# Logistic Regression

Nov 15 2025
Giảng viên: TS. Lưu Phúc Lợi
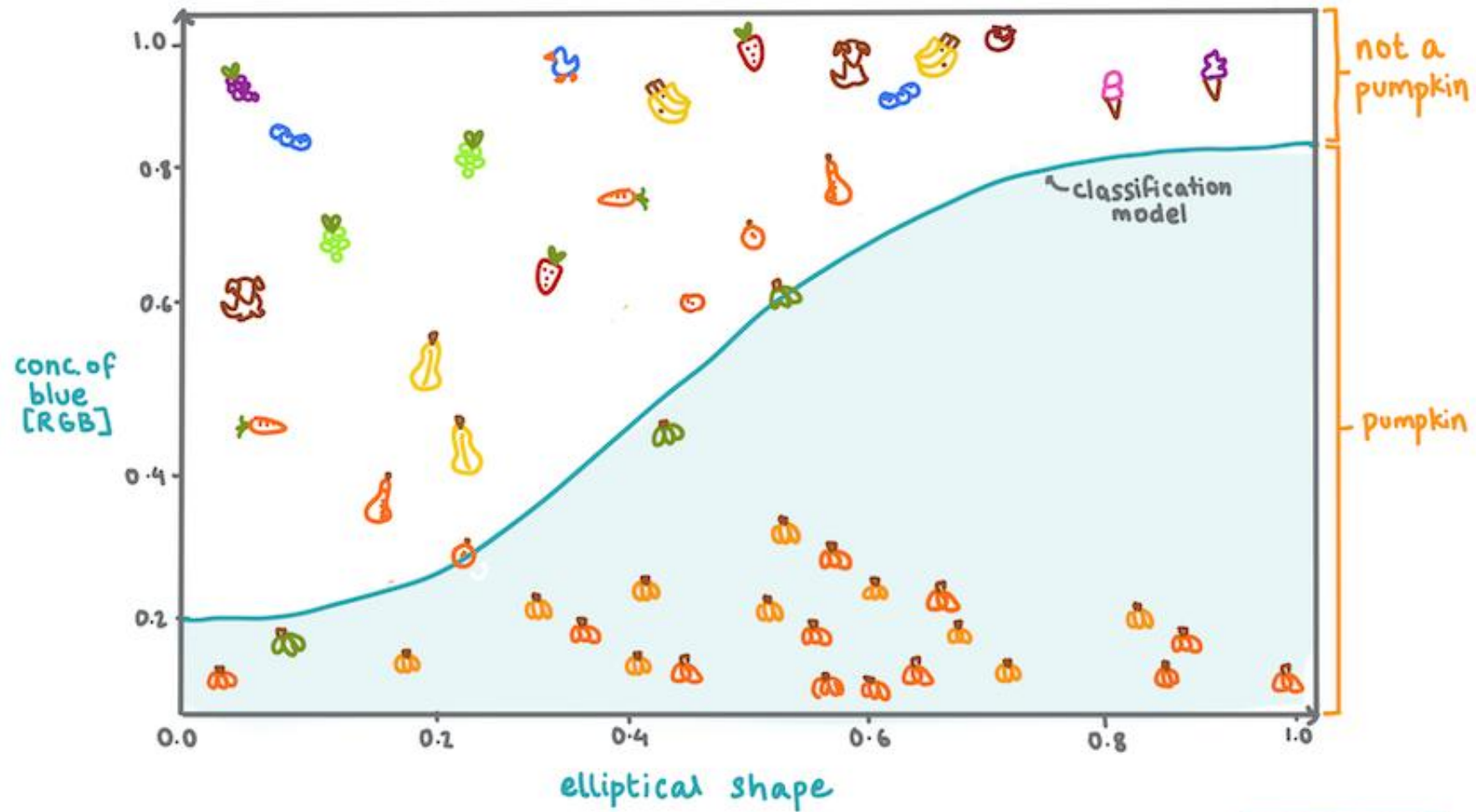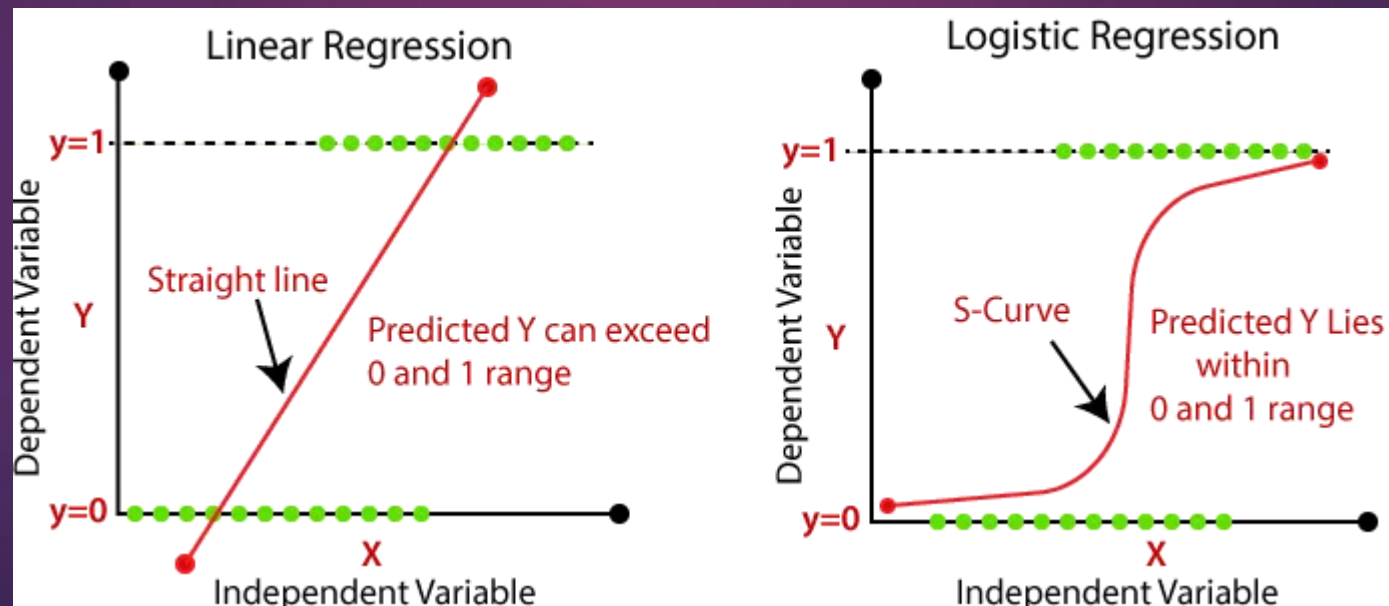Email: Luu.p.loi@googlemail.com
Zalo: 0901802182

