LECTURE 5

CEIC6789: NOTES

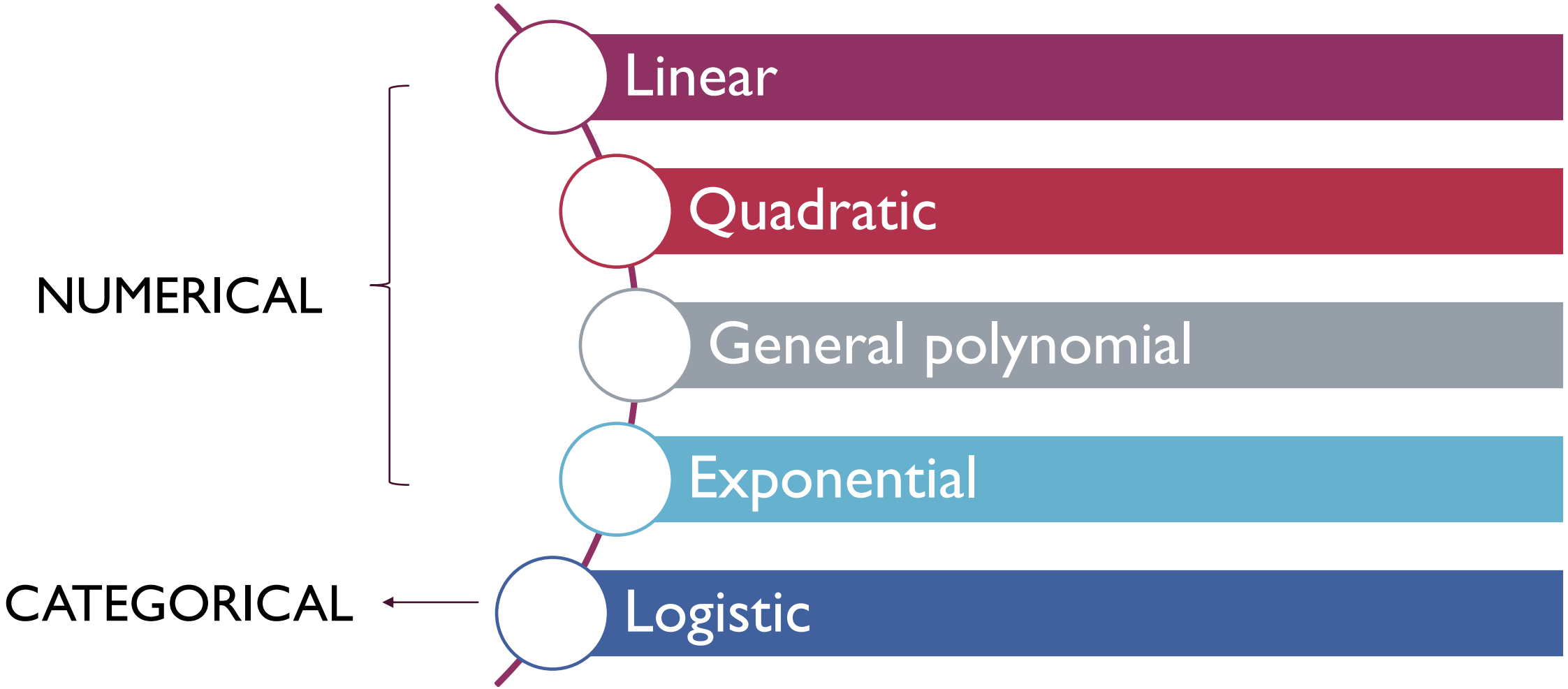
LINEAR REGRESSION

NONLINEAR

NONLINEAR

- Quadratic
- Quartic
- General polynomial
- Exponential
- Logistic

WHY LOGISTIC REGRESSION?



