INSTITUT TEKNOLOGI BANDUNG

1920

01

Logistic Regression

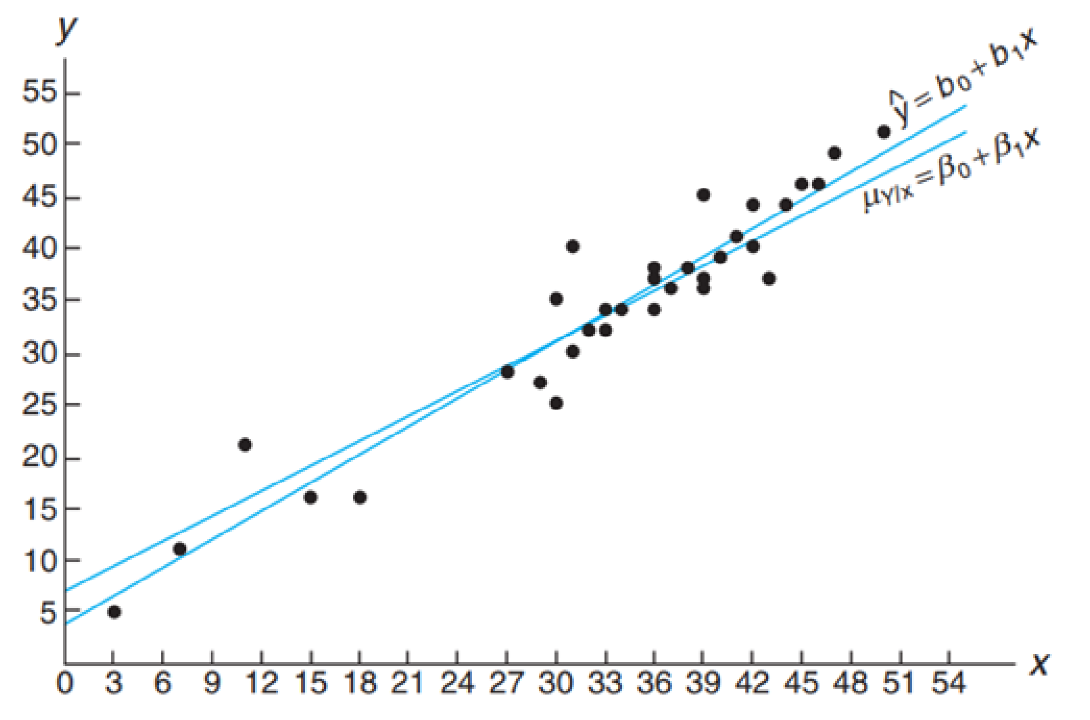
What & Why

Masayu Leylia Khodra (masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Pembelajaran Mesin (*Machine Learning*)

REGRESSION ANALYSIS



Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probability and Statistics for engineering and sciences. *Pearson Education*, 430-435.

