Artificial Intelligence 2: Logistic Regression and Maximum Likelihood

Shan He

School for Computer Science
University of Birmingham

Outline of Topics

Classification Problems

- **Classification**: determining the most likely class that an input pattern belongs to.
- **Formally**: modelling the posterior probabilities of class membership (dependent variable) conditioned on the input (independent) variables.
- Artificial neural networks: one output unit for each class, and for each input pattern we have
  - 1 for the output unit corresponding to that class
  - 0 for all the other output units
- The simplest case: binary classification $\rightarrow$ one output unit

Example 1: Hepatocellular carcinoma diagnosis

Hepatocellular carcinoma (HCC): a common complication of chronic liver disease (CLD), and is conventionally diagnosed by radiological means.

HCC diagnosis:

- Histologic examination of tumor tissue: invasive
- Ultrasound, Computed Tomography (CT) or MRI scanning: inaccurate and subjective
- Serum biomarker: α-fetoprotein (AFP): limited sensitivity for smaller HCC tumours

GALAD score [1]:

- Based on age, sex, and three serum biomarkers
- Recently approved by FDA and now used in Roche's instruments.
- Developed using Logistic Regression

[1] The Detection of Hepatocellular Carcinoma Using a Prospectively Developed and Validated Model Based on Serological Biomarkers, Philip J. Johnson, et al. Cancer Epidemiology, Biomarkers & Prevention, 2014

Toy Example: Time spent and passing Math exams

- The Time spent and Math score/grade problem:

Student ID	1	2	3	4	5	6	7	8	9	10
Hours spent	4	6	10	14	4	7	12	22	3	15
Math Score	39	58	65	73	41	50	60	79	40	64
Grades	F	F	P	P	F	F	P	P	F	P

- Task: to train a classification model based on the hours spent and grades to predict unseen students' grades based on the hours spent

Logistic regression

Logistic regression: also known as logit regression, is a regression model where the prediction (dependent variable) is categorical, e.g., binary.

- **Goal**: to predict the probability that a given example belongs to the "1" class versus the probability that it belongs to the "0" class.
- **Importance**: An fundamental machine learning algorithm used in medical and social sciences
- **How**: use the logarithm of the **odds** (called **logit** or **log-odds**) to models the binary prediction (dependent variable) as a linear combination of independent variables, then use **logistic function**: converts log-odds to probability.

Odd ratio vs Probability

Probability: a number $p \in [0, 1]$ between 0 to 1 to describe how likely an event is to occur, or how likely it is that a proposition is true.
Odds: A number to describe to the probability of **a binary outcome**, which either present or absent, e.g., mortality. Specifically, the odds are the ratio of the probability that an outcome present, i.e., p to the probability that the outcome absent, i.e., $1 - p$

$$odds = \frac{p}{1-p}$$

Logit (aka. log-odds): the logarithm of the odds:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = -\log\left(\frac{1}{p} - 1\right) .$$

Odds ratio vs Probability

Example: We have a deck of 52 cards, so

- Q1: What is the **probability** (sometimes called risk) of drawing a card randomly from the deck and getting spades?
- Q2: What is the **odds** of drawing a spade?

Answer, there are 13 spades, so,

- A1: the probability (sometimes called risk) of drawing a card randomly from the deck and getting spades is $13/52 = 0.25$.
- A2: Since the probability of not drawing a spade is 1 - 0.25, so the odds is 0.25/0.75 or 1:3 (or 0.33 or 1/3 pronounced 1 to 3 odds).

Main assumptions of Logistic Regression

How: use the logarithm of the **odds** (called **logit** or **log-odds**) to models the binary prediction (dependent variable) as a linear combination of independent variables.

Details: Given n independent variables x_1 to x_n, and one dependent variable Y which is a random variable that follow Bernoulli distributin (binary) which is denoted as $p = P(Y = 1)$. The logistic model can be formally defined as:

$$\text{logit} = \log\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots + \theta_n x_n$$

Linear Regression

- **Linear Regression** (function approximation): approximating an underlying linear function from a set of noisy data
- **Problem definition**: we have N observations, i.e., $\{(\boldsymbol{x}_i, y_i)\}$, $i = 1, \cdots, N$, where
  - $\boldsymbol{x}_i$: independent variables, also called regressor which is a K dimensional vector $\boldsymbol{x}_i \in \mathbb{R}^K$, and $\boldsymbol{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{iK} \end{bmatrix}^\mathsf{T}$, e.g., hours spent
  - y_i: dependent variable, e.g., math score, which is a scalar $y_i \in \mathbb{R}^1$
- **Aim**: To approximate a linear regression model to predict the dependent variable

$$\hat{y}_i = \theta_0 + \sum_{k=1}^{K} \theta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \cdots, N$$

where θ_0 is the the interception and θ_k is the weight of the kth dimensional data. ε_i is the disturbance term or error variable.

