Machine Learning for Breast Cancer Diagnosis
A Proof of Concept

Ravinder Singh
Sumit Ananad
Akshay Kumar

Introduction

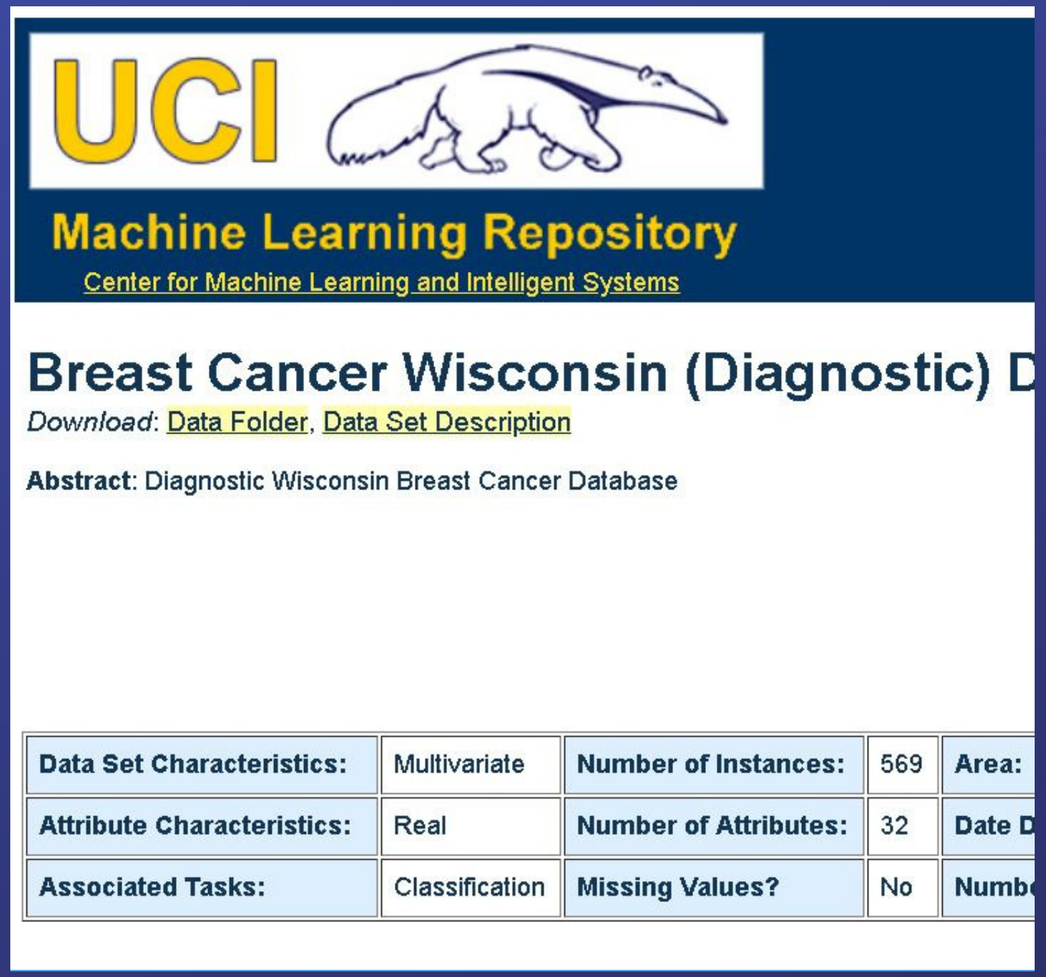
- Machine learning is branch of Data Science which incorporates a large set of statistical techniques.
- These techniques enable data scientists to create a model which can learn from past data and detect patterns from massive, noisy and complex data sets.
- Researchers use machine learning for cancer prediction and prognosis.
- Machine learning allows inferences or decisions that otherwise cannot be made using conventional statistical methodologies.
- With a robustly validated machine learning model, chances of right diagnosis improve.
- It specially helps in interpretation of results for borderline cases.