# Content

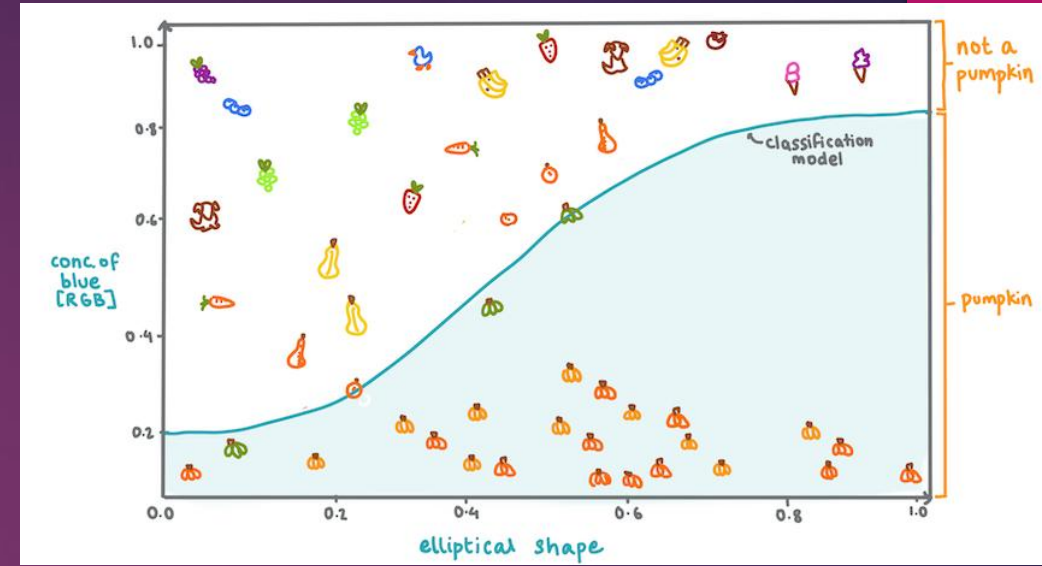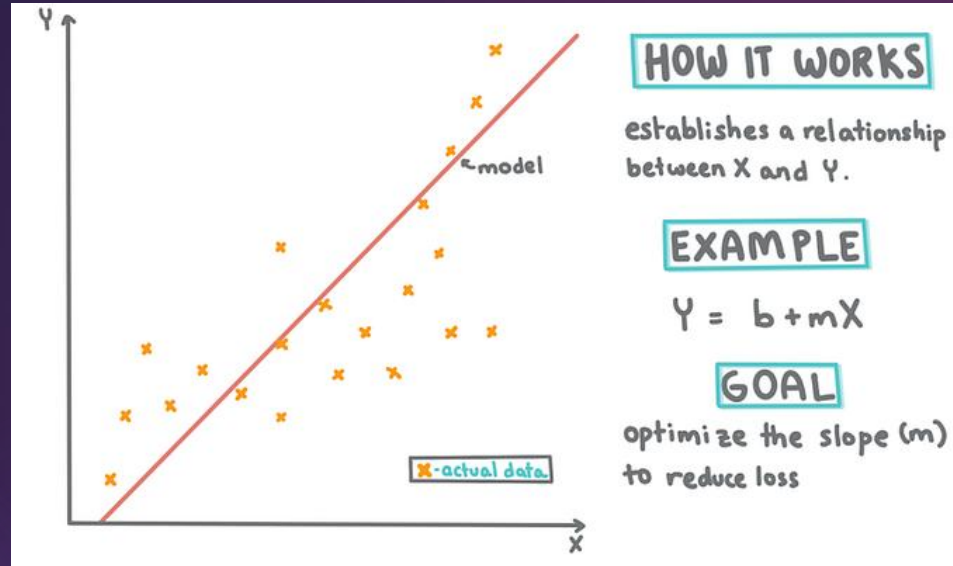1. Recap of Logistic Regression
2. Logistic Regression loss function
3. Practice Logistic Regression
4. Odds
5. Odds Ratio
6. Odds ratio or exp(B) of Logistic Regression