Logistic regression

Let $x_0 = 1$, we have

$$\hat{y}_i = \theta_0 x_0 + \sum_{k=1}^{K} \theta_k x_{ik} + \varepsilon_i = \sum_{k=0}^{K} \theta_k x_{ik} + \varepsilon_i = \boldsymbol{\theta}^\intercal \boldsymbol{x}_i + \varepsilon_i,$$

where $\boldsymbol{\theta}$ is a vector $\boldsymbol{\theta} = \begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_K \end{bmatrix}^\intercal$, $\boldsymbol{x}_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{iK} \end{bmatrix}^\intercal$.

Logistic regression: for each pair of train samples, we aim to learn a function of the form:

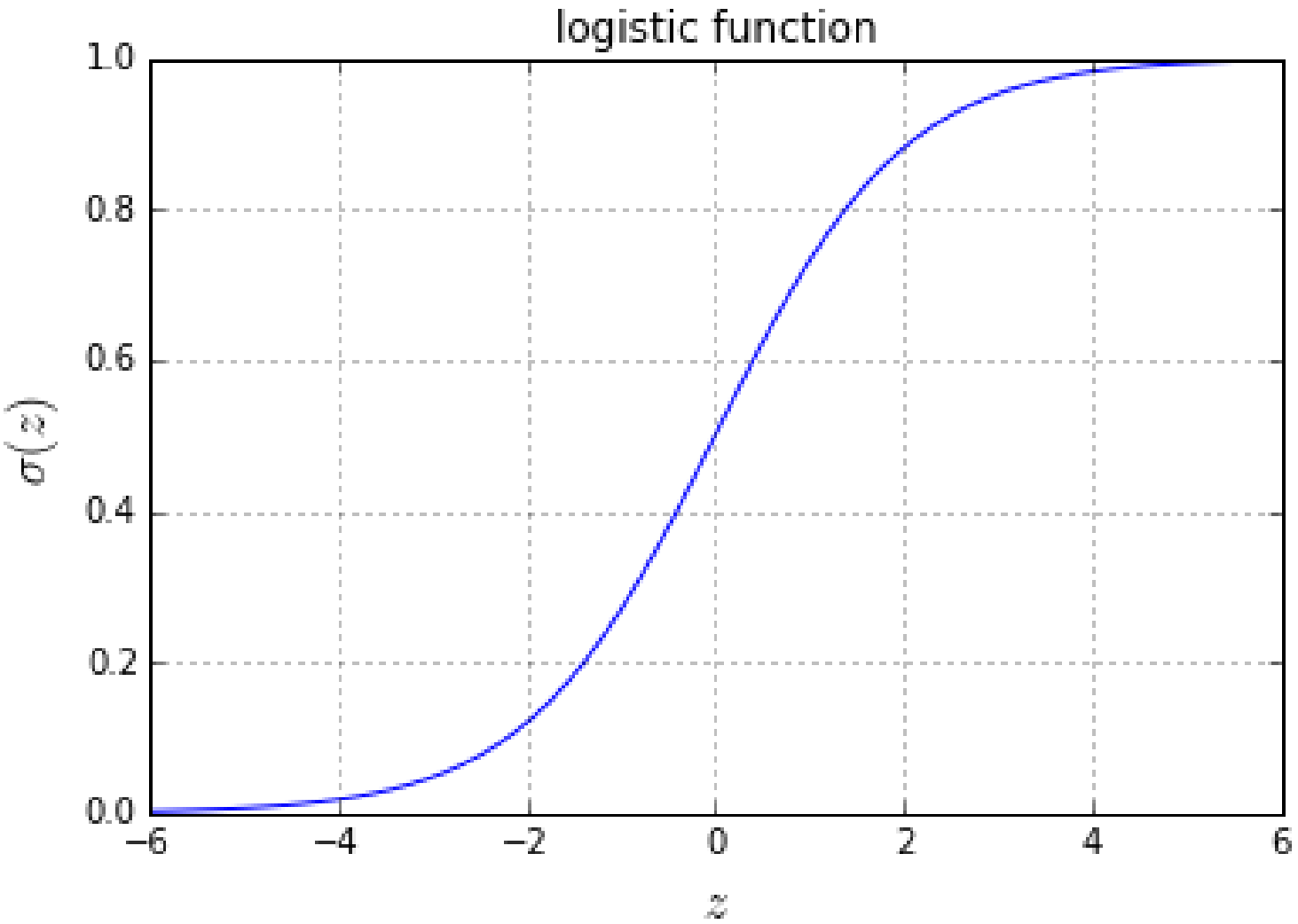
$$P(y_i = 1 \,|\, \boldsymbol{x}_i) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\intercal \boldsymbol{x}_i)}, \quad i = 1, 2, \cdots, N \tag{1}$$

