Beyond Linear Regression

From Logistic Regression to Neural Networks

Question

How to recognize hand-writing digits?

All Models are wrong, but some are useful.

George Box, British Statistician

Motivating Example

Suppose that we want to predict whether or not a student can pass an exam.

Here, the dependent variable, Y, is the outcome of the exam (1 = pass, 0 = failure).

The independent variables X are, for example, hours studied in class participation, background, …

QUESTION

Can we simply regress Y on X and use the regression result to predict a student's performance in the exam?

Numerical Example

Suppose that our independent variables are hours studied for the exam and in class participation. Based on the regression analysis, we come up with the following result:

$$\text{Outcome} = -0.12 + 0.075 \times \text{Hours} + 0.015 \times \text{Participation}$$

And the coefficients for hours and participation are both significant.

Numerical Example

Let us make predictions based on the linear regression.

Suppose that Alice studies for 10 hours and her participation is 4, then her outcome will be 0.69 --- well, you can say Alice is likely to pass the exam.

Suppose that Bob studies 5 hours and his participation is 0, then his outcome will be 0.225 --- you can say Bob is less likely to pass the exam.

Numerical Example

Let us make predictions based on the linear regression.

Suppose that Carol studies for 20 hours and her participation is 10, then her outcome will be 1.53 --- well, what can you say now?

Suppose that Denis studies 0 hours and his participation is 0, then his outcome will be -0.12 --- again, what can you say about this result?

Issue with linear regression

In the above example, the only possible outcome is 1 or 0. However, linear regression can make predictions like 2 or -1. They do not really make sense and cannot be used for our understanding.

This is not a small issue. Binary variables are everywhere around us!

Binary Variables

Binary variables only take two values (e.g., 0 and 1). Here are some examples:

Example outcome: pass vs. fail
Trial result: guilty vs. not guilty
Football result: win vs. lose
Medical testing: positive vs. negative
Gender detection: male vs. female

QUESTION

We already know linear regression does not work well for binary dependent variables. So, what would you do to solve this issue?

Solution

Suppose that we are predicting Y (binary with two values 1 and 0) based on X.

Because Y is a binary variable, instead of predicting the value of Y directly, we would like to predict the probability of $\Pr[Y = 1|X]$. For example, you can say that "If Alice studies for 10 hours for the exam and her attendance is 4, then she will pass the exam with probability 55%". This sounds much more reasonable.

Solution

Here, we need a function f to help us make the prediction:

$$\Pr[Y = 1] = f(X).$$

In contrast, in linear regression, we are doing this:

$$Y = f(X), \qquad f(X) = a + bX$$

Solution

In specifying the function f for making predictions for binary variables, we must be careful with the followings:

$f(X) \geq 0$ for all X: probability is nonnegative.
$f(X) \leq 1$ for all X: probability cannot exceed 1.
$f(X)$ is monotone in X: when X becomes larger, Y always becomes larger or smaller
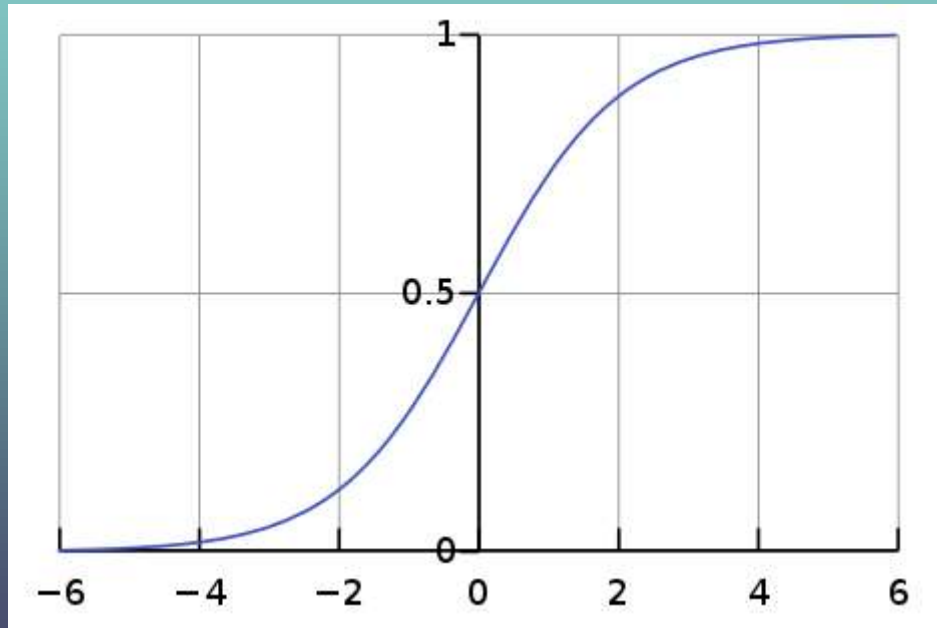
Any idea in mind?