Finding the best relationship between Y and x: not deterministic, random error

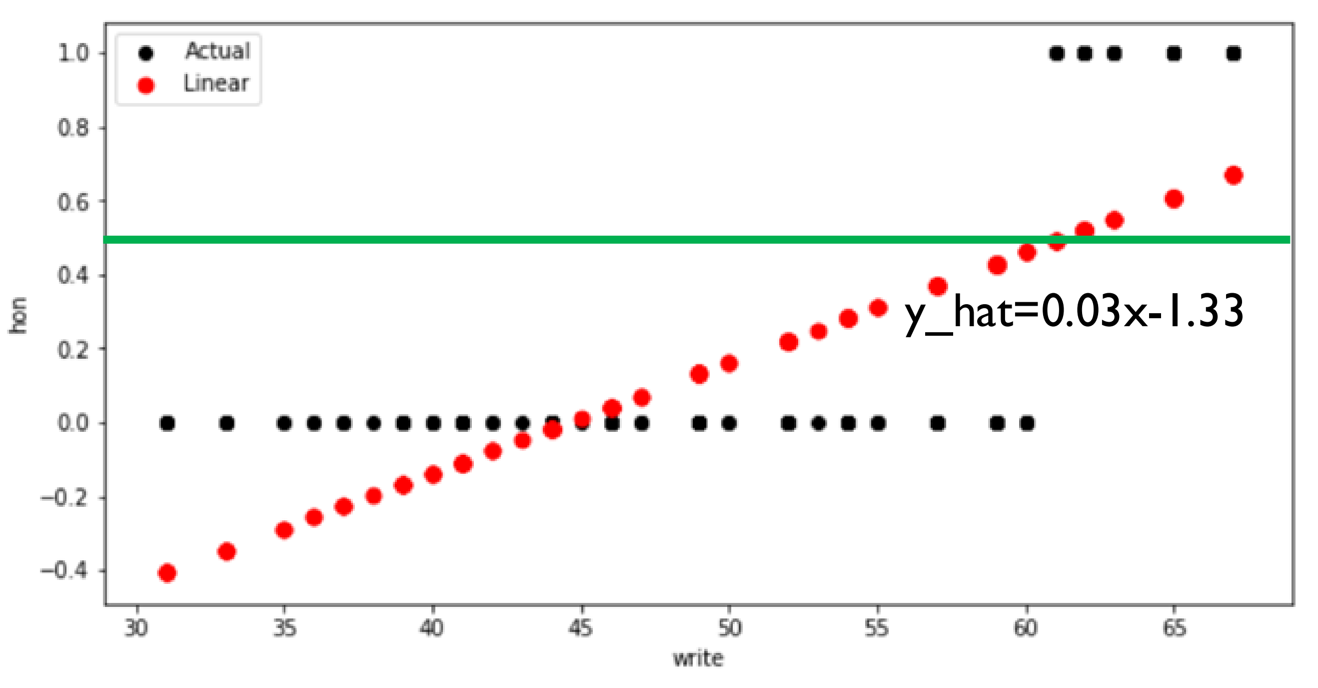
Quantifying the strength of that relationship. Random error with E(error)=0 and homoscedasticity. Least Squared Estimation (LSE)

Predicting of the response value y given values of the regressor x.

CLASSIFICATION USING LINEAR REGRESSION

LINEAR REGRESSION + THRESHOLD

Example: Dataset has 200 observations (5 attributes), and the target attribute is **hon**, indicating if a student is in an honors class or not.

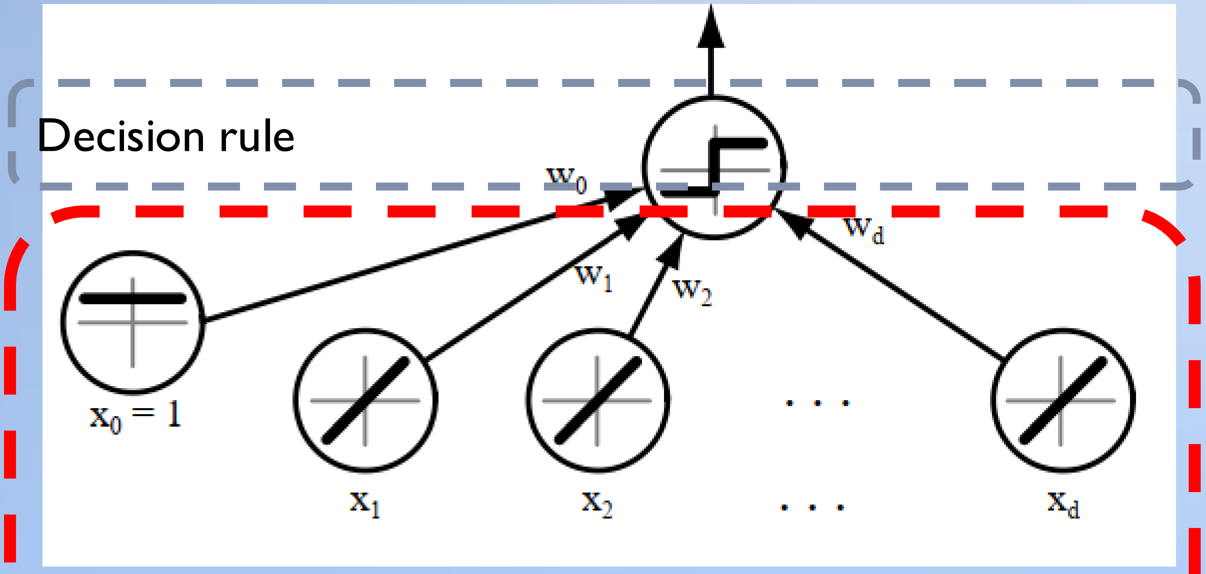


We are able to predict the value along the Y-axis. If Y is greater than 0.5 (above the green line), predict that the student is in honor class otherwise not in honor class.