# Linear and Logistic Regression Models



https://medium.com/@tushar.babbar08/understanding-the-difference-between-linear-and-logistic-regression-models-bd81e30eb92b
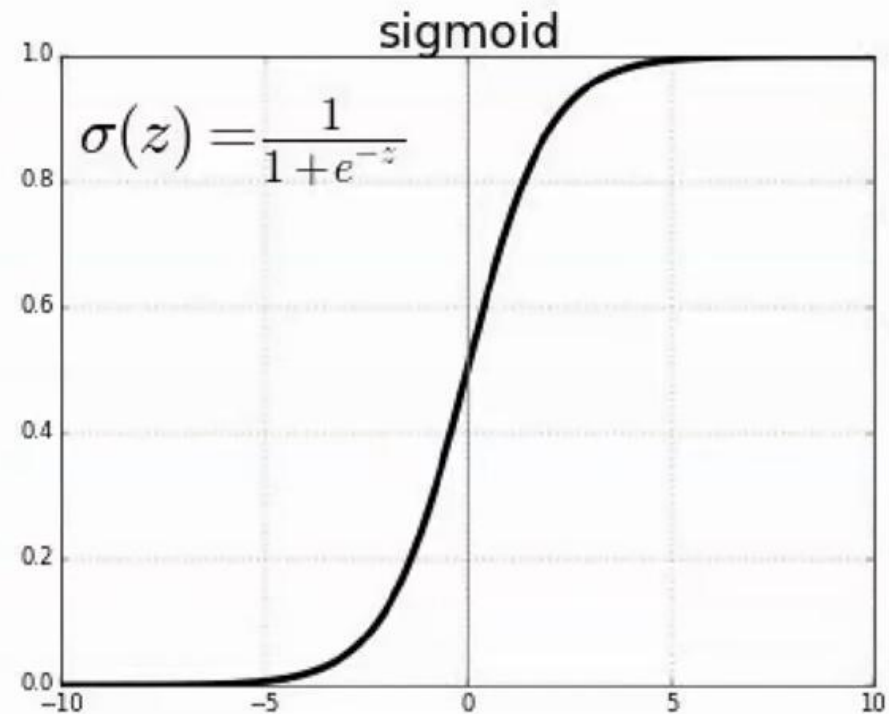
# Let's continue with binary classification

- Binary Classification is the simplest form of classification.

- It consists of 2 possible outcomes.

  1. Yes or No          2. Dog or Cat

- We normally represent the two possible options either as

- 1. O or 1          2. -1 or 1

- The **Sigmoid function** is commonly used with binary classification.

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

- As shown in the graph, the sigmoid function forces the output to always be between 0 or 1.

- Why is this useful for classification?

sigmoid

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The output of a sigmoid function is **always** between 0 and 1.

# $\sigma(x)$ models the probability of 1 happening

- Since $\sigma(x)$ function outputs numbers between 0 and 1, we use it to model the probability of 1 (success) given input of $x$.

$$\underbrace{q(x = 1|w) = \frac{1}{1 + e^{-\phi(x)^\top w}}}_{\textbf{prob of success}} \quad \text{and} \quad \underbrace{q(x = 0|w) = 1 - \frac{1}{1 + e^{-\phi(x)^\top w}}}_{\textbf{prob of failure}}.$$

- Note that the probability of failure is simply 1 - prob of success.