Solution

Here, the function we choose is

$$f(Z) = \frac{1}{1 + \exp(-Z)}, Z = a + bX.$$

It is called the logistic function, and it satisfies the three properties that are previously noted.

The Logistic Function

Historical Note

The logistic function was first developed by French mathematician Laplace to model population growth: He used the logistic function to model the growth of populations, which has an S-shaped curve that starts slowly, accelerates, and then slows down.

Logistic Regression

Logistic Regression

Recall that the logistic regression specifies the following relationship:

$$\Pr[Y = 1] = \frac{1}{1 + \exp(-a - bX)}.$$

The real challenge is, given X and Y, how should we find out the values of a and b?

Idea

Recall what you did in linear regression…

You try to find out values of a and b that minimize the "square of errors", i.e., you want to minimize

$$\sum_i (Y_i - a - bX_i)^2$$

For all data points (X_i, Y_i).

Idea

Actually, we seek to find a and b that maximize the following:

$$Y_i f(X_i) + (1 - Y_i)\big(1 - f(X_i)\big)$$

Note that a and b are part of your function f. But why?

Idea

But least square does not work well here... We want to define a and b in other ways. Note that we seek to find out a and b such that

When $Y_i = 1$, $f(X_i) = \Pr[Y_i = 1|X_i]$ is as large as possible --- this means your prediction is accurate.

When $Y_i = 0$, $f(X_i)$ is as small as possible.

Idea

Let us translate the above reasoning into the following mathematics. For each observation (X_i, Y_i), we hope the following is as large as possible:

$$Y_i f(X_i) + (1 - Y_i)\big(1 - f(X_i)\big)$$

.

$$\text{Likelihood} = Y_i f(X_i) + (1 - Y_i)\big(1 - f(X_i)\big)$$

If $Y_i = 1$: The above likelihood is equal to $f(X_i)$, which we want to maximize. So we want to maximize the likelihood.

If $Y_i = 0$: The above likelihood is equal to $1 - f(X_i)$. We want to minimize $f(X_i)$ and again we want to maximize the likelihood.

In either case, we want the likelihood to be as large as possible.

Idea

Let us translate the above reasoning into the following mathematics. For each observation (X_i, Y_i), we hope the following is as large as possible:

$$Y_i f(X_i) + (1 - Y_i)\big(1 - f(X_i)\big)$$

.

And for all observations, we want to maximize the following:

$$\Pi_i Y_i f(X_i) + (1 - Y_i)\big(1 - f(X_i)\big)$$

Idea

This is called "maximum likelihood" --- we want to find out values of a and b that maximize the likelihood that was defined previously.

It is not easy to estimate the values of a and b that maximize the likelihood, and we skip the details here. However, if you are interested, you can check here.

Logistic Regression in R

Let us use an online database on university admission.

```r
mydata <-
read.csv("https://ximarketing.github.io/class
/ABOM/binary.csv")
head(mydata)
```

Logistic Regression in R

Let us use an online database on university admission.

```
mydata <-
read.csv("https://ximarketing.github.io/class
/ABOM/binary.csv")
head(mydata)
```

We have variables admit (binary), GMAT, GPA, and rank.

Logistic Regression in R

Next, we regress admission on GMAT, GPA and rank to see how these factors affect the admission decision, using logistic regression:

```r
logit <- glm(admit ~ gmat + gpa + rank, data = mydata, family = "binomial")
summary(logit)
```

Logistic Regression in R

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.449548   1.132846  -3.045  0.00233 **
gmat         0.002294   0.001092   2.101  0.03564 *
gpa          0.777014   0.327484   2.373  0.01766 *
rank        -0.560031   0.127137  -4.405 1.06e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic Regression in R

This result means that the best logistic model (i.e., the one with maximum likelihood) is

$$\text{Prob}(\text{admit} = 1) = \frac{1}{1 + \exp(-z)}$$

$$z = -3.44 + 0.0023\text{GMAT} + 0.777\text{GPA} - 0.56\text{Rank}$$

And all independent variables are significant at 5% level.