Breast Cancer: An overview

- The most common cancer in women worldwide.
- The principle cause of death from cancer among women globally.
- Early detection is the most effective way to reduce breast cancer deaths.
- Early diagnosis requires an accurate and reliable procedure to distinguish between benign breast tumors from malignant ones
- Breast Cancer Types - three types of breast tumors: Benign breast tumors, In-situ cancers, and Invasive cancers.
- The majority of breast tumors detected by mammography are benign.
- They are non-cancerous growths and cannot spread outside of the breast to other organs.
- In some cases, it is difficult to distinguish certain benign masses from malignant lesions with mammography.
- If the malignant cells have not gone through the basal membrane but is completely contained in the lobule or the ducts, the cancer is called in-situ or noninvasive.
- If the cancer has broken through the basal membrane and spread into the surrounding tissue, it is called invasive.
- This analysis assists in differentiating between benign and malignant tumors.

Data Source :

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)



UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Breast Cancer Wisconsin (Diagnostic) D

Download: Data Folder, Data Set Description

Abstract: Diagnostic Wisconsin Breast Cancer Database

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date D**
Associated Tasks:	Classification	**Missing Values?**	No	**Numb**

- **Citation**: This breast cancer databases was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg.

- **Reference :** O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.
- William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.
- O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.
- K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

Loading the dataset dictionary

```
In [2]: from sklearn.datasets import load_breast_cancer
        data = load_breast_cancer()

In [3]: data.keys()

Out[3]: dict_keys(['data', 'target', 'target_names', 'DESCR', 'feature_names'])

In [4]: data.target_names

Out[4]: array(['malignant', 'benign'],
              dtype='<U9')
```

Data Description

```
[5]: print(data.DESCR)
```

```
Breast Cancer Wisconsin (Diagnostic) Database
=============================================

Notes
-----
Data Set Characteristics:
    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
        - symmetry
        - fractal dimension ("coastline approximation" - 1)
```

```
:Summary Statistics:

===================================== ====== ======
                                       Min    Max
===================================== ====== ======
radius (mean):                        6.981  28.11
texture (mean):                       9.71   39.28
perimeter (mean):                     43.79  188.5
area (mean):                          143.5  2501.0
smoothness (mean):                    0.053  0.163
compactness (mean):                   0.019  0.345
concavity (mean):                     0.0    0.427
concave points (mean):                0.0    0.201
symmetry (mean):                      0.106  0.304
fractal dimension (mean):             0.05   0.097
radius (standard error):              0.112  2.873
texture (standard error):             0.36   4.885
perimeter (standard error):           0.757  21.98
area (standard error):                6.802  542.2
smoothness (standard error):          0.002  0.031
compactness (standard error):         0.002  0.135
concavity (standard error):           0.0    0.396
concave points (standard error):      0.0    0.053
symmetry (standard error):            0.008  0.079
fractal dimension (standard error):   0.001  0.03
```

```
radius (standard error):                  0.112  2.873
texture (standard error):                 0.36   4.885
perimeter (standard error):               0.757  21.98
area (standard error):                    6.802  542.2
smoothness (standard error):              0.002  0.031
compactness (standard error):             0.002  0.135
concavity (standard error):               0.0    0.396
concave points (standard error):          0.0    0.053
symmetry (standard error):                0.008  0.079
fractal dimension (standard error):       0.001  0.03
radius (worst):                           7.93   36.04
texture (worst):                          12.02  49.54
perimeter (worst):                        50.41  251.2
area (worst):                             185.2  4254.0
smoothness (worst):                       0.071  0.223
compactness (worst):                      0.027  1.058
concavity (worst):                        0.0    1.252
concave points (worst):                   0.0    0.291
symmetry (worst):                         0.156  0.664
fractal dimension (worst):                0.055  0.208
========================================  ======  ======
```

How this Data was collected

This is a copy of UCI ML Breast Cancer Wisconsin (Diagnostic) datasets.
https://goo.gl/U2Uwz2

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:

Data Structure Used : DataFrame

```
In [8]: df.head()
```

Out[8]:

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430

Warnings

- area_error is highly correlated with perimeter_error (ρ = 0.93766) Rejected
- concave points_error has 13 / 2.3% zeros Zeros
- concavity_error has 13 / 2.3% zeros Zeros
- mean_area is highly correlated with mean_perimeter (ρ = 0.98651) Rejected
- mean_concave points is highly correlated with mean_concavity (ρ = 0.92139) Rejected
- mean_concavity has 13 / 2.3% zeros Zeros
- mean_perimeter is highly correlated with mean_radius (ρ = 0.99786) Rejected
- perimeter_error is highly correlated with radius_error (ρ = 0.97279) Rejected
- worst_area is highly correlated with worst_perimeter (ρ = 0.97758) Rejected
- worst_concave points is highly correlated with mean_concave points (ρ = 0.91016) Rejected
- worst_concavity has 13 / 2.3% zeros Zeros
- worst_perimeter is highly correlated with worst_radius (ρ = 0.99371) Rejected
- worst_radius is highly correlated with mean_area (ρ = 0.96275) Rejected
- worst_texture is highly correlated with mean_texture (ρ = 0.91204) Rejected

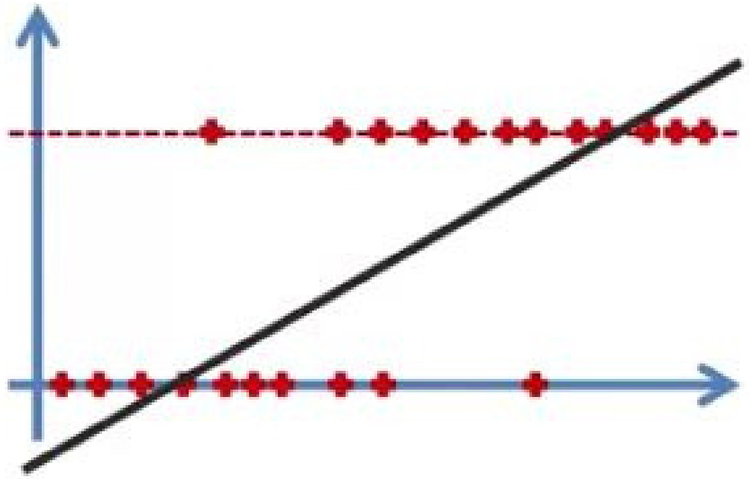
Algorithm Used : Logistic Regression

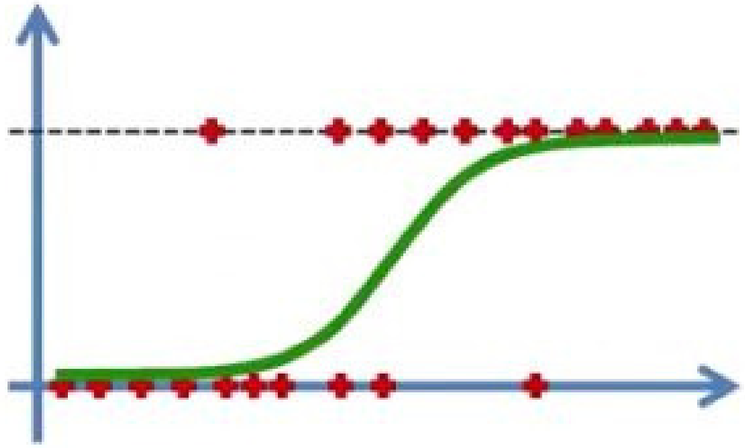
$$y = b_0 + b_1 * x$$

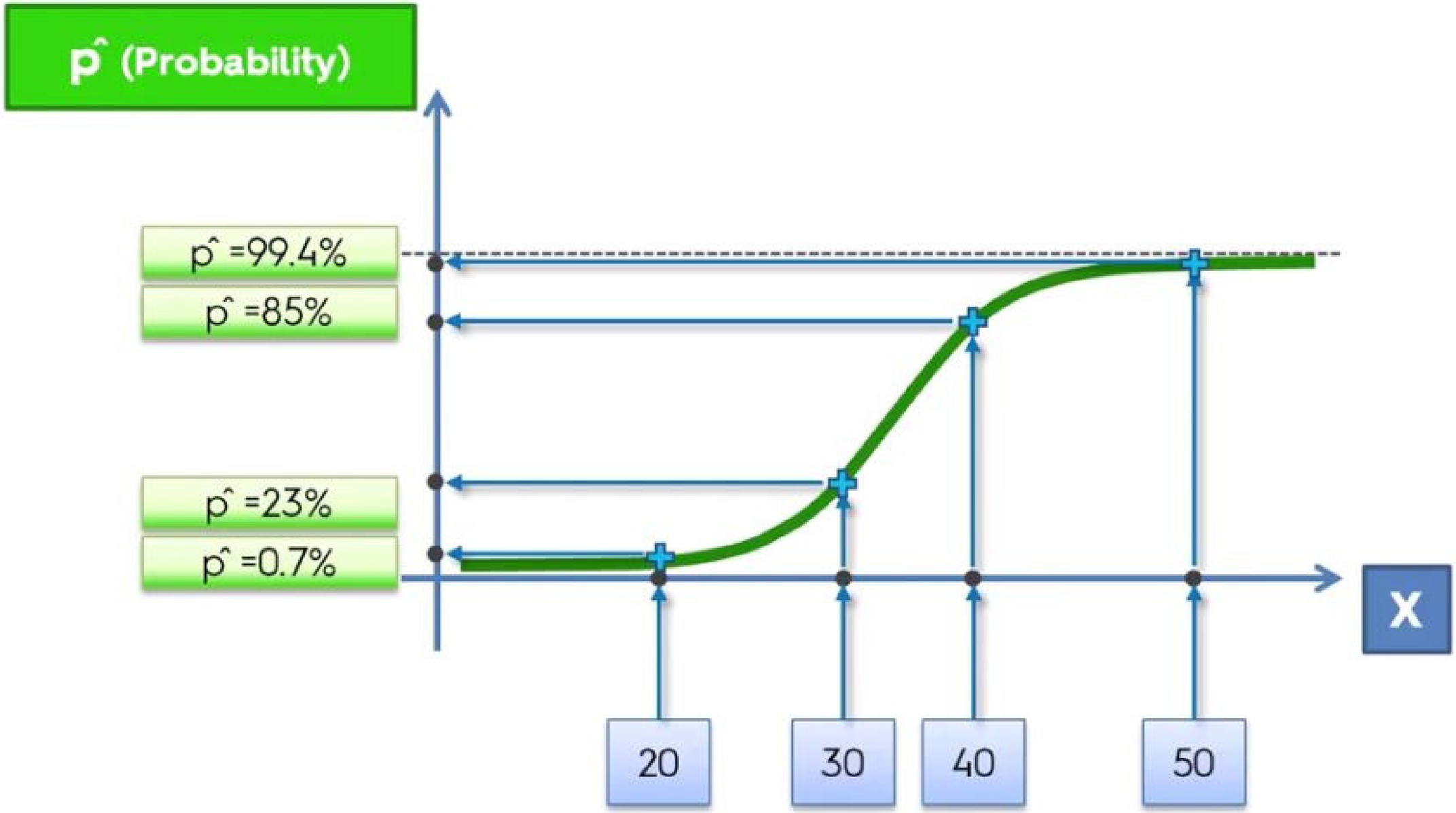
Sigmoid Function

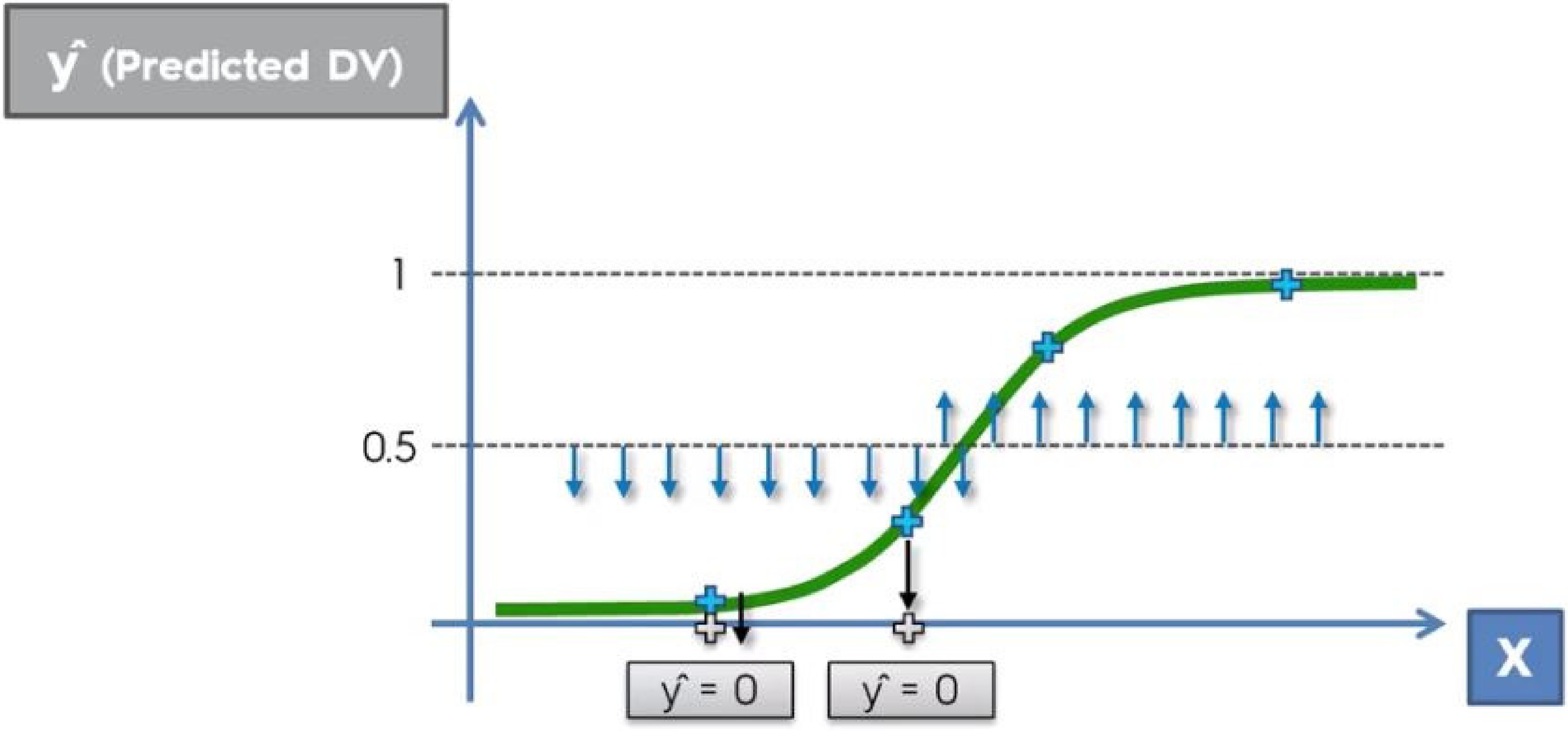
$$p = \frac{1}{1 + e^{-y}}$$

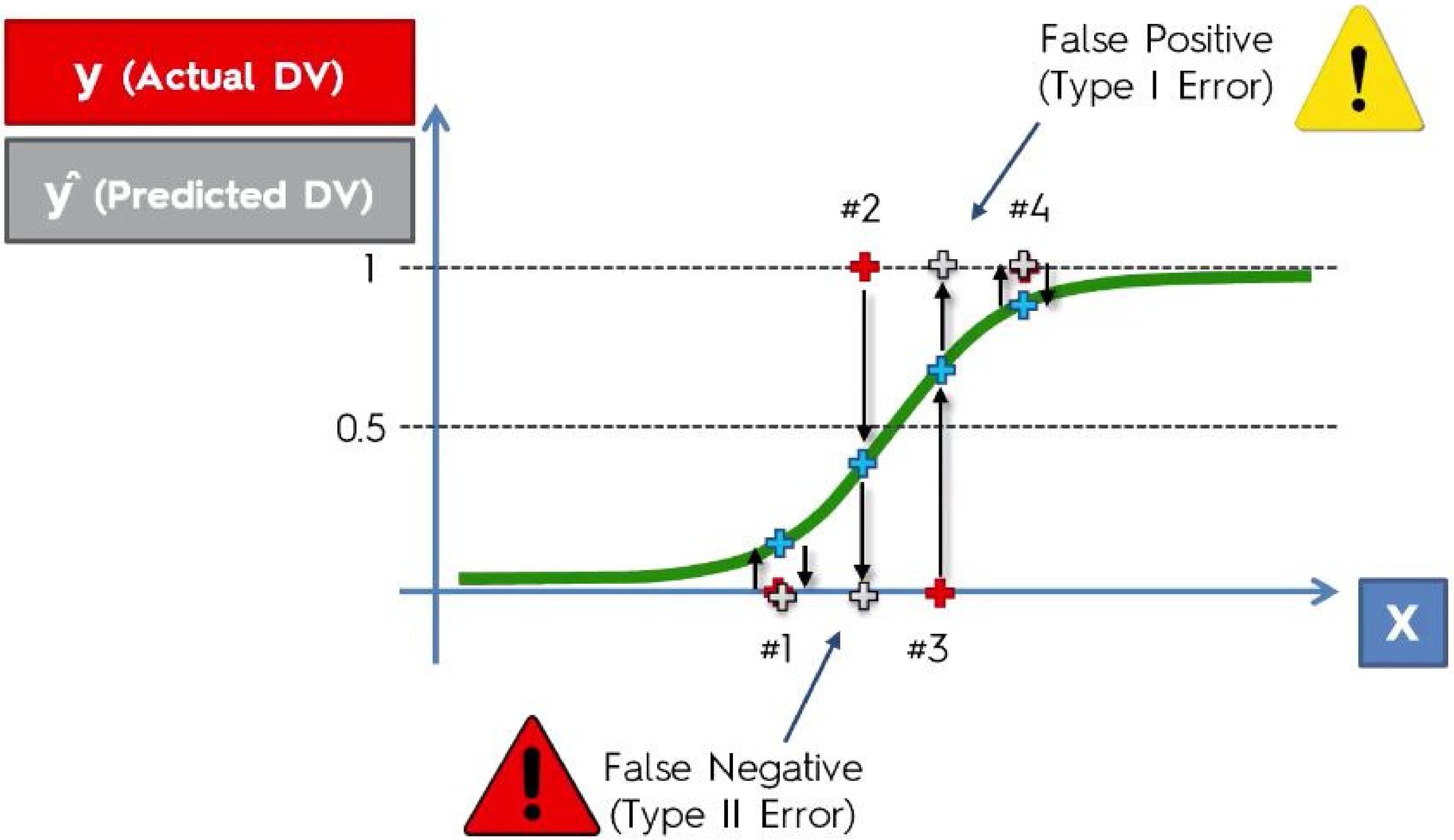
$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$






p̂ (Probability)
p̂ =99.4%
p̂ =85%
p̂ =23%
p̂ =0.7%
X
20
30
40
50


ŷ (Predicted DV)
1
0.5
ŷ = 0
ŷ = 0
X


y (Actual DV)
ŷ (Predicted DV)
False Positive
(Type I Error)
#2
#4
1
0.5
#1
#3
X
False Negative
(Type II Error)

		$\hat{y}$ (Predicted DV)	
		0	1
y (Actual DV)	0	35	5
	1	10	50

False Positive
(Type I Error)

False Negative
(Type II Error)

Calculate two rates

1. Accuracy Rate = Correct / Total
AR = 85/100 = 85%