and

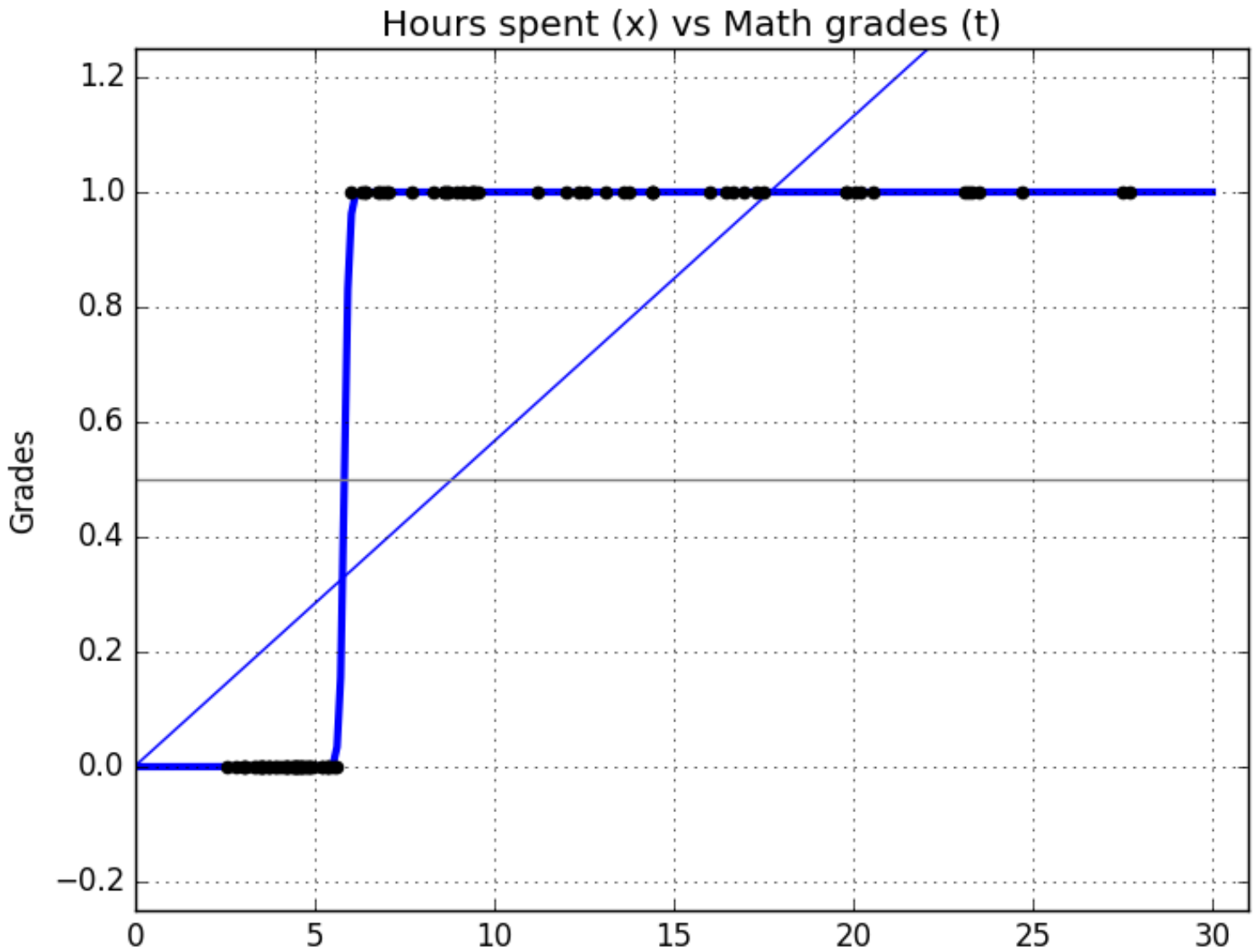
$$P(y_i = 0 \,|\, \boldsymbol{x}_i) = 1 - P(y_i = 1 \,|\, \boldsymbol{x}_i)$$

We define $\sigma(\boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{\theta}^\intercal \boldsymbol{x}_i)}$ as the "logistic" or "sigmoid" function

Logistic function



Logistic function



Maximum Likelihood Estimation of Logistic Regression

Solving logistic regression: We need to estimate the $K + 1$ unknown parameters θ equation 1.

Method: Maximum likelihood estimation – finding the set of parameters for which the probability of the observed data is greatest.

Question: How to derived the maximum likelihood estimator for logistic regression from the probability distribution of binary random variable Y?

Maximum Likelihood Estimation of Logistic Regression

Answer: Y is a binary random variable, we use Bernoulli distribution
Derivation: write equation 1 as a generalized linear model function parametrised by $\boldsymbol{\theta}$,

$$h_{\boldsymbol{\theta}}(X) = P(Y = 1 \mid X; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^\intercal X)}$$

and

$$P(Y = 0 \mid X; \boldsymbol{\theta}) = 1 - h_{\boldsymbol{\theta}}(X)$$

From Bernoulli distribution, we have

$$P(y \mid X; \theta) = \text{Bernoulli}(h_{\boldsymbol{\theta}}(X)) = h_{\boldsymbol{\theta}}(X)^y (1 - h_{\boldsymbol{\theta}}(X))^{1-y}$$

Maximum Likelihood Estimation of Logistic Regression

Derivation: Now we can define the likelihood function

$$\begin{aligned} L(\boldsymbol{\theta} \mid \boldsymbol{y}; \boldsymbol{x}) &= P(Y \mid X; \boldsymbol{\theta}) \\ &= \prod_i P(y_i \mid x_i; \boldsymbol{\theta}) \\ &= \prod_i h_{\boldsymbol{\theta}}(x_i)^{y_i} (1 - h_{\boldsymbol{\theta}}(x_i))^{(1-y_i)} \end{aligned}$$

We then use the negative logarithm of the likelihood function as the cost function, i.e., $-\log(L(\boldsymbol{\theta} \mid \boldsymbol{y}; \boldsymbol{x}))$:

$$-\log(L(\boldsymbol{\theta} \mid \boldsymbol{y}; \boldsymbol{x})) = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(h_{\boldsymbol{\theta}}(x_i)) + (1 - y_i) \log(1 - h_{\boldsymbol{\theta}}(x_i))] \tag{2}$$

Finally minimise it to obtain $\hat{\boldsymbol{\theta}}_{MLE} = \operatorname{argmin}_{\boldsymbol{\theta}} -\log(L(\boldsymbol{\theta} \mid \boldsymbol{y}; \boldsymbol{x}))$

Example 1: Time spent and passing Math exams

Let's solve example 1 using logistic regression. Denoting hours spent as the independent variable x_1 and the grades, i.e., fail or pass as binary dependent variable y, 1 means pass. To solve this problem, we implement logistic regression with Python Scikit Learn (See my source code here)

Example 1: Time spent and passing Math exams

Results interpretation:

- Model intercept, i.e., θ_0: -1.2725
- Model coefficients, i.e., θ_1: 0.2064

From the results, we calculate

- Odds:

$$\begin{aligned} odds &= \frac{p}{1-p} = \exp(\theta_0 + \theta_1 x_1) = \exp(-1.2725 + 0.2064 x_1) \\ &= \exp(0.2064 \cdot (-6.1652 + x_1)) \end{aligned} \tag{3}$$

- Probability:

$$\begin{aligned} P(Y=1) &= \frac{1}{1+\exp(-(\theta_0 + \theta_1 x_1))} \\ &= \frac{1}{1+\exp(-0.2064 \cdot (-6.1652 + x_1))} \end{aligned} \tag{4}$$

Example 1: Time spent and passing Math exams

Results interpretation:
Model intercept: The intercept is the log-odds of the event that you pass the math exam when you spend 0 hour. Using equations 3 and 4, we know $odds = 0.28$, roughly 1 to 3 odds. The probability is 0.22
Examples:

- From equation 3, if we increase our study time by one hour, we will increase the odds of passing exam by $\exp(\theta_1) = \exp(0.2064) \approx 1.2292$
- From equation 4, to just pass the the exam, i.e., even odds or probability $1/2$, we need to study at least $-\exp(\frac{\theta_0}{\theta_1}) = 6.1652$ hours
- Given some hours of a student spent, we can predict the probability the student will pass the exam. For example, for student who studies 10 hours, we have

$$P(Y = 1) = \frac{1}{1 + \exp(-0.2064 \cdot (-6.1652 + 10))} = 0.6881$$

Main assumptions of Logistic Regression

Main assumptions of Logistic Regression

- Binary outcomes
- Independent observations: observations are independent of each other. In other words, the observations should not come from repeated measurements or matched data.
- Low or no multicollinearity among the independent variables: the independent variables are not too highly correlated with each other.
- Linearity of independent variables and log odds. Note: this does not mean logistic regression assumes the dependent and independent variables are related linearly
- A large sample size.
  - **Rule of ten**: to fit a logistic regression model, we need at least 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 10*5 / .10 = 500.

Further reading

More about Logistic Regression

- An Introduction to Statistical Learning: Chapters 4.1-4.3