Logistic Regression in R

$$\text{Prob}(\text{admit} = 1) = \frac{1}{1 + \exp(-z)}$$

$$z = -3.44 + 0.0023\text{GMAT} + 0.777\text{GPA} - 0.56\text{Rank}$$

This result means: A higher GMAT or GPA helps you get admitted, while a larger Rank hurts (e.g., rank 2 is larger and worse than rank 1), which is intuitive.

Logistic Regression in R

	Dependent variable:
	admit
gmat	0.002**
	(0.001)
gpa	0.777**
	(0.327)
rank	-0.560***
	(0.127)
Constant	-3.450***
	(1.133)
Observations	400
Log Likelihood	-229.721
Akaike Inf. Crit.	467.442
Note:	*p<0.1; **p<0.05; ***p<0.01

We can organize the logistic regression output using the stargazer package. The details can be found in the "Introduction to R" lecture.

```
getwd()
library(stargazer)
stargazer(logit, out = "out.html", type = "html")
```

Probit Regression

In logistic regression, we adopt the logistic function to estimate $\Pr[Y = 1|X]$, which satisfies the properties that we listed. However, the logistic function is not the only function that satisfies those properties. Now, we introduce another function that can also make predictions about binary outcomes.
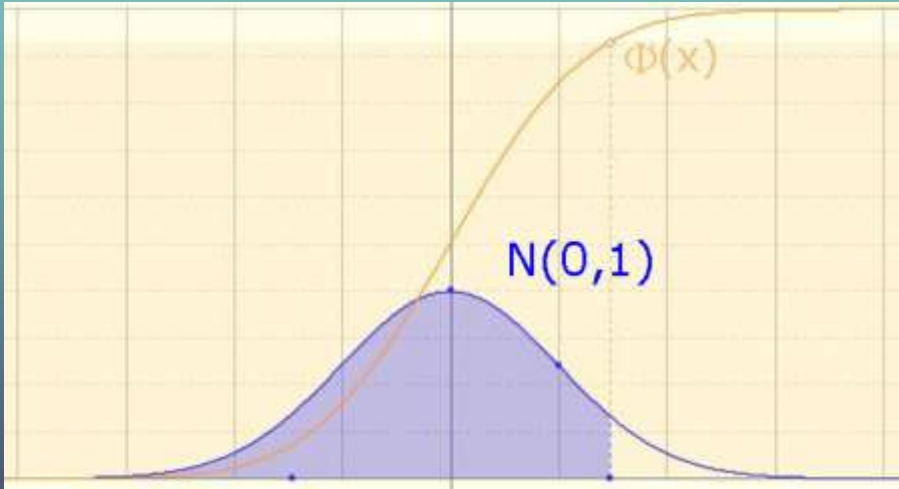
Probit Regression

Here, we use the cumulative distribution function of the standard normal distribution. Mathematically, suppose that $v \sim N(0,1)$ is a standard normal random variable, then we can define the cumulative distribution function Φ as

$$\Phi(Z) = \Pr[v \leq Z]$$

Probit Regression

The shape of the cumulative function is very similar to the logistic function that we have discussed previously.

Probit Regression

In Probit regression, we assume that the data satisfies the following model:

$$\Pr[Y = 1] = \Phi(a + b\,X)$$

And again, we want to find out the values of a and b.

Probit Regression in R

```r
probit <- glm(admit ~ gmat + gpa + rank, data = mydata, family = binomial(link = "probit"))
summary(probit)
stargazer(probit, out = "out.html", type = "html")
```

Contrasting Logistic and Probit Regression

	Dependent variable:
	admit
gmat	0.002**
	(0.001)
gpa	0.777**
	(0.327)
rank	-0.560***
	(0.127)
Constant	-3.450***
	(1.133)
Observations	400
Log Likelihood	-229.721
Akaike Inf. Crit.	467.442
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	admit
gmat	0.001**
	(0.001)
gpa	0.464**
	(0.195)
rank	-0.332***
	(0.075)
Constant	-2.092***
	(0.672)
Observations	400
Log Likelihood	-229.740
Akaike Inf. Crit.	467.481
Note:	*p<0.1; **p<0.05; ***p<0.01

Probit vs. Logistic Regression

They are similar models that yield similar (though not identical) inferences.

Logistic regression is more popular in healthcare.
Probit regression is more popular in political science.

But in most situations, it does not matter which method you choose to go with. Working with either will be fine.