- The goal is to find the weight $w$ such that it gives us the probability of 1 and 0 given input $x$

  1. If $q(x = 1|w) > 0.5$, then we say 1 is more likely
  2. If $q(x = 1|w) < 0.5$, then we say 0 is more likely

- The idea of logistic Regression is just like Regression (instead of finding $f(x|w)$, we find $q(x|w)$

- We also put $q(x)$ into an error function such that minimizes prediction error

$$\min_w \text{ Prediction Error}(q(x|w)).$$

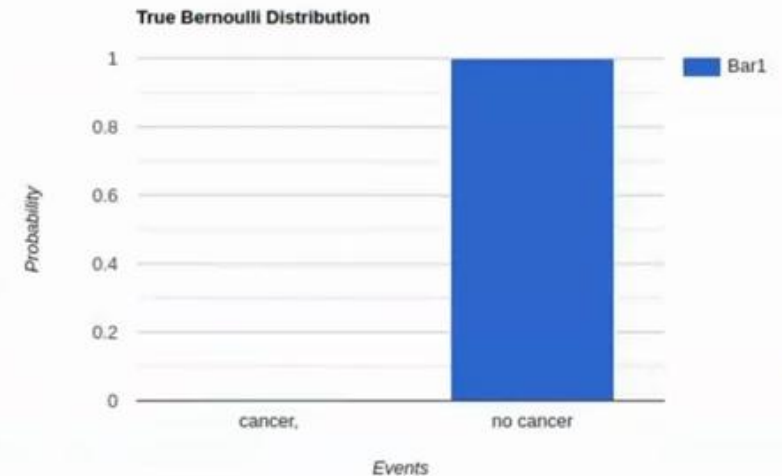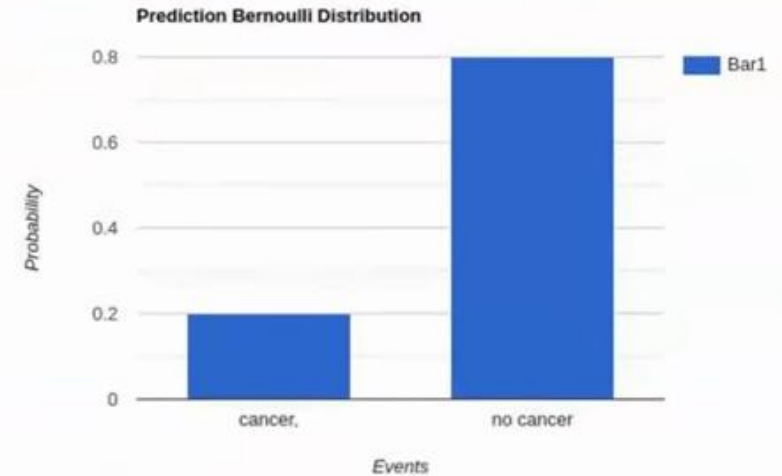# Once we know $w$, we can then make predictions!!

- For example, does this person $\hat{x}$ have cancer? (0 as no, 1 as yes)

$$q(\hat{x} = 1|w) = \frac{1}{1 + e^{-\phi(\hat{x})^\top w}}.$$

- If we get output of $q(\hat{x} = 1|w) = 0.2$ it tells us that

  - The probability of a "yes" event $= 1$ is 0.2
  - The probability of a "no" event $= 0$ is 0.8
  - Therefore, it is more likely that it is a no

- Let's assume that the truth is that this person doesn't have cancer, and therefore the probability should be 0, then we can see the result as 2 probability tables.

|  | Cancer | No Cancer |
|---|---|---|
| prediction probability | 0.2 | 0.8 |

|  | Cancer | No Cancer |
|---|---|---|
| Truth probability | 0 | 1 |

- This implies that we have 2 distinct Bernoulli distributions.

- The distance between them is the error.

- How do we go about measuring the error of our prediction?



Prediction Bernoulli Distribution



True Bernoulli Distribution

# Measuring the Error between distributions

- We can measure the error between our prediction and the truth by measuring the distance between the 2 Bernoulli Distributions.

- The standard method to measure the distance between 2 distributions is **Cross Entropy**.

- You may remember the equation from the last class as

$$H(p, q) = \mathbb{E}_{p(x)} \left[ \log \left( \frac{1}{q(x)} \right) \right] = \begin{cases} \sum_x \log \left( \frac{1}{q(x)} \right) p(x) & X \text{ is discrete} \\ \int_{\mathcal{X}} \log \left( \frac{1}{q(x)} \right) p(x) \, dx & X \text{ is continuous} \end{cases} \tag{1}$$

- We normally call the true distribution as $p(x)$ and the predicted distribution here as $q(x)$. The cross-entropy distance between our 2 distributions would have the equation

$$error = H(p, q) = \sum_{i=1}^{2} p(x) \log \left( \frac{1}{q(x)} \right) \tag{2}$$

You may remember from our lecture on information theory.
Cross-entropy is one of the most used objective in Machine Learning.

