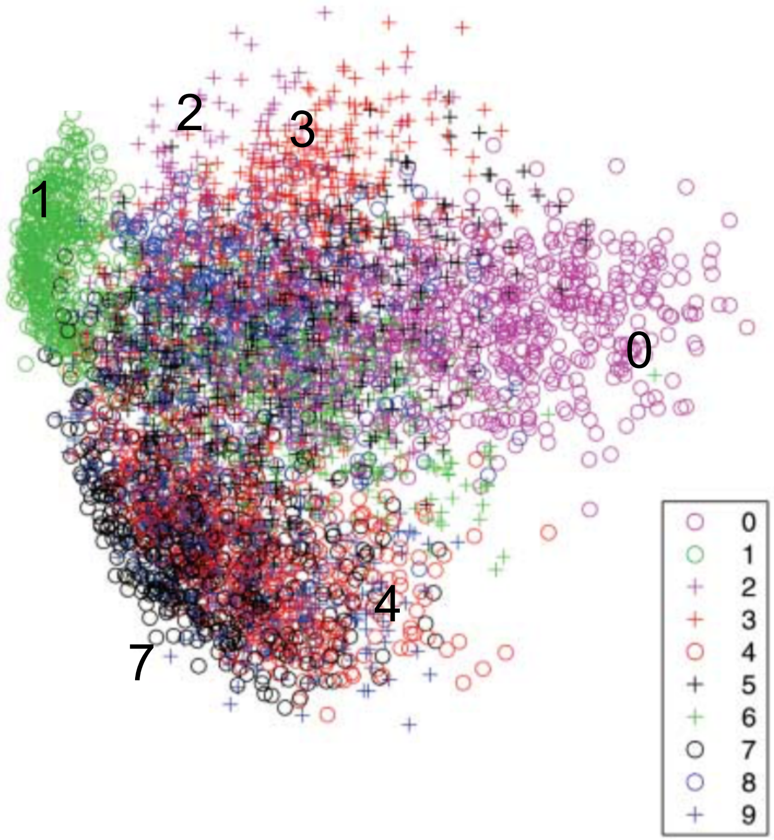
Number of classes:
k=10

Each square is a class centroid (or class mean)



*Digits 4, 5, 7 are not represented as individual classes.*
*There are three classes that look like a mixture of 4, 7 and 9*

# MNIST Results with PCA




The two first principal components for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. The labels were not used for PCA, they are just posted on the PCA results.

# First Session Conclusion

- **Supervised vs Unsupervised Learning**
- **Regression: The Elementary Machine Learning Approach**
- **Logistic Regression: The Elementary Supervised Classification Approach**
- **K-Means and PCA: The Elementary Unsupervised Classification Approaches**
- **Mathematical Notations are Important.**

***Neural Nets and Deep Learning are going to be a generalization of the above to more complex (non-linear) approaches applied to huge datasets.***