Problem

In the above discussion, we discuss how to proceed when our dependent variable is binary. Now consider the following scenario:

Suppose that you are analyzing high school students' choice. Each student chooses one of the following options: A general program, an academic program, or a vocational program.

Career Choices

General   Vocational   Academic

Problem

For each high school students, we have some characteristics such as their test scores.

Question: How can we use the independent variables to make predictions for a student's career choice?

A Historical Note

Transportation infrastructures often involve massive investments. For instance, the HK-Zhuhai-Macau Bridge cost more than RMB 127 Billion, but there is little traffic on the bridge. It is a huge waste of money and resources.

A Historical Note

单是香港，港府至今合共向立法会申请了超过1200亿港元拨款，用作大桥相关的工程，包括大桥的香港接线、人工岛填海、口岸设施、屯门至赤鱲角的连接路等等。如果以香港人口约740万人作计算，即是每位香港人为大桥工程支付1.6万港元。近日正式通车的高铁香港段耗资844亿，两项工程均被称为"大白象工程"。

A Historical Note

时任广东省省长黄华华2008年曾经表示，港珠澳大桥36年便可以回本。

根据2008年估算，大桥使用初期车流量为每天9.2千架次，到2030年会升至3.3万架次，但2016年，三地政府调低评估，2030年车流量会预计会是2.9万架次。

政府在2008年興建港珠澳大橋可行性研究報告中指出，大橋於開通初期，車流量估計為每日9,200至14,000架次，事實上，大橋的使用量不似預期。政府在今年5月透露，大橋每日平均雙向車流量只有5,300多架次。

It is predicted that when the bridge is launched, the demand will be at least 9,200 vehicles per day. Now, in 2023, the actual traffic is still around 5,000 vehicles per day.

A Historical Note

There is also a lot of Airports with zero demand in Mainland China:

These Chinese Airports do not have even a single flight:

Tianshui 天水; Shennongjia 神农架; Liupanshui 六盘水; Kaili 凯里; Xinyang 信阳…

A Historical Note

This raises the question of how to predict the transportation demand before making such massive investments…

Daniel McFadden, an American economist, first noticed this issue. He developed a model to predict individual's demand for different transportation modes (e.g., driving, train, flying), and used the idea to predict the demand for the BART (Bay Area Rapid Transportation) services, a public transportation service in the Bay Area.

A Historical Note

A Historical Note

Daniel McFadden's model became so popular, and he won the Nobel Prize in Economics in 2000 for "his development of theory and methods for analyzing discrete choice."

Problem

For each high school student, we have some characteristics such as their test scores.

Question: How can we use the independent variables to make predictions for a student's career choice?

Idea

Denote the independent variables (e.g., test scores and gender) as X.

Students' choices are denoted by (Y_G, Y_V, Y_A): For example, $Y_A = 1$ implies that student joins an academic program.

Idea

Moreover, we place the following two restrictions on the dependent variables Y:

$Y_i \in \{0, 1\}$: A student either chooses one career or does not choose that career.

$Y_A + Y_V + Y_G = 1$: A student must choose one of the three careers.

Idea

Like in logistic regression, instead of predicting the value of Y_i directly, we want to find out the probability $\Pr[Y_i = 1]$. More specifically, we need three functions f_A, f_V, f_G that estimate $\Pr[Y_i = 1] = f_i(X)$.

Idea

These functions must satisfy the following properties:

$f_i \geq 0$: probabilities are nonnegative

$f_i \leq 1$: probabilities are below 100%

f_i is monotone with X

$f_G + f_V + f_A = 1$: A student must choose one of the three career paths.

Function

While the restrictions are strict, we can consider the following functions:

$$f_A = \frac{\exp(a_A + b_A X)}{\exp(a_A + b_A X) + \exp(a_G + b_G X) + \exp(a_V + b_V X)}$$

$$f_G = \frac{\exp(a_G + b_G X)}{\exp(a_A + b_A X) + \exp(a_G + b_G X) + \exp(a_V + b_V X)}$$

$$f_V = \frac{\exp(a_V + b_V X)}{\exp(a_A + b_A X) + \exp(a_G + b_G X) + \exp(a_V + b_V X)}$$

Function

Note that we can simplify the model as follows:

$$f_A = \frac{1}{1 + \exp(a_G - a_A + (b_G - b_A)X) + \exp((a_V - a_A) + (b_V - b_A)X)}$$

$$f_G = \frac{\exp((a_G - a_A + (b_G - b_A)X))}{1 + \exp(a_G - a_A + (b_G - b_A)X) + \exp((a_V - a_A) + (b_V - b_A)X)}$$

$$f_V = \frac{\exp((a_V - a_A + (b_V - b_A)X))}{1 + \exp(a_G - a_A + (b_G - b_A)X) + \exp((a_V - a_A) + (b_V - b_A)X)}$$

Function

So, we can normalize the model by taking academic as a benchmark. Thus, our model often has four parameters:

$$a'_G = a_G - a_A, b'_G = b_G - b_A$$

$$a'_V = a_V - a_A, b'_V = b_V - b_A$$

Again, we estimate the model using maximum likelihood method.