The **mean square error** is most common for regression.
The **Cross-Entropy** is the most common for classification.

# Example of Calculating the Cross-Entropy Error

Given the output of a sigmoid function as

| $q(x)$ | Cancer | No Cancer |
|---|---|---|
| prediction probability | 0.2 | 0.8 |

| $p(x)$ | Cancer | No Cancer |
|---|---|---|
| Truth probability | 0 | 1 |

The equation for cross-entropy as

$$error = H(p,q) = \sum_{j=0}^{1} p(x=j) \log\left(\frac{1}{q(x=j)}\right)$$

We can calculate the error as

$$H(p,q) = p(\text{cancer}) \log\left(\frac{1}{q(\text{cancer})}\right) + p(\cancel{\text{cancer}}) \log\left(\frac{1}{q(e=\cancel{\text{cancer}})}\right) = 0 \log\left(\frac{1}{0.2}\right) + 1 \log\left(\frac{1}{0.8}\right) = 0.22$$

If we had the perfect prediction of

- $q(\text{cancer}) = 0$ and $q(\cancel{\text{cancer}}) = 1$

Then the distance between prediction and truth would be 0, implying 0 error. Here, notice 0 error with a perfect prediction.

$$H(p,q) = p(x=\cancel{\text{cancer}}) \log\left(\frac{1}{q(x=\cancel{\text{cancer}})}\right) = 0 \log\left(\frac{1}{0}\right) + 1 \log\left(\frac{1}{1}\right) = 0.$$

# Calculating the Cross-Entropy Error with Multiple Samples

- In the last example, we computed the error from a single sample.

- However, in real life, we have thousands and thousands of samples.

- In these cases, we normally calculate the average error from the samples.

- Let $p(x_i)$ be the true probability distribution from samples $x_1, x_2, ....$

- Let $q(x_i)$ be the predicted probability distribution from samples $x_1, x_2, ...$

- Then, the average cross-entropy error $\bar{H}$ of $n$ samples would be

$$\underbrace{\bar{H}}_{\text{Average Error}} = \frac{1}{n} \sum_{i=1}^{n} H(p(x_i), q(x_i)) = \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j=1}^{n} p(x_i = j) \log \left( \frac{1}{q(x_i = j)} \right) \right] \quad (1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left[ p(x_i = 0) \log \left( \frac{1}{q(x_i = 0)} \right) + p(x_i = 1) \log \left( \frac{1}{q(x_i = 1)} \right) \right] \quad (2)$$

- In this example, we only have 3 samples ($n = 3$). We would essentially calculate each individual error and average them together.

| $q(x_i)$ | $x_i$ = Cancer | $x_i$ = No Cancer |
|----------|----------------|-------------------|
| $q(x_1)$ | 0.2 | 0.8 |
| $q(x_2)$ | 0.1 | 0.9 |
| $q(x_3)$ | 0.9 | 0.1 |

| $p(x_i)$ | $x_i$ = Cancer | $x_i$ = No Cancer |
|----------|----------------|-------------------|
| $p(x_1)$ | 0 | 1 |
| $p(x_2)$ | 0 | 1 |
| $p(x_3)$ | 1 | 0 |

# Practice Calculating Average Cross Entropy

Given the following predictions and the true labels, what's is the average cross-entropy error of the prediction? (use log base 2)

| $q(x_i)$ | $x_i$ = Cancer | $x_i$ = No Cancer |
|---|---|---|
| $q(x_1)$ | 0.2 | 0.8 |
| $q(x_2)$ | 0.1 | 0.9 |
| $q(x_3)$ | 0.9 | 0.1 |

| $p(x_i)$ | $x_i$ = Cancer | $x_i$ = No Cancer |
|---|---|---|
| $p(x_1)$ | 0 | 1 |
| $p(x_2)$ | 0 | 1 |
| $p(x_3)$ | 1 | 0 |

# Practice Calculating Average Cross Entropy

Given the following predictions and the true labels, what's is the average cross-entropy error of the prediction? (use log base 2)

| $q(x_i)$ | $x_i =$ Cancer | $x_i =$ No Cancer |
|----------|----------------|-------------------|
| $q(x_1)$ | 0.2 | 0.8 |
| $q(x_2)$ | 0.1 | 0.9 |
| $q(x_3)$ | 0.9 | 0.1 |

| $p(x_i)$ | $x_i =$ Cancer | $x_i =$ No Cancer |
|----------|----------------|-------------------|
| $p(x_1)$ | 0 | 1 |
| $p(x_2)$ | 0 | 1 |
| $p(x_3)$ | 1 | 0 |