Linear Model or
Linear Discriminant Function
or Linear Classifier

LINEAR DISCRIMINANT FUNCTION

LINEAR COMBINATION + DECISION RULE

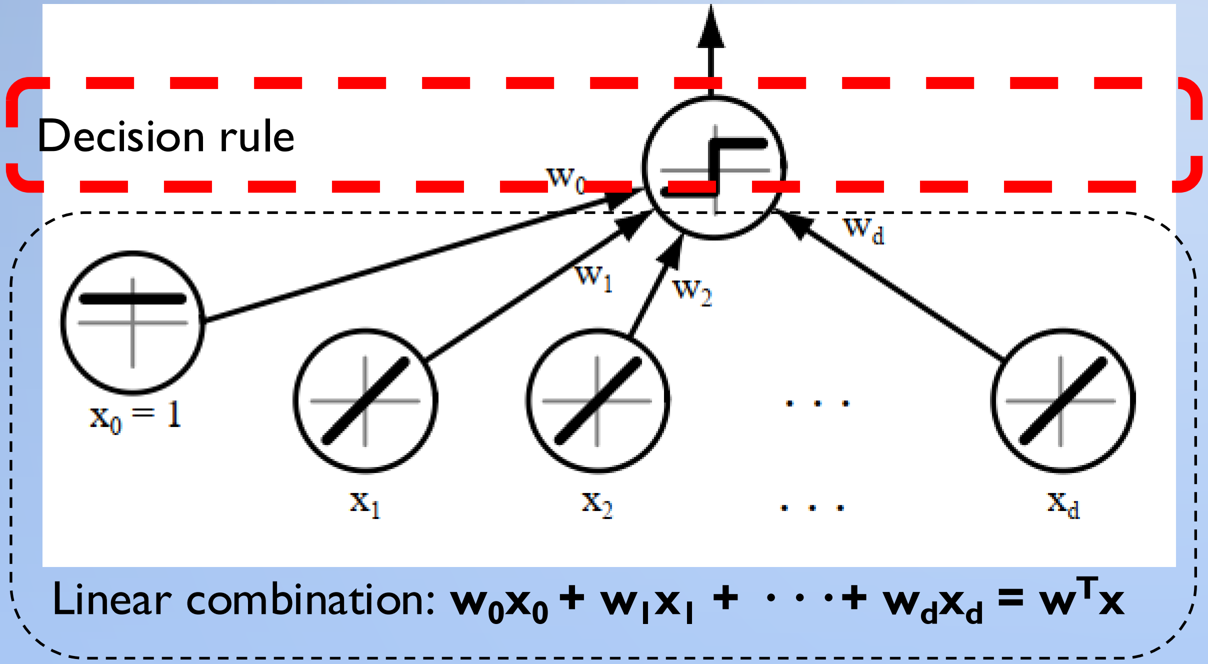


Linear combination: $\mathbf{w_0x_0 + w_1x_1 + \cdots + w_dx_d = w^Tx}$

Discriminant function:
a linear combination of the components of x
$g(\mathbf{x})=w_0x_0 + w_1x_1 + \cdots + w_dx_d = \mathbf{w}^T\mathbf{x}$

RO Duda, PE Hart, and DG Stork, Pattern Classification, 2nd edition, John Wiley & Sons, 2001
https://www.cs.toronto.edu/~rgrosse/courses/csc321_2018/readings/L03%20Linear%20Classifiers.pdf

HYPERPLANE DECISION SURFACE



Discriminant function:
$g(\mathbf{x})=\mathbf{w}^T\mathbf{x}$

Decision rule:

Decide ω_1 if $g(x)>0$ and ω_2 if $g(x)<0$

$g(x)=0$: decision surface

ω_1 and ω_2 are target classes

HYPERPLANE DECISION SURFACE

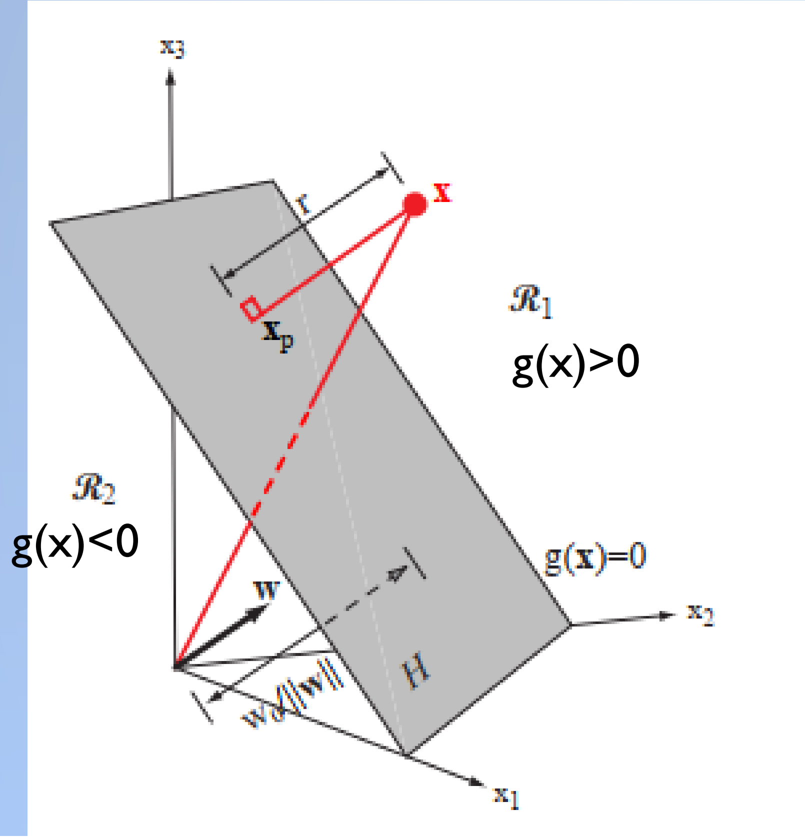


Figure 5.2: The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$).

Discriminant function:

$g(\mathbf{x})=\mathbf{w}^T\mathbf{x}$

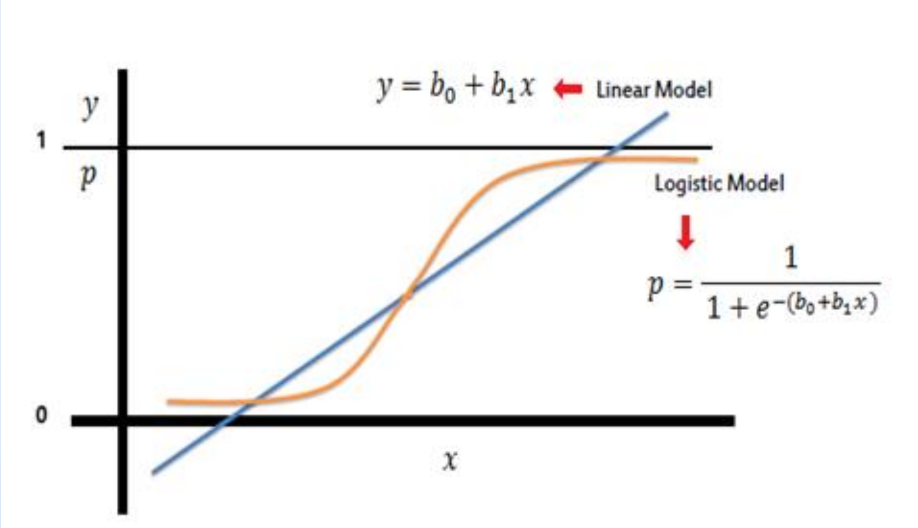
Decision rule:

Decide ω_1 if $g(x)>0$ and ω_2 if $g(x)<0$

$g(x)=0$: decision surface

ω_1 and ω_2 are target classes

Linear Regression	Logistic Regression
	Estimate the probability of classes
Least squares	Maximum likelihood
Constant variance	Non-constant variance
Response variable is normal	Response variable is binary

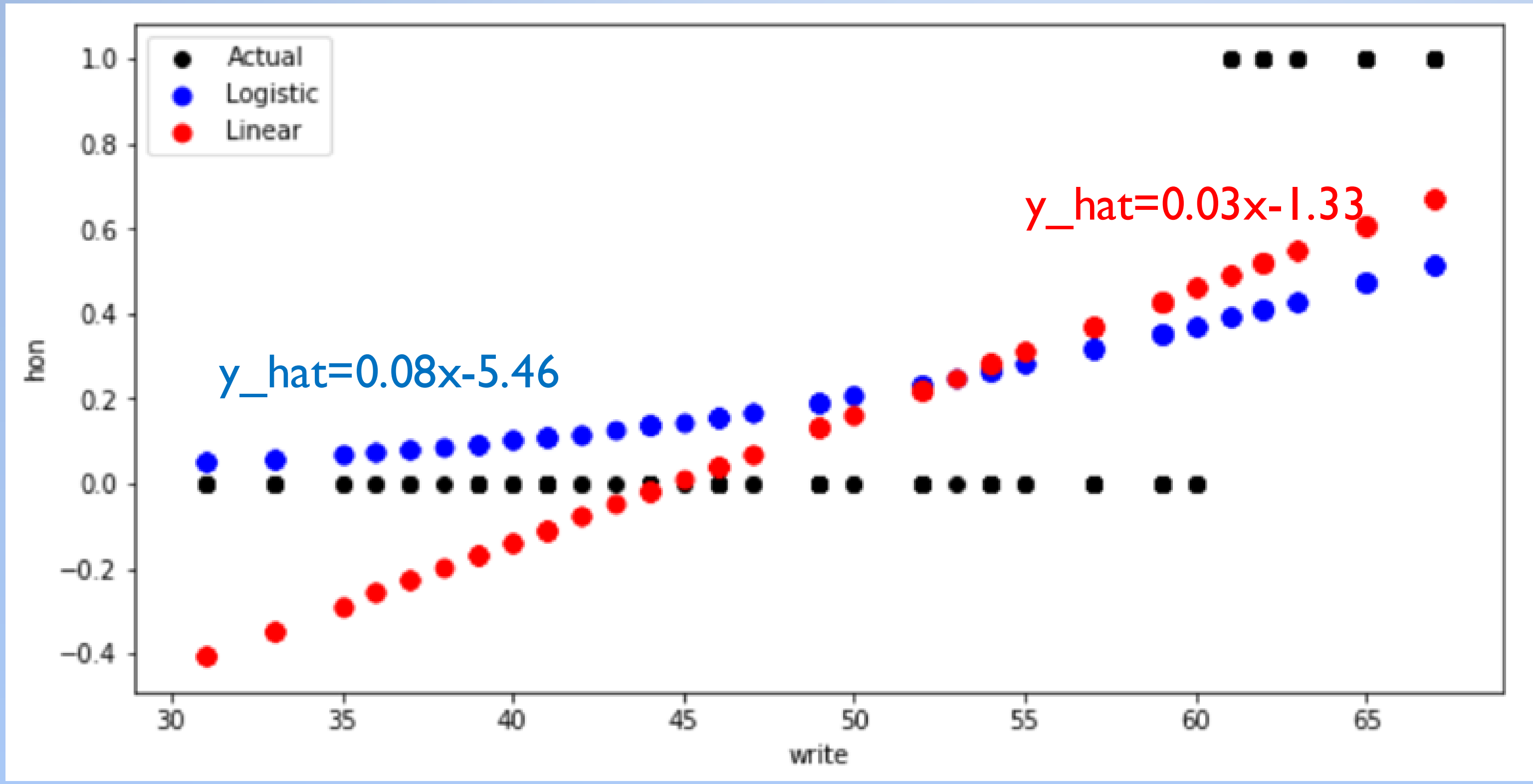


http://www.saedsayad.com/logistic_regression.htm

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probability and Statistics for engineering and sciences. *Pearson Education*, 430-435.

LOGISTIC REGRESSION

$$\hat{y} = \log\left(\frac{p}{1-p}\right) = \boldsymbol{b}^T\boldsymbol{x} = b_0.1 + b_1x_1 + \cdots + b_dx_d$$



Walpole et al. (2012):
Odds of success = p/(1-p)
Logistic regression estimates the probability of classes:

$$p = P(y = 1|\boldsymbol{x}, \boldsymbol{b}) = \frac{1}{1 + e^{-\boldsymbol{b}^T\boldsymbol{x}}}$$

Coefficients in logistic regression are in terms of the log odds, that is, the coefficient 0.08 implies that a one unit change in "write" results in a 0.08 unit change in the log of the odds.

https://stats.idre.ucla.edu/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/

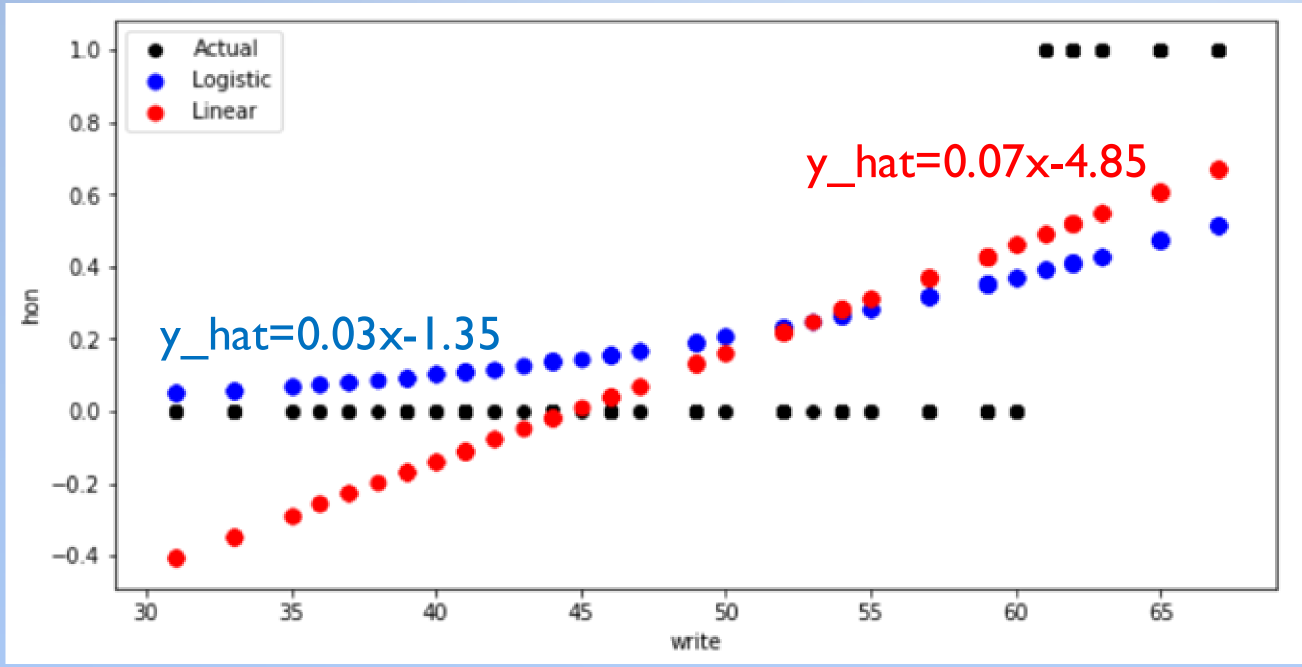
SUMMARY: LOGISTIC REGRESSION

- ☑ Linear model
- ☑ Binary classification
- ☑ Soft classification

$$\hat{y} = \log\left(\frac{p}{1-p}\right) = \boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{x} = b_0.1 + b_1x_1 + \cdots + b_dx_d$$

$$p = P(y = 1|\boldsymbol{x}, \boldsymbol{b}) = \frac{1}{1 + e^{-\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{x}}}$$

EXERCISE



Given these linear regression and logistic regression models, determine whether a student that has write score 65 is in honors class.

12 REFERENCES

- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probability and Statistics for engineering and sciences. *Pearson Education*, 430-435. Chapter 11 & 12.12, 9.14
- RO Duda, PE Hart, and DG Stork, Pattern Classification, 2nd edition, John Wiley & Sons, 2001. Chapter 5
- Charles Elkan (2014). Maximum Likelihood, Logistic Regression, and Stochastic Gradient Training. https://cseweb.ucsd.edu/~elkan/250B/logreg.pdf
- Russell, S., & Norvig, P. (2010). Artificial intelligence: a modern approach. 3rd edition. Chapter 18.6.4.

Logistic Regression

Stochastic Gradient Ascent

Masayu Leylia Khodra (masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Pembelajaran Mesin (*Machine Learning*)

LOGISTIC REGRESSION

$$\hat{y} = \log\left(\frac{p}{1-p}\right) = \boldsymbol{b}^T\boldsymbol{x} = b_0.1 + b_1x_1 + \cdots + b_dx_d$$

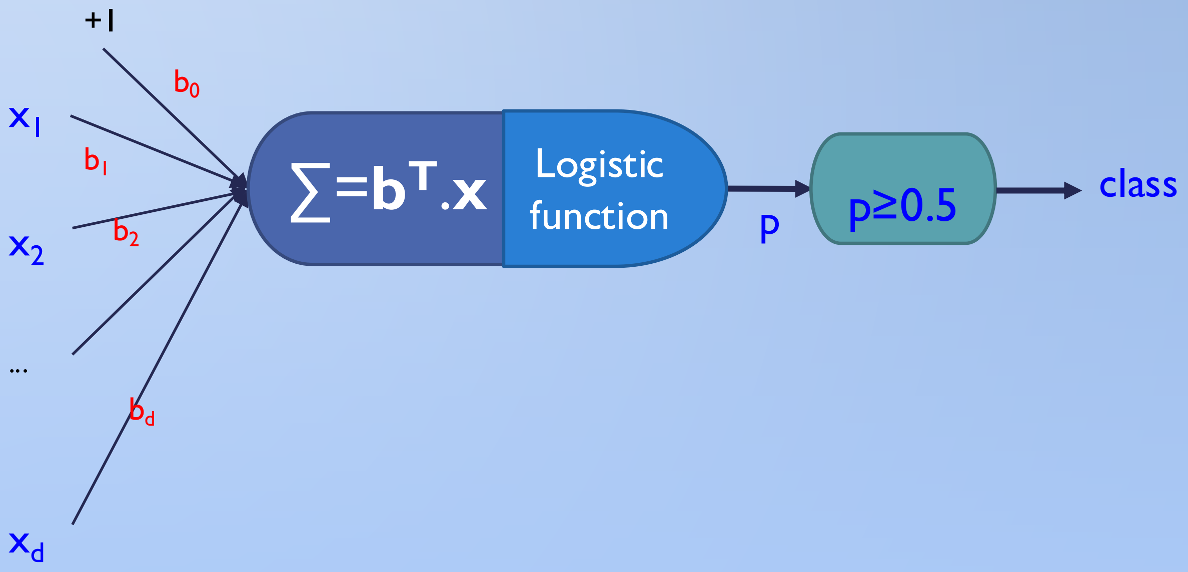
$$p = P(y = 1|\boldsymbol{x}, \boldsymbol{b}) = \frac{1}{1 + e^{-\boldsymbol{b}^T\boldsymbol{x}}}$$

Input $\mathbf{x}=(1, x_1, x_2, \ldots, x_d)$

Model $\mathbf{b}=(b_0, b_1, b_2, \ldots, b_d)$

$\sum = \mathbf{b}^T\boldsymbol{x} = b_0.1 + b_1x_1 + \cdots + b_dx_d$

Output = $\sigma(\sum)$



15

MAXIMUM LIKELIHOOD ESTIMATOR FOR LOGISTIC REGRESSION

ESTIMATOR THAT RESULTS IN A MAXIMUM VALUE FOR ITS JOINT PROBABILITY OR MAXIMIZES THE LIKELIHOOD OF THE SAMPLE

Formal definition (Elkan, 2014):

Given the training set {<x_1,y_1>..<x_n,y_n>}, learn logistic regression classifier by maximizing the log joint conditional likelihood, that is the sum of log conditional likelihood (LCL) for each training example. x_{ij} is the value of the jth feature of the ith training example.

$$LCL = \sum_{i=1}^{n} \log L(\theta; y_i|x_i) = \sum_{i=1;y_i=1}^{n} \log p_i + \sum_{i=1;y_i=0}^{n} \log(1-p_i)$$

$$\rightarrow \frac{\partial LCL}{\partial b_j} = \sum_i (y_i - p_i) x_{ij}$$

Stochastic gradient **ascent** is optimization method that changes the coefficient values (as random approximation to true derivative) to **increase** the log likelihood based on a randomly chosen example at a time.

Stochastic gradient update of b_j is:

$$b_j = b_j + \eta(y_i - p_i)x_{ij}$$

η : learning rate

https://cseweb.ucsd.edu/~elkan/250B/logreg.pdf

16 STOCHASTIC GRADIENT ASCENT FOR LOGISTIC REGRESSION

INPUT: TRAINING DATA $D=\{<X_1,Y_1>\ldots<X_N,Y_N>\}$; MAX-ITER T; LEARNING RATE η

Initialize **b**

For t=1, …, T:

- For each example <xi,yi>: #randomly chosen example
  - pi=prediction for xi using the current coefficients **b**
  - For each non-zero feature of xij: $b_j = b_j + \eta(y_i - p_i)x_{ij}$

One iteration = one epoch

Return **b**

D: $\{<[52,41],0>,<[62,58],1>\}$;T=1; η=0.1

b=[0,0,0] #b0=0; b1=0; b2=0

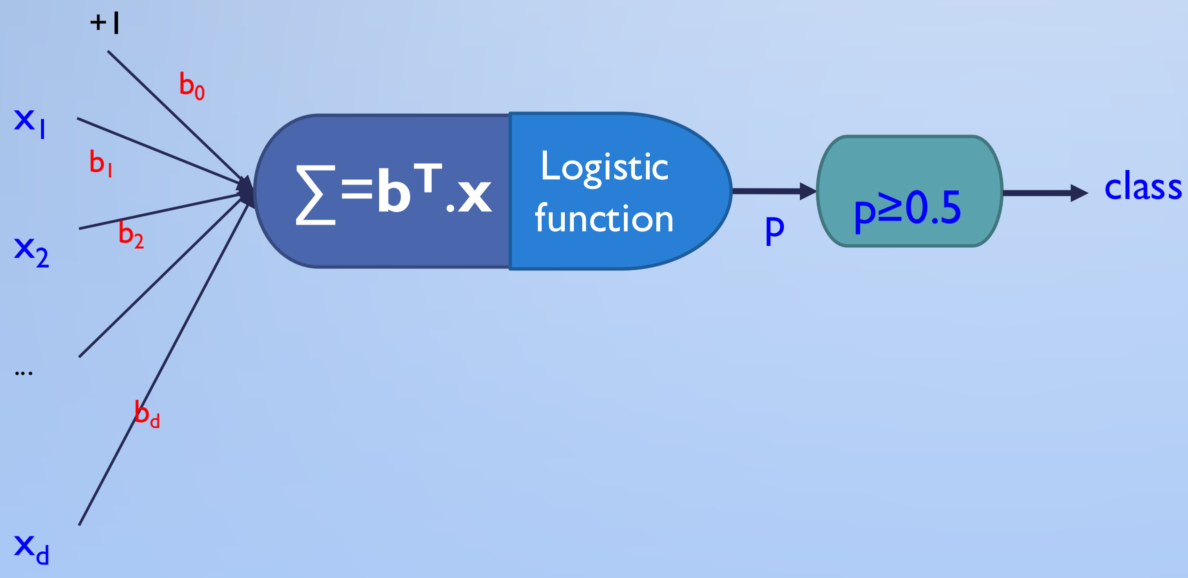
t=1:

```
<[62,58],1>:
pi=1/(1+e^-0)=0.5
b0=0+0.1(1-0.5)*1=0.05
b1=0+0.1(1-0.5)*62=3.1
b2=0+0.1(1-0.5)*58=2.9
<[52,41],0>:
pi=1/(1+e^-(0.05+3.1*52+2.9*41))
=1/(1+e^-280.15)=1
b0=0.05+0.1(0-1)*1= -0.05
b1=3.1+0.1(0-1)*52= -2.1
b2=2.9+0.1(0-1)*41= -1.2
```

PREDICTION

- $b0= -0.05$; $b1= -2.1$; $b2= -1.2$
- D: {<[52,41],0>,<[62,58],1>}
  - **x1**=[52,41]: $p1= 1/(1+e^{-(-0.05-2.1*52-1.2*41)}) =1/(1+e^{158.45})=1.53*10^{-69}$ class=0 ($p1<0.5$)
  - **x2**=[62,58]: $p2= 1/(1+e^{-(-0.05-2.1*62-1.2*58)}) =1/(1+e^{199.85})=1.61*10^{-87}$ class=0 ($p2<0.5$)
- Akurasi training = ½ = 50%

SUMMARY: LOGISTIC REGRESSION



- Model: $\mathbf{b} \in \Re^{d+1}$
- Maximum Likelihood estimator
- Stochastic gradient ascent

REFERENCES

- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probability and Statistics for engineering and sciences. *Pearson Education*, 430-435. Chapter 11 & 12.12, 9.14
- RO Duda, PE Hart, and DG Stork, Pattern Classification, 2nd edition, John Wiley & Sons, 2001. Chapter 5
- Charles Elkan (2014). Maximum Likelihood, Logistic Regression, and Stochastic Gradient Training. https://cseweb.ucsd.edu/~elkan/250B/logreg.pdf
- Russell, S., & Norvig, P. (2010). Artificial intelligence: a modern approach. 3rd edition. Chapter 18.6.4.

INSTITUT TEKNOLOGI BANDUNG

1920