Function

Now, our new model becomes

$$f_A = \frac{1}{1 + \exp(a_G' + b_G'X) + \exp(a_V' + b_V'X)}$$

$$f_G = \frac{\exp(a_G' + b_G'X)}{1 + \exp(a_G' + b_G'X) + \exp(a_V' + b_V'X)}$$

$$f_V = \frac{\exp(a_V' + b_V'X)}{1 + \exp(a_G' + b_G'X) + \exp(a_V' + b_V'X)}$$

Estimation

We want to find values of a_G', b_G', a_V', a_G' that maximize

$$\Pi_i\ f_V Y_V + f_G Y_G + f_A Y_A$$

Multinomial Models in R

```r
library(foreign)
library(nnet)

mydata <- read.dta("https://ximarketing.github.io/class/ABOM/hsbdemo.dta")
```

Multinomial Models in R

```r
mydata$career <- relevel(mydata$prog, ref =
"academic")
```

Here, we set academic as a benchmark and compare other careers with it.

```r
result <- multinom(career ~ read + write + math +
science, data = mydata)
summary(result)
```

Multinomial Models in R

```
multinom(formula = career ~ read + write + math + science, data = mydata)

Coefficients:
         (Intercept)         read        write        math      science
general     4.393599  -0.05638993  -0.03242985  -0.09976556  0.09049715
vocation    8.701066  -0.05844241  -0.06060226  -0.12352885  0.05879027

Std. Errors:
         (Intercept)        read        write        math      science
general     1.437971  0.02806080  0.02792886  0.03302162  0.02913435
vocation    1.549432  0.03046132  0.02781245  0.03578399  0.02936508
```

Multinomial Models in R

	Dependent variable:	
	general (1)	vocation (2)
read	-0.056** (0.028)	-0.058* (0.030)
write	-0.032 (0.028)	-0.061** (0.028)
math	-0.100*** (0.033)	-0.124*** (0.036)
science	0.090*** (0.029)	0.059** (0.029)
Constant	4.394*** (1.438)	8.701*** (1.549)
Akaike Inf. Crit.	356.823	356.823
Note:	*p<0.1; **p<0.05; ***p<0.01	

Again, we can organize the result using R's stargazer package.

Multinomial Models in R

	Dependent variable:	
	general (1)	vocation (2)
read	-0.056** (0.028)	-0.058* (0.030)
write	-0.032 (0.028)	-0.061** (0.028)
math	-0.100*** (0.033)	-0.124*** (0.036)
science	0.090*** (0.029)	0.059** (0.029)
Constant	4.394*** (1.438)	8.701*** (1.549)
Akaike Inf. Crit.	356.823	356.823
Note:	*p<0.1; **p<0.05; ***p<0.01	

We can see that when a student's read score is higher, he or she is less likely to join general and vocation programs. In other words, he or she is more likely to join an academic program. Interestingly, when one's science score is higher, he or she is likely to join a general program.

Multinomial Models in R

	Dependent variable:	
	general (1)	vocation (2)
read	-0.057**	-0.059*
	(0.028)	(0.031)
write	-0.031	-0.054*
	(0.028)	(0.028)
math	-0.101***	-0.125***
	(0.033)	(0.036)
science	0.090***	0.059**
	(0.029)	(0.030)
factor(schtyp)private	-0.643	-1.781**
	(0.539)	(0.800)
Constant	4.529***	8.694***
	(1.457)	(1.568)
Akaike Inf. Crit.	353.849	353.849
Note:	*p<0.1; **p<0.05; ***p<0.01	

Now we revise the model to include school type into the logistic regression. We can see that, compared with public school students, private school students are less likely to join vocation programs.

Other Applications of Multinomial Model

Multinomial models are commonly used for discrete choice problems.

Which brand of milk is the consumer going to buy?

Is the consumer going to buy an EV, a traditional car, or simply do not buy?

Which university to apply for?