**Solution:**
On the exam, you can simply write it out without simplifying

$$H(p(x_1), q(x_1)) = 0 \log\left(\frac{1}{0.2}\right) + 1 \log\left(\frac{1}{0.8}\right)$$

$$H(p(x_2), q(x_2)) = 0 \log\left(\frac{1}{0.1}\right) + 1 \log\left(\frac{1}{0.9}\right)$$

$$H(p(x_3), q(x_3)) = 1 \log\left(\frac{1}{0.9}\right) + 0 \log\left(\frac{1}{0.1}\right)$$

$$\bar{H} = \frac{1}{3} \sum_i H(p(x_i), q(x_i)) \approx 0.2086.$$

# Let's Write out the entire Logistic Regression Optimization Problem

Using the sigmoid function, we know that

$$q(x = 1|w) = \frac{1}{1 + e^{-\phi(x)^\top w}} \quad \text{and} \quad q(x = 0|w) = 1 - \frac{1}{1 + e^{-\phi(x)^\top w}}$$

Given $n$ samples, and the true distribution as $p(x_i = 0)$ and $p(x_i = 1)$. The logistic regression error we wish to minimize is

$$\min_w \mathbb{E}[\text{error}] = \min_w \frac{1}{n} \sum_i^n H(p(x_i), q(x_i))$$

$$= \min_w \frac{1}{n} \sum_i^n \left[ p(x_i = 0) \log \left( \frac{1}{q(x_i = 0|w)} \right) + p(x_i = 1) \log \left( \frac{1}{q(x_i = 1|w)} \right) \right]$$

$$= \min_w \frac{1}{n} \sum_i^n \left[ p(x_i = 0) \log \left( \frac{1}{1 - \frac{1}{1 + e^{-\phi(x_i)^\top w}}} \right) + p(x_i = 1) \log \left( \frac{1}{\frac{1}{1 + e^{-\phi(x_i)^\top w}}} \right) \right]$$

A closed-form solution doesn't exist for this problem, so we must use **gradient descent**.

# Practice Logistic Regression
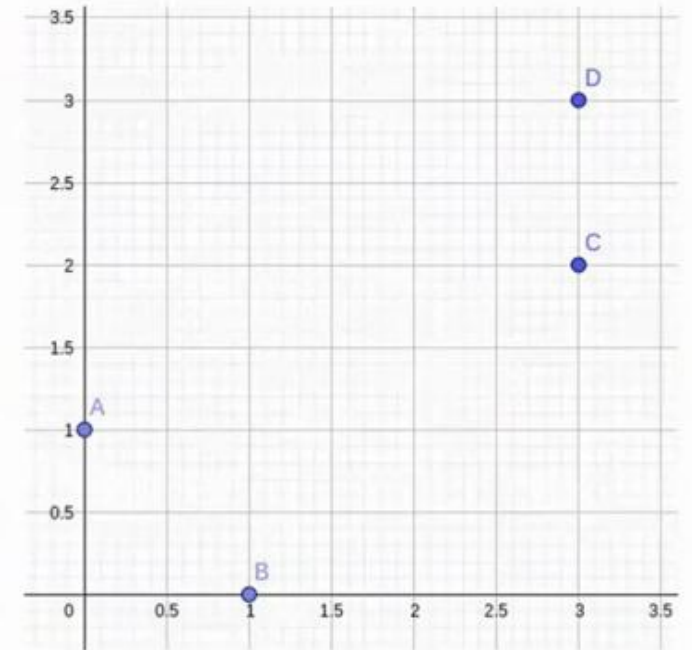
- Given the data

| $x_1$ | $x_2$ | cancer |
|-------|-------|--------|
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 3 | 2 | 1 |
| 3 | 3 | 1 |

- Write the code for logistic regression and identify the $\theta$ that will make the correct prediction.

# Performing Logistic Regression with Sklearn is super easy.

```python
#!/usr/bin/env python
import numpy as np
from sklearn.linear_model import LogisticRegression

X = np.array([  [0,1,1],
                [1,0,1],
                [3,2,1],
                [3,3,1]])

y = np.array(   [0,
                 0,
                 1,
                 1])
clf = LogisticRegression(random_state=0).fit(X, y)
print(clf.coef_)
print(clf.predict(X))
```

```
[[ 8.10680190e-01  6.28644615e-01 -1.86720392e-06]]
[0 0 1 1]
```

## Load the famous breast cancer data and perform binary classification on it.

```python
#!/usr/bin/env python
from sklearn.datasets import load_breast_cancer
data = load_breast_cancer()

print(data.data.shape)
```

```
(569, 30)
```

```python
print(data.target.shape)
print(data.target[0:10])
```

```
(569,)
[0 0 0 0 0 0 0 0 0 0]
```

```python
print(data.feature_names)
```

```
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
```

```python
print(list(data.target_names))
```

```
['malignant', 'benign']
```

# Example binary logistic regression

Let's say we are researchers, and we want to know whether a particular medication and a person's age have an influence on whether a person gets a certain disease or not.