CATEGORICAL OUTCOMES

YES or NO

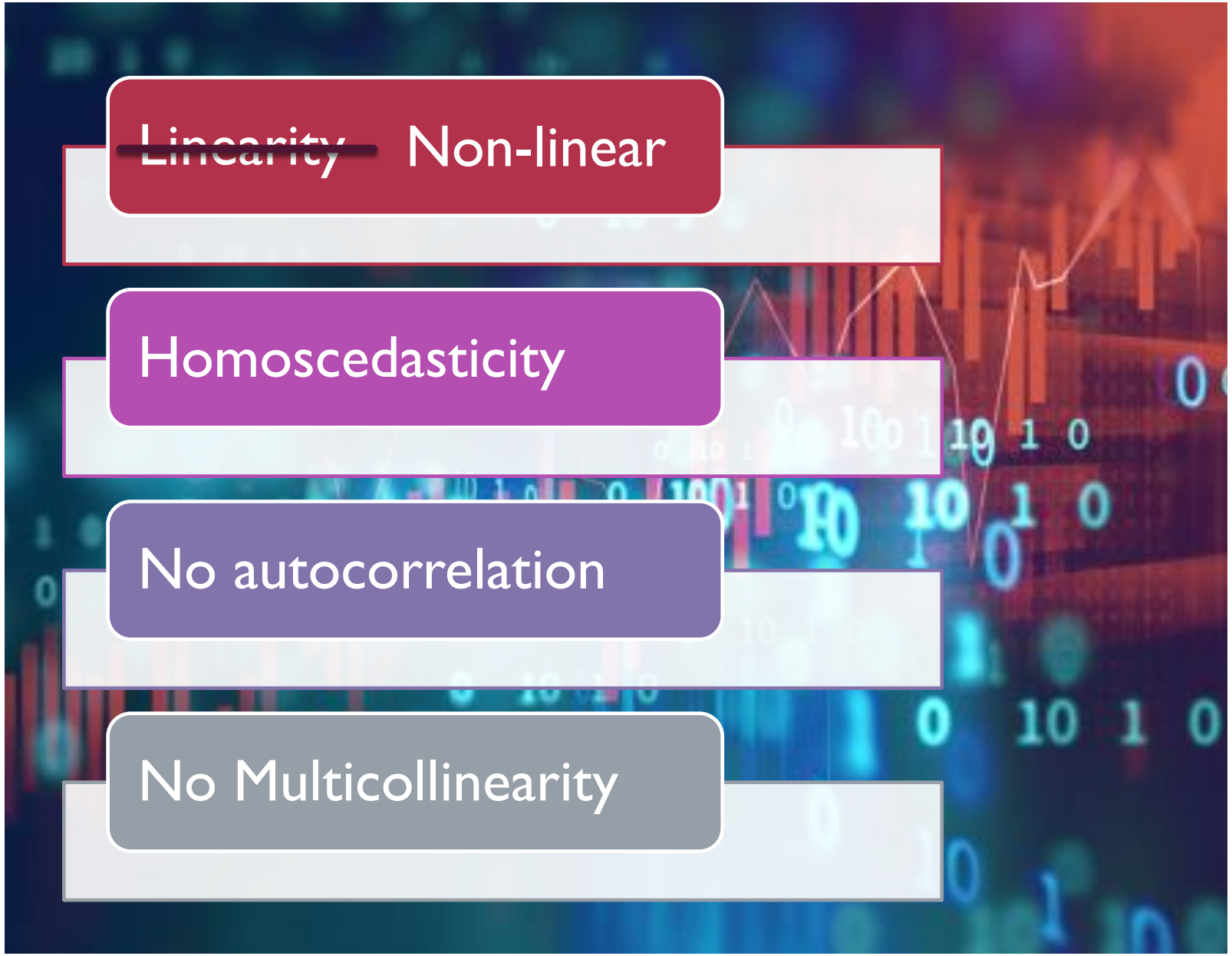
WILL BUY or WON'T BUY

DECISION MAKING

- Should we manufacture this product? Yes or No
- Will a consumer buy this product? Yes or No
- Should we implement a new reactor design? Yes or No

WOULD PEOPLE BUY A LAPTOP GIVEN ITS PRICE?

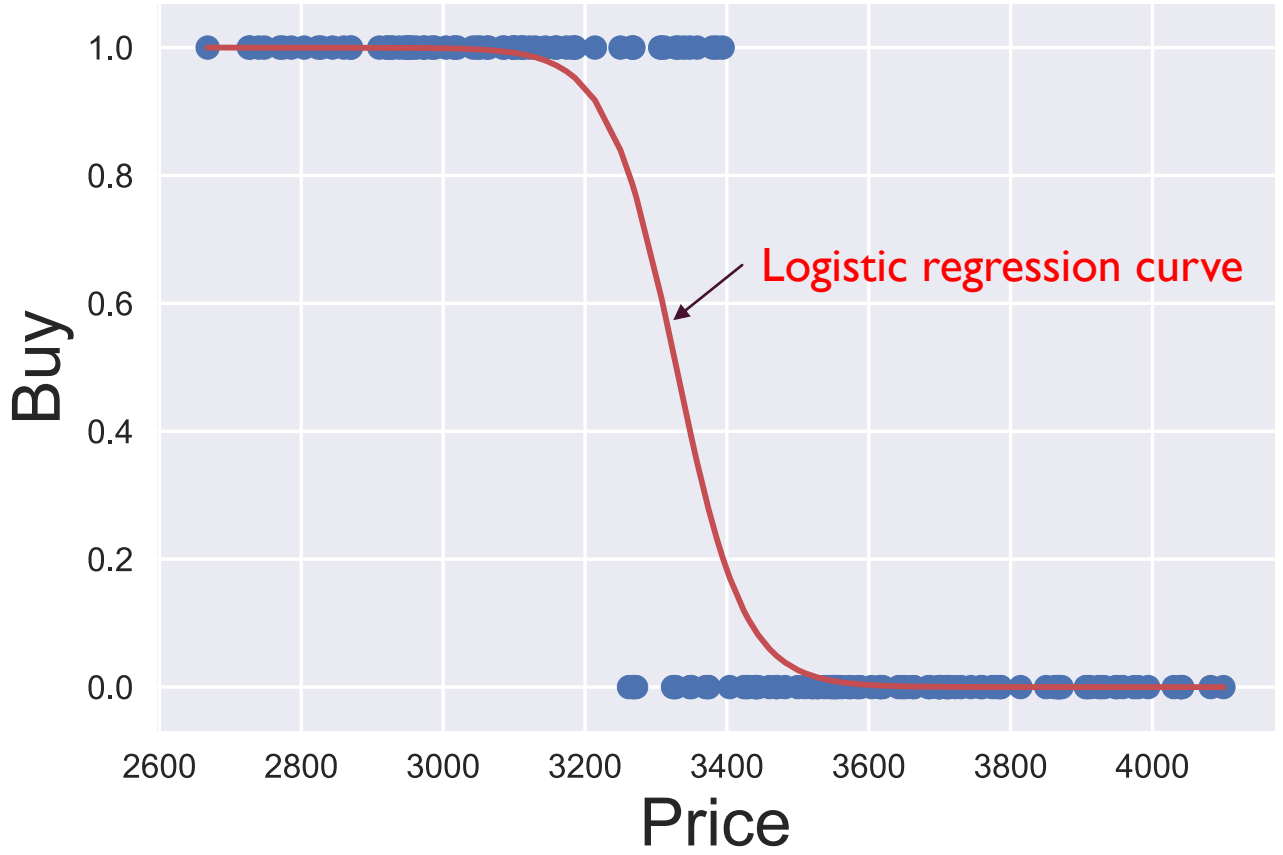
YES OR NO?



LOGISTIC REGRESSION

Logistic regression predicts the probability of an event occuring

What is the probability of a customer buying the laptop?



LOGISTIC

$$p = \frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

LOGIT

$$\underbrace{\frac{p}{1-p}}_{\text{odds}} = e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}$$

ODDS

Odds of getting a head

p (getting a head) = 0.5

p (not getting a head) = 1 – 0.5 = 0.5

Odds = p / (1- p) = 0.5 / 0.5 = 1

Odds of getting a 6

p (getting a 6) = 1/6

p (not getting a 6) = 1 – 1/6 = 5/6

Odds = p / (1- p) = 1/6 / 5/6 = 1/5

LOGISTIC

$$p = \frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

LOGIT

$$\frac{p}{1-p} = e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}$$

$$\log\frac{p}{1-p} = \underbrace{b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k}$$

LINEAR REGRESSION

LINEAR REGRESSION (OLS)

How to get the best fitting line?

Algebraic solution (Direct)

- min. SSE
- $b = (X^T X)^{-1} X^T Y$

Numerical solution (Indirect)

- Objective function $f(b)$
- $f(b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Minimize $f(b)$
- Optimizers or solvers to minimize $f(b)$

LOGISTIC REGRESSION

How to get the best fitting line?

Numerical solution (Indirect)

- Objective function $f(b)$
- $f(b) = \sum_{i=1}^{n} y^{(i)} \log \sigma\left(b^T x^{(i)}\right) + (1 - y^{(i)}) \log\left(1 - \sigma\left(b^T x^{(i)}\right)\right)$
- Maximize $f(b)$ - Maximum likelihood estimation
- Optimizers or solvers to maximize $f(b)$

ACTIVITY

- You can go through the mathematics behind how the objective function is obtained for logistic regression in the document named "Logistic-regression-theory_Courtesy_WillMonroe_Stanford". This is given in the folder "Lecture resources".
- If you are interested in understanding about underfitting and overfitting in machine learning, and the type of penalty terms employed to prevent overfitting, please go through the following links (given in Lecture resources):
  - https://towardsdatascience.com/over-fitting-and-regularization-64d16100f45c
  - https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c
  - https://www.datacamp.com/community/tutorials/towards-preventing-overfitting-regularization?utm_source=adwords_ppc&utm_campaignid=9942305733&utm_adgroupid=100189364546&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=229765585183&utm_targetid=aud-299261629574:dsa-929501846124&utm_loc_interest_ms=&utm_loc_physical_ms=20605&gclid=Cj0KCQjw3Nv3BRC8ARIsAPh8hgL2XdlBbtBVFx8eGeV85tUOY6tZsde1Y1F5NDOMQf_DPdJroaPQmToaAg_TEALw_wcB

MULTIPLE LOGISTIC REGRESSION

$$\log(odds) = 66.40 - 0.02 \times Price + 1.94 \times Gender$$

$$\log(odds_2) = 66.40 - 0.02 \times Price_2 + 1.94 \times Gender_2$$

$$- \log(odds_1) = 66.40 - 0.02 \times Price_1 + 1.94 \times Gender_1$$

$$\log(odds_2/odds_1) = 1.94 \times (Gender_2 - Gender_1)$$

$$\log(odds_2/odds_1) = 1.94$$

$$odds_2 = 7 * odds_1$$

ACTIVITY

Divide the dataset into training and testing datasets. You can then perform regression on the training set, and compute the confusion matrix and accuracy on the testing dataset.