Occupational choice…

Other Dependent Variables

So far, we have linear regression models and logistic/multinomial logit models. How to choose your models?

Y is a real number: linear regression

Y = A, B, C, … multinomial logit models.

Y = 0, 1, 2, 3: what should we choose?

$Y \geq 0$: what should we choose?

Other Dependent Variables

When considering decisions such as purchase quantity (i.e., number of bottle waters to purchase from the supermarket), your dependent variable is something like $Y = 0, 1, \ldots$

In this case, you can use Poisson/Negative Binomial models.

Poisson Regression

In many scenarios our dependent variable is a nonnegative integer. Examples include:

The number of children in a family
The number of classes you take in a semester
The number of books a consumer purchases

Poisson Regression

Consider the following model: We have independent variables X and we want to use them to predict the dependent variable, Y. Here, the dependent variable is a nonnegative integer.

For example, given the consumer's age, gender, income, we want to predict how many bottles of milk the consumer will purchase in a week.

Poisson Regression (Optional Topics)

Idea: we assume that the dependent variables follows the Poisson distribution:

$$\Pr[Y_i = y | \lambda_i] = \frac{\exp(-\lambda_i)\lambda_i^y}{y!}$$

where $y!$ is the factorial function and

$$\lambda_i = \exp(\beta X)$$

We are looking for the parameter β.

Poisson Regression in R

```
mydata <-
read.csv("https://ximarketing.github.io/class/ABOM/poisson_
sim.csv")
```

Here, prog refers to the type of the program: 1 for general program, 2 for academic program, and 3 for vocational program.

Poisson Regression in R

```
result = glm(num_awards ~ factor(prog) + math,
family="poisson", data=mydata)
summary(result)
```

Because prog is simply a notation and does not have any numerical meaning, we take it as a fixed effect, and run the Poisson regression.

Other Dependent Variables

There are also occasions in which your parameters are "censored". For example, we want to predict how various factors affect a person's salary in Canada. The salary database in Canada is public, but it only discloses individuals whose salary is above CAD 100,000. In other words, salaries below 100,000 are "censored".

In this case, you can use Tobit models.

Regression Analysis
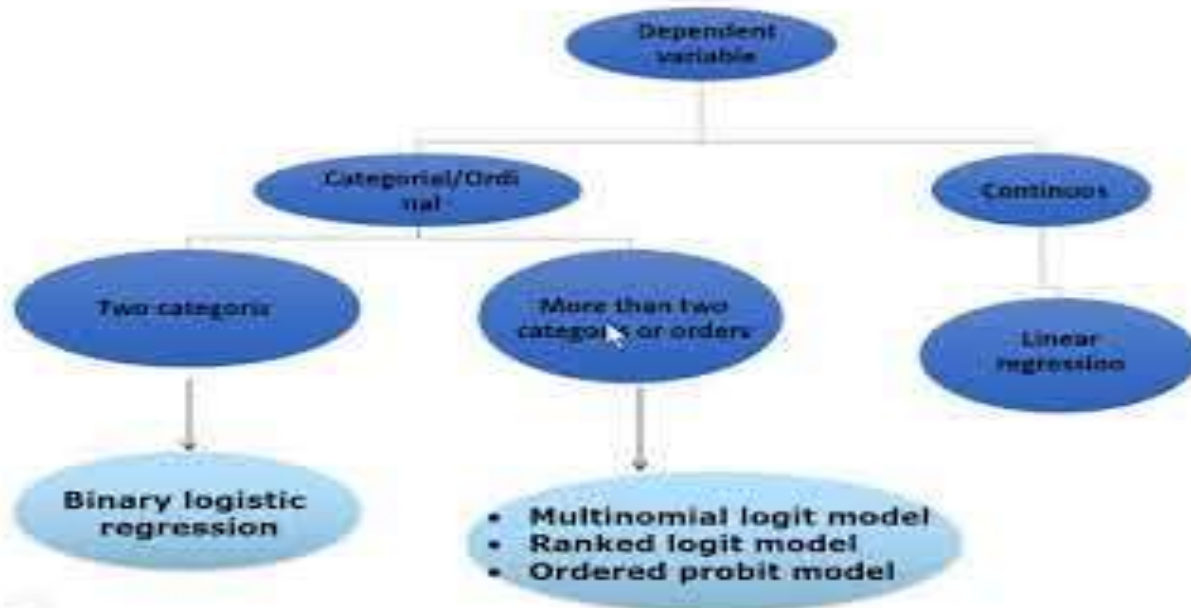
Going Beyond

Now, let's go beyond logistic regression in the marketing setting.

Let's consider some "machine learning" problems.

Going Beyond

Medical imaging:

Input: medical scans
Output: whether the patient has the disease

This is exactly what logistic regression is doing!

Going Beyond

Recognizing hand-writing digits ---

Input: hand-writing
Output: Y_0 to Y_9

This is exactly what multinomial models are doing!

Going Beyond

Autonomous driving ---

    Input: road condition
    Output: Y =
    turning left, turining right, ...

This is like multinomial logit models.

Going Beyond

Alpha Go ---

Input: current situation
Output: Y =
the place to move next

This is also like multinomial logit models.

Going Beyond

There are also a large number of similar scenarios such as

Gender detection (male or female?)

Facial recognition (Alice or not Alice?)

Voice recognition (which word?)

Neural Network

While multinomial logit models can help us make predictions about discrete outcomes, they are not powerful enough to address all the problems described above. We need to strengthen the model to make better predictions of these complex models.

One of such powerful models is neural network.

Neural Network

A basic neural network has three layers: an input layer, a hidden layer, and an output layer.

The input layer reads an input (e.g., the current road condition).

The hidden layer makes some calculation.

The output layer generates an output (e.g., turning right).

Neural Network
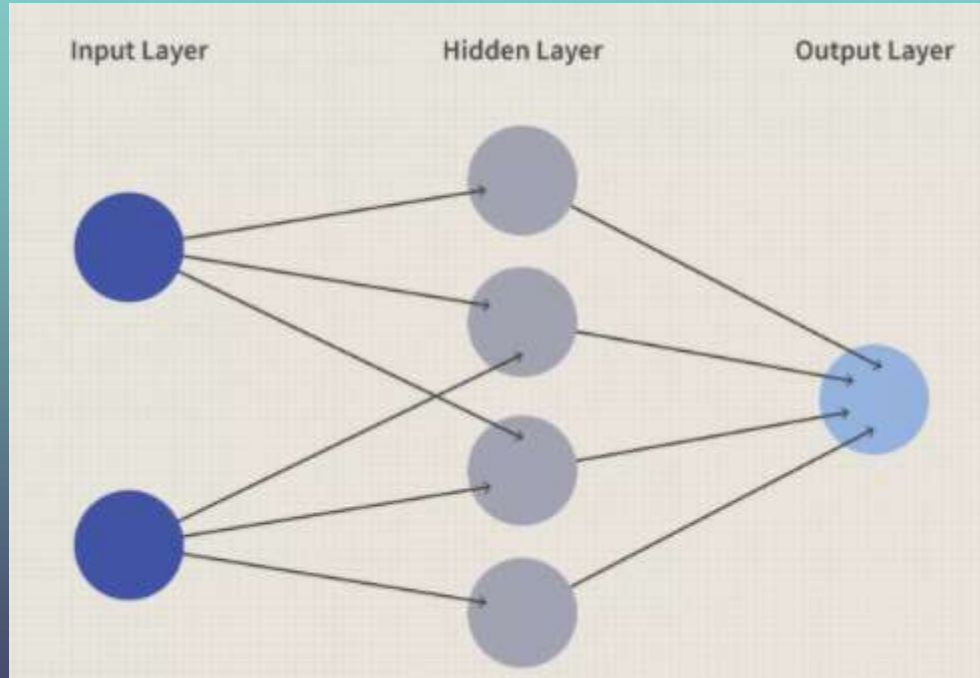
Neural Network

As shown in the figure, in a neural network, each layer consists of a number of units called "neurons". Neurons in the hidden layers are responsible of receiving information from the input layer, make calculation, and send output to the output layer.

What type of calculations does the hidden layer computes? It is essentially logistic function.

Neural Network

Neural networks are inspired by human brain. As human beings, we rely on our brain to make complex calculation such as "Is this person Alice?" "How to play the chess?" "Should I slow down the car a little bit?"

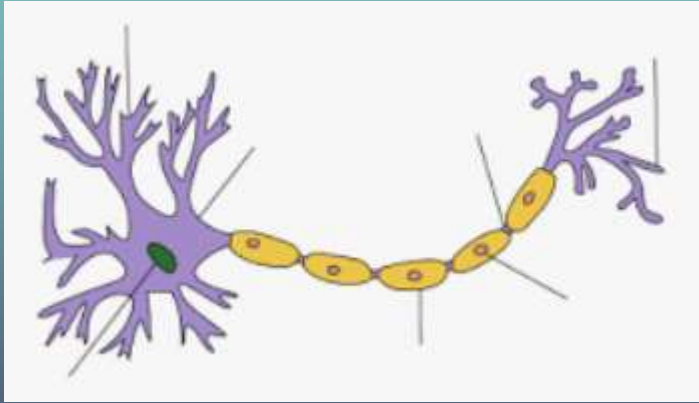While we can fluently handle these tasks, it is not yet clear how our brain makes these calculation…

Neural Network
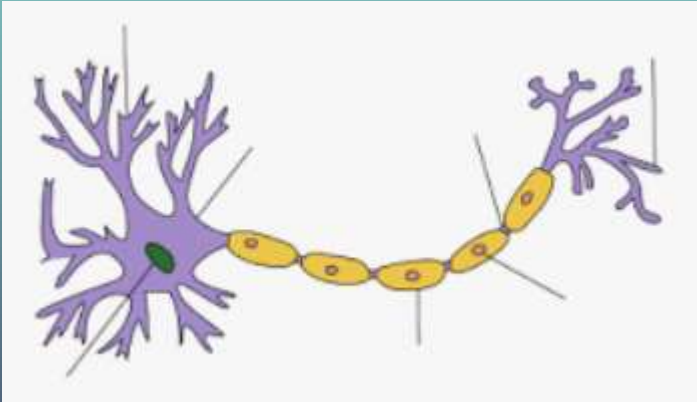
In our brain, there are billions of cells which are called "neurons". These neurons are capable of making calculations.

Neural Network

Some neurons receive input (e.g., light/sound) and convert the input into electric pulse. Some other neurons make calculations (i.e., convert one type of electric pulse into another type). Finally, some neurons output certain electric pulse.

Neural Network

The key is the calculations made in our brain. What type of calculation has been done by our neurons?

Surprisingly, neurons calculate the logistic function!

So, neuron network is just a simulation of our brain. This is why it has the word "neuron" in its name.

The Universal Approximation Theorem

In neural network, there is a key result called the universal approximation theorem.

Basically, it says that, when our neural network is large enough, it would be able to approximate any functions.

So, theoretically, neural networks should be capable of driving a car, recognizing human voice, writing a book or teaching a class.

The Universal Approximation Theorem

The issue is we have billions of neurons in our brain, and they make very complex calculations.

Due to the technology limit, we cannot build a neural network with billions of neurons --- it requires considerable memory and computing power that we cannot afford.

But this may become possible in the future.

Neural Network

To overcome the technology limit, scientists have also revised the basic neural network to make it easier to handle complex tasks. It can have multiple hidden layers, but the idea is more or less the same.
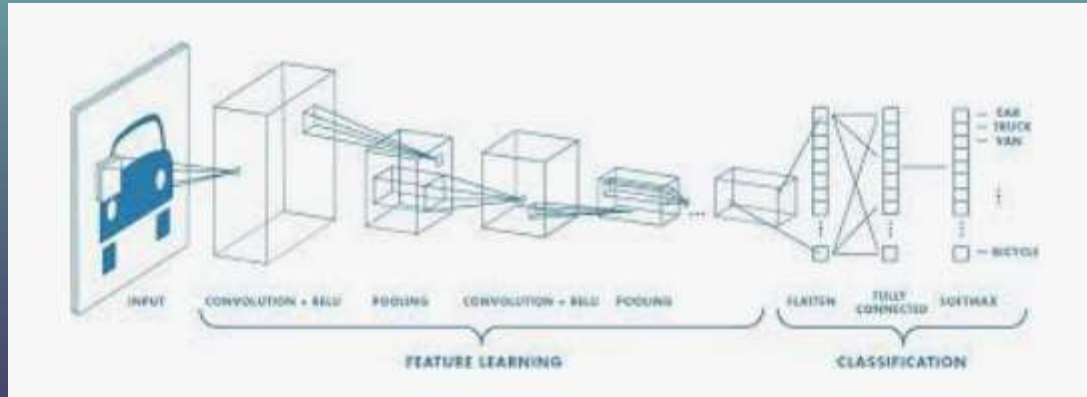
Type: Perceptron
Data Set: MNIST
Hidden Neurons: 2000
Synapses: 1191000
Synapses shown: 2%
Learning: WCor

0 1 2 3 4 5 6 7 8 9

www.cybercontrols.org

Next Class

Data workshop!

Bring your laptop with you.

If you are presenting next class, get your discussion topic ready.