Medication + Age = Diseased or not

So, the outcome we're interested in is whether the patients developed the disease or did not develop it. And our independent variables are medication and age of a person. Now with the help of a logistic Regression, we want to infer or predict the outcome variable based on the independent variables. Now let's take a look at what odds mean.

# Odds in logistic regression

Let's say we have two possible outcomes of something: success and failure. For example, if a therapy is successful or not. The probability that the therapy is successful is 0.7 (or 70%) and thus the probability of failure is 1 - 0.7 = 0.3.

| | Success | Failure |
|---|---|---|
| Probability | 0.7 | 0.3 |

https://numiqo.com/tutorial/odds-ratio-logistic-regression

# What are the odds?

Odds are defined as the ratio of the probability of success and the probability of failure. In other words, odds represent the ratio of the probability of an event happening to the probability of it not happening.

$$\frac{\text{Success}}{\text{Failure}}$$

If we look at our example, the odds are 0.7 divided by 0.3, which equals 2.33. This means the event "success" is 2.33 times more likely to happen than not.

$$\text{Odds} = \frac{\checkmark \; \text{Success}}{\times \; \text{Failure}} = \frac{0.7}{0.3}$$

So odds give us a measure of the likelihood of an event happening versus not happening.

$$\text{Odds} = \frac{\text{Probability of event } \textbf{happening}}{\text{Probability of it } \textbf{not happening}}$$

https://numiqo.com/tutorial/odds-ratio-logistic-regression
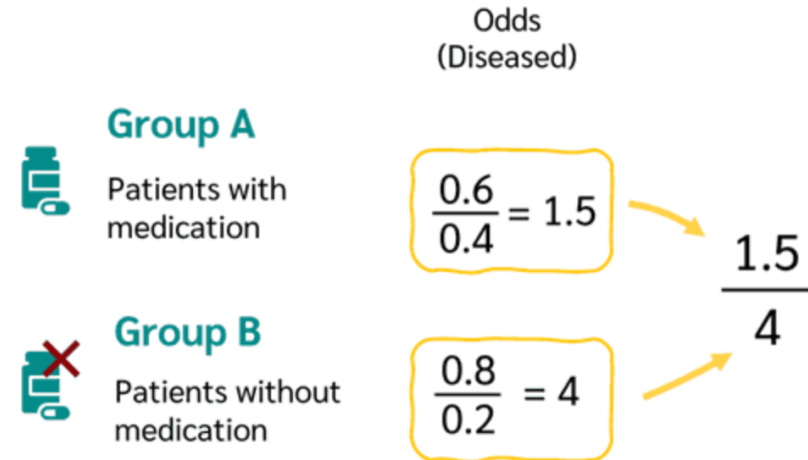
# What are Odds Ratios?

Let's say we have a Group A (Patients with medication) and a Group B (Patients without medication). In Group A, we calculated a probability of 60% (or 0.6) of getting diseased. So the odds of getting diseased is 0.6 divided by 0.4, which is 1.5.

In Group B, where the patients didn't get the medication, the probability of getting diseased is 80% (or 0.8). So the odds in Group B of getting diseased are 0.8 divided by 0.2, which is 4.

|  | Probability of Diseased | Odds (Diseased) |
|---|---|---|
| **Group A** Patients with medication | 0.6 | $\frac{0.6}{0.4} = 1.5$ |
| **Group B** Patients without medication | 0.8 | $\frac{0.8}{0.2} = 4$ |

With the odds ratio, we can now compare the two groups. To do this, we can compare the odds of getting the disease in Group A relative to the odds of getting the disease in Group B.

The odds ratio is simply calculated by dividing the odds in Group A by the odds in Group B. This results in an odds ratio of 0.38.

Odds (Diseased)

**Group A** Patients with medication $\frac{0.6}{0.4} = 1.5$

**Group B** Patients without medication $\frac{0.8}{0.2} = 4$

$\frac{1.5}{4}$

The odds ratio of 0.38 means that the odds of being diseased in Group A are 0.38 times the odds of being diseased in Group B.

Group A

$OR = \frac{1.5}{4} = 0.38$

Group B

https://numiqo.com/tutorial/odds-ratio-logistic-regression

# What are Odds Ratios?

- Of course we can also switch the order, then the odds ratio would be the odds in Group B divided by the odds in Group A. In this case, the odds ratio of approximately 2.67 means that the odds of being diseased in Group B are 2.67 times higher than the odds of being diseased in Group A.

- If the odds ratio is greater than 1, the event is more likely to occur in the first group. If it's less than 1, the event is less likely in the first group.

# Odds ratio in Logistic Regression

First of all, to calculate a logistic regression we need data. Let's say we have data from 50 patients. Our outcome variable is Disease, which is coded as 0 for 'not diseased' and 1 for 'diseased.' And we have two independent variables: Medication and Age. For the Medication variable, 0 indicates 'no medication,' and 1 indicates 'medication taken.'

| Medic. | Age | Disease |
| --- | --- | --- |
| 0 | 25 | 0 |
| 0 | 28 | 0 |
| 1 | 28 | 1 |
| 0 | 64 | 1 |
| 0 | 34 | 1 |
| 0 | 44 | 1 |
| 1 | 46 | 1 |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| 1 | 25 | 0 |

Now we can use this data to calculate a logistic regression.

Now we get the results of the logistic regression. Here we see the table that we will now take a closer look at.

| | Coefficient B | Standard error | z | p | Odds Ratio |
|---|---|---|---|---|---|
| Constant | -1.45 | 1.16 | 1.25 | .211 | 0.23 |
| Medication | -0.45 | 0.59 | 0.76 | .447 | 0.64 |
| Age | 0.04 | 0.02 | 1.69 | .09 | 1.04 |

In the first column, we can see the coefficients that define our model. These coefficients can be entered into the logistic regression formula.

| | Coefficient B | Standard error | z | p | Odds Ratio |
|---|---|---|---|---|---|
| Constant | -1.45 | 1.16 | 1.25 | .211 | 0.23 |
| Medication | -0.45 | 0.59 | 0.76 | .447 | 0.64 |
| Age | 0.04 | 0.02 | 1.69 | .09 | 1.04 |

$$P = \frac{1}{1 + e^{-(-1.45 - 0.45 \cdot Medication + 0.04 \cdot Age)}}$$

Now, we just need to enter a value for Medication—such as 1, indicating the patient received medication—and a value for Age, for example, 50.

$$P = \frac{1}{1 + e^{-(-1.45 - 0.45 \cdot 1 + 0.04 \cdot 50)}} = 0.55$$

Medication    Age

Then we can calculate the probability. In this case, the probability of being diseased is 0.55, or 55%. Okay, but we're not interested in the odds alone—we're interested in the odds ratio.

Again, the odds ratio is simply a comparison of the odds of an event occurring in two different groups.

Therefore, we just need to compare the odds of a person who took the medication with the odds of a person who did not take the medication. So to get the odds ratio, we just need to divide the odds of a person who took the medication by the odds of a person who did not take the medication. This results in an odds ratio of 0.64.

| | Value |
|---|---|
| Medication | 1 |
| Age | 50 |
| Probability | 0.55 |
| Odds | 0.55 / (1 - 0.55) = 1.22 |

| | Value |
|---|---|
| Medication | 0 |
| Age | 50 |
| Probability | 0.66 |
| Odds | 0.66 / (1 - 0.66) = 1.91 |

$$OR = \frac{1.22}{1.91} = 0.64$$

The odds ratio of 0.64 for Medication indicates that for individuals who took the medication, the odds of the outcome diseased are 0.64 times the odds for those who did not take the medication.

# Odds ratios of continuous variables

- With medication, we have two groups to compare. But what about a continuous variable like age? In this case, we simply look at what happens when we increase age by one unit. For example, we might compare the odds of the outcome for someone aged 50 versus someone aged 51. This allows us to calculate the odds ratio by comparing the two odds.

- In this case, we get an odds ratio of 1.04. So for each one-year increase in age, the odds of the outcome 'diseased' increase by a factor of 1.04.

# Odds ratio or exp(B)

There is one important thing: The odds ratio can actually be calculated simply by exponentiating each coefficient. So, exp(-0.45) is 0.64, which is the odds ratio of medication. And exp(0.04) is 1.04. which is the odds ratio for age.

$$e^{-0.45} = 0.64$$

| | Coefficient B | Standard error | z | p | Odds Ratio |
|---|---|---|---|---|---|
| Constant | -1.45 | 1.16 | 1.25 | .211 | 0.23 |
| Medication | -0.45 | 0.59 | 0.76 | .447 | 0.64 |
| Age | 0.04 | 0.02 | 1.69 | .09 | 1.04 |

$$e^{0.04} = 1.04$$

# Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM