

## Part 3: Applications

### L15: Logistic regression (1/2)

[Connections (multinomial) logistic regression, softmax, maximum entropy models, Lagrange multipliers]

Wolfgang Gatterbauer, Javed Aslam

cs7840 Foundations and Applications of Information Theory (fa24)

<https://northeastern-datalab.github.io/cs7840/fa24/>

10/28/2024

# Deriving Logistic regression (and the SoftMax) from Max Entropy and the Balance Equations

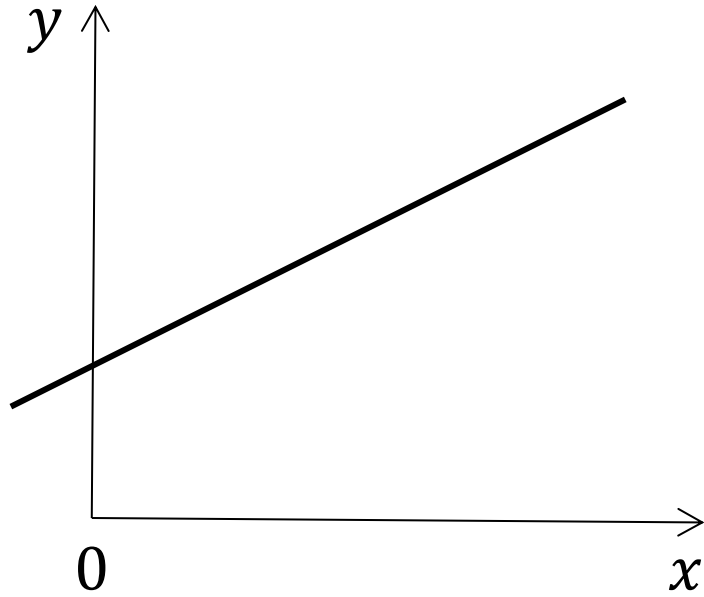
Following derivation and terminology of "balance equations" is based on following nice write-up: "John Mount. The equivalence of logistic regression and maximum entropy models, 2011:

<https://github.com/WinVector/Examples/blob/main/dfiles/LogisticRegressionMaxEnt.pdf>

"Balance equations" is the terminology used in Markov chains: [https://en.wikipedia.org/wiki/Balance\\_equation](https://en.wikipedia.org/wiki/Balance_equation)

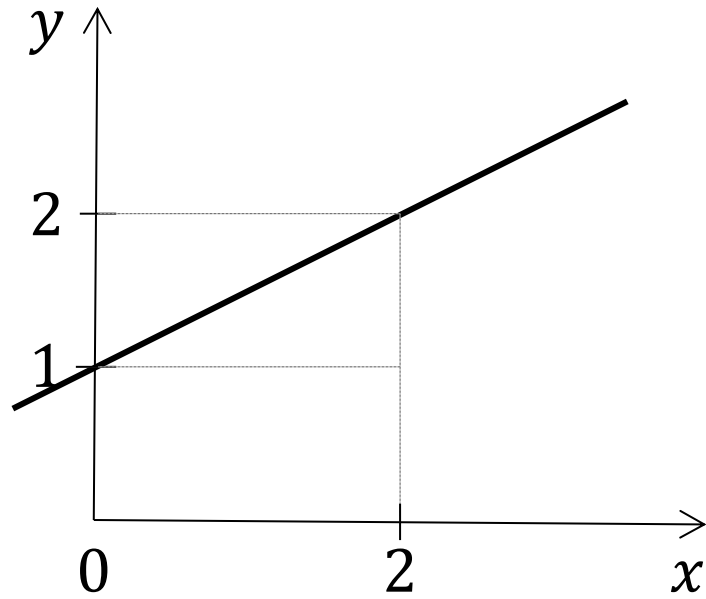
Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

# Non-homogenous linear combination of features



$$y = ax + b$$

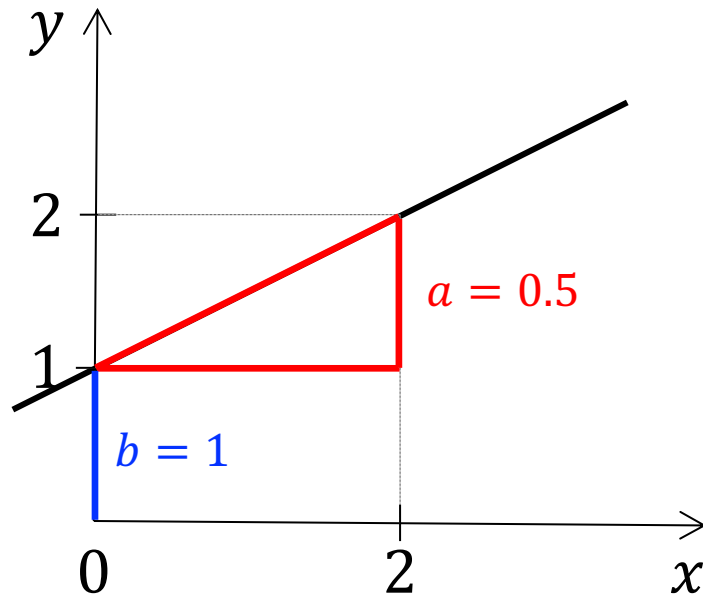
# Non-homogenous linear combination of features



$$y = ax + b$$

$$a = ?$$
$$b = ?$$

# Non-homogenous linear combination of features



$$y = ax + b$$

$$y = 0.5x + 1$$

$$y = a_1x_1 + a_2x_2 + \cdots a_mx_m + a_0$$

$$y = \mathbf{ax}$$

$$y = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}$$

$$y = \mathbf{ax} = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_m \end{pmatrix}$$

$(m+1)$ -dimensional  
parameter vector

# Our earlier formal setup from decision trees

EXAMPLE: Classifying days based on weather conditions.

Class label  $y_i$  denotes weather a particular event happened.

Columns denote  $m = 4$  features  $\{X_j\}_{j=1}^m$ .  
Domain  $\mathcal{X}_H$  of feature  $X_H$  is {high, normal}

day	Predictors				Response
	(O)utlook	(T)emp.	(H)umidity	(W)ind	(C)lass
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

$\langle \mathbf{x}_4, y_4 \rangle$

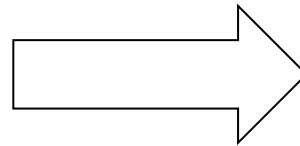
- Problem Setting
  - Set of possible instances  $\mathcal{X} = \mathcal{X}_O \times \dots \times \mathcal{X}_W$
  - Set of possible labels  $\mathcal{Y} = \{\text{yes, no}\}$  with size  $k = |\mathcal{Y}| = 2$  (binary)
  - Unknown target function  $f: \mathcal{X} \rightarrow \mathcal{Y}$
  - Set of function hypotheses  $H = \{h | h: \mathcal{X} \rightarrow \mathcal{Y}\}$
- Input: training examples of unknown target function  $f$   
 $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^n = \{\langle \mathbf{x}_1, y_1 \rangle, \dots, \langle \mathbf{x}_n, y_n \rangle\}$
- Output: Hypothesis  $h \in H$  that best approximates  $f$

Rows denote labeled instances  $\langle \mathbf{x}_i, y_i \rangle$ .

# Preparing the Tennis classification example for regression

Converting categorical variables into indicators

day	Predictors				Response
	(O)utlook	(T)emp.	(H)umidity	(W)ind	(C)lass
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

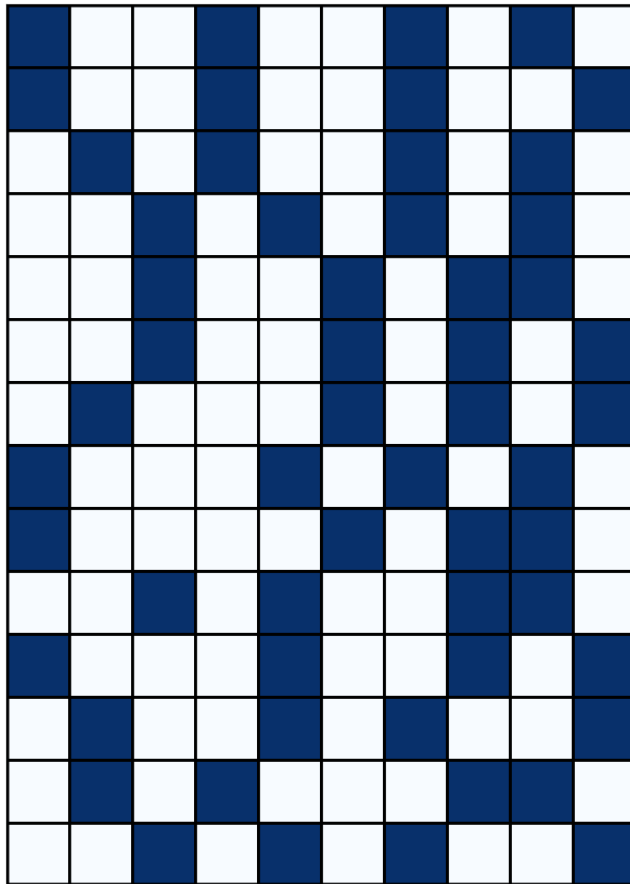


day	Predictors										Resp.	
	Outlook			Temp.			Hum.		Wind		Play	
	s	o	r	h	m	c	h	n	w	s	n	y
1	1			1			1		1		1	
2	1			1			1			1	1	
3		1		1			1		1			1
4			1		1		1		1			1
5			1			1		1	1			1
6			1			1		1		1	1	
7		1				1		1		1		1
8	1				1		1		1		1	
9	1					1		1	1			1
10			1		1			1	1			1
11	1				1			1		1		1
12		1			1		1			1		1
13		1		1				1	1			1
14			1		1		1			1	1	

# Preparing the Tennis classification example for regression

Converting categorical variables into indicators

X



Y

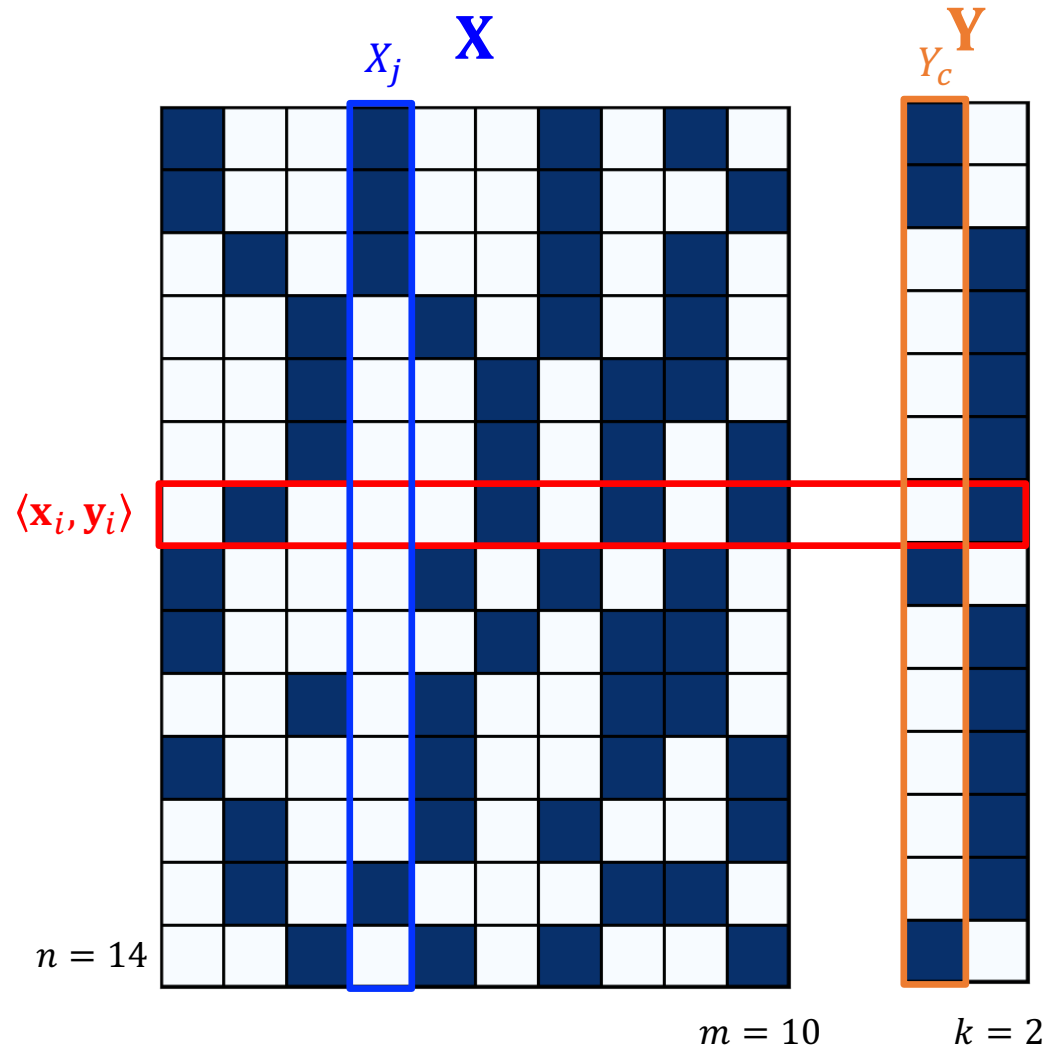


day	Predictors									Resp.		
	Outlook			Temp.			Hum.		Wind		Play	
	s	o	r	h	m	c	h	n	w	s	n	y
1	1			1			1		1		1	
2	1			1			1			1	1	
3		1		1			1		1			1
4			1		1		1		1			1
5			1			1		1	1			1
6			1			1		1		1	1	
7		1				1		1		1		1
8	1				1		1		1		1	
9	1					1		1	1			1
10			1		1			1	1			1
11	1				1			1		1		1
12		1			1		1			1		1
13		1		1				1	1			1
14			1		1		1			1	1	



# Preparing the Tennis classification example for regression

Converting categorical variables into indicators



We want to learn a target function that fits the labels well:

$$f(\mathbf{x}_i) \approx y_i$$

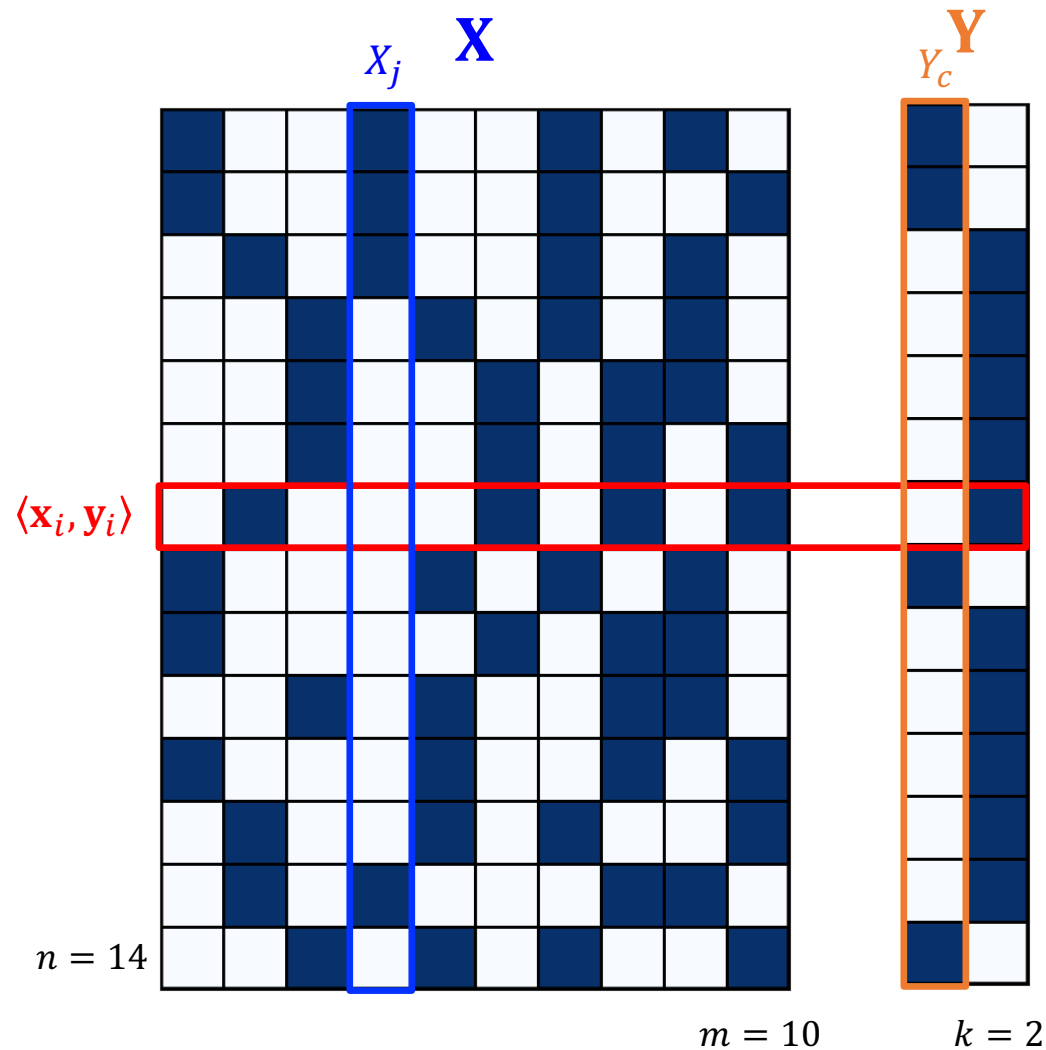
This is the original label  $\in [k]$

We will actually approximate the label indicators and fit an  $k$ -dimensional estimate function:

$$\mathbf{p}(\mathbf{x}_i)_c \approx y_{i,c}$$

This is an indicator  $\in \{0,1\}$

# Preparing the Tennis classification example for regression



What you get if you put it into  
sklearn LogisticRegression

Why so bad?



We want to learn a target function  
that fits the labels well:

$$f(\mathbf{x}_i) \approx y_i$$

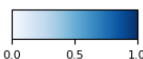
This is the original label  $\in [k]$

We will actually approximate the label  
indicators and fit an  $k$ -dimensional  
estimate function:

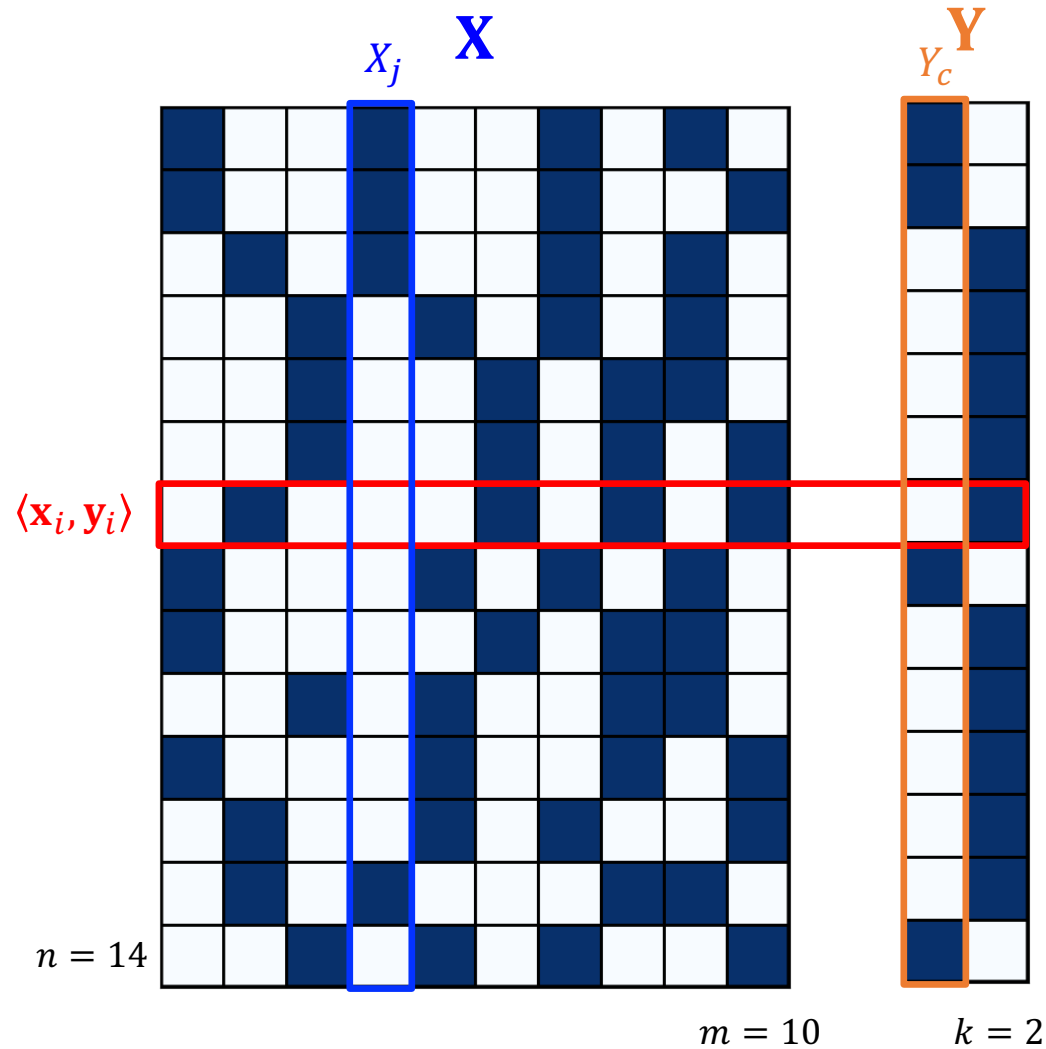
$$\mathbf{p}(\mathbf{x}_i)_c \approx y_{i,c}$$

This is an indicator  $\in \{0,1\}$

0.62	0.38
0.70	0.30
0.28	0.72
0.35	0.65
0.07	0.93
0.10	0.90
0.05	0.95
0.54	0.46
0.14	0.86
0.10	0.90
0.26	0.74
0.29	0.71
0.07	0.93
0.43	0.57



# Preparing the Tennis classification example for regression



Actually fits perfectly 😊  
Just need to tone down regularization...

We want to learn a target function that fits the labels well:

$$f(\mathbf{x}_i) \approx y_i$$

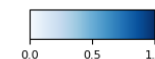
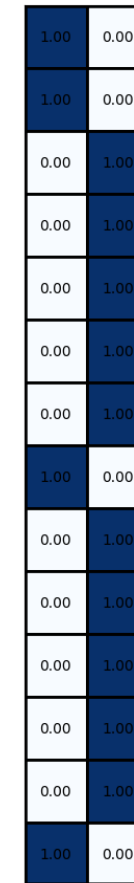
← This is the original label  $\in [k]$

We will actually approximate the label indicators and fit an  $k$ -dimensional estimate function:

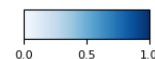
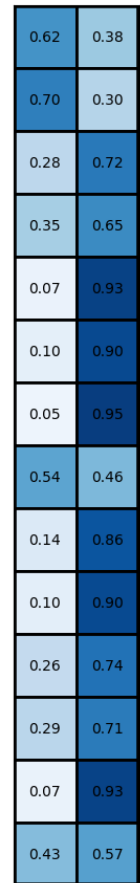
$$\mathbf{p}(\mathbf{x}_i)_c \approx y_{i,c}$$

← This is an indicator  $\in \{0,1\}$

Almost no regularization:  
 $C = 1e5$



Default regularization:  
 $C = 1$



# A more interesting example

X

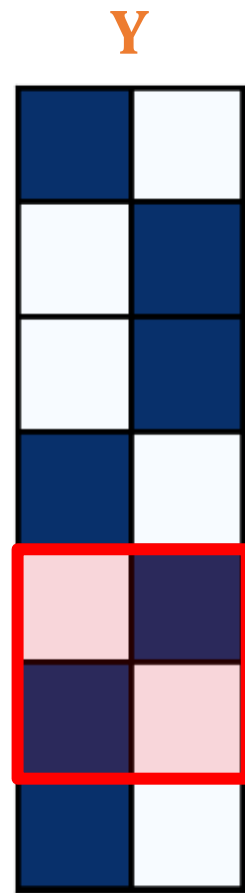
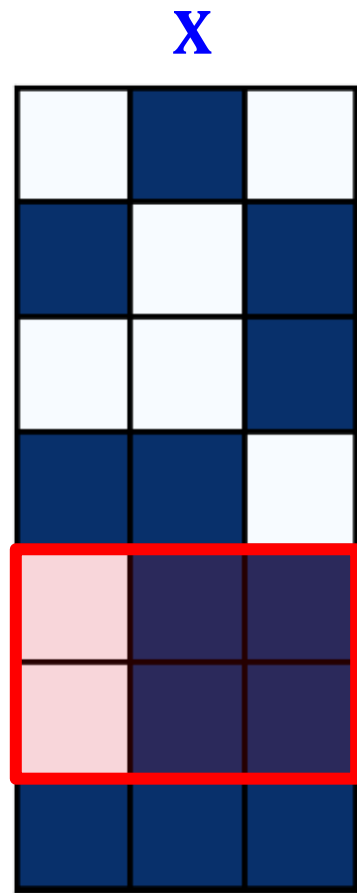
White	Dark Blue	White
Dark Blue	White	Dark Blue
White	White	Dark Blue
Dark Blue	Dark Blue	White
White	Dark Blue	Dark Blue
White	Dark Blue	Dark Blue
Dark Blue	Dark Blue	Dark Blue

Y

Dark Blue	White
White	Dark Blue
White	Dark Blue
Dark Blue	White
White	Dark Blue
Dark Blue	White
Dark Blue	White

Why more interesting ?

# A more interesting example



Default regularization

$C = 1$

**p(X)**



# A more interesting example

Almost no regularization  
 $C = 1e5$

$C = 100$

Default regularization  
 $C = 1$

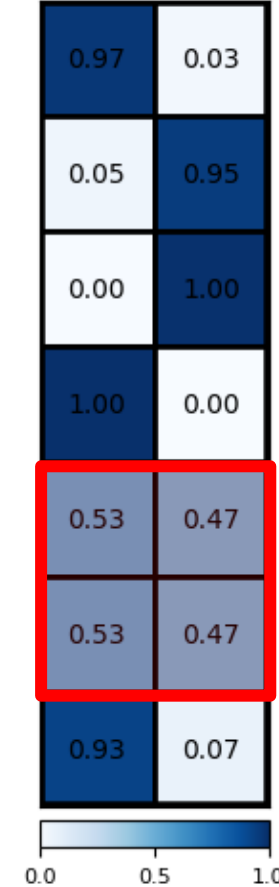
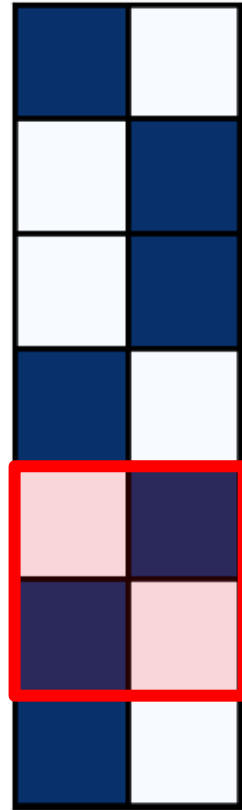
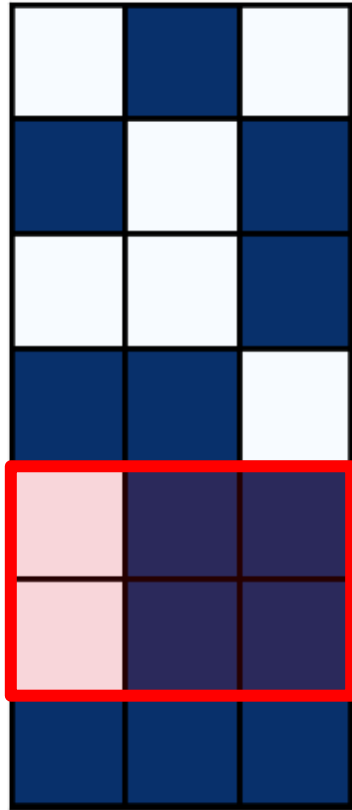
**X**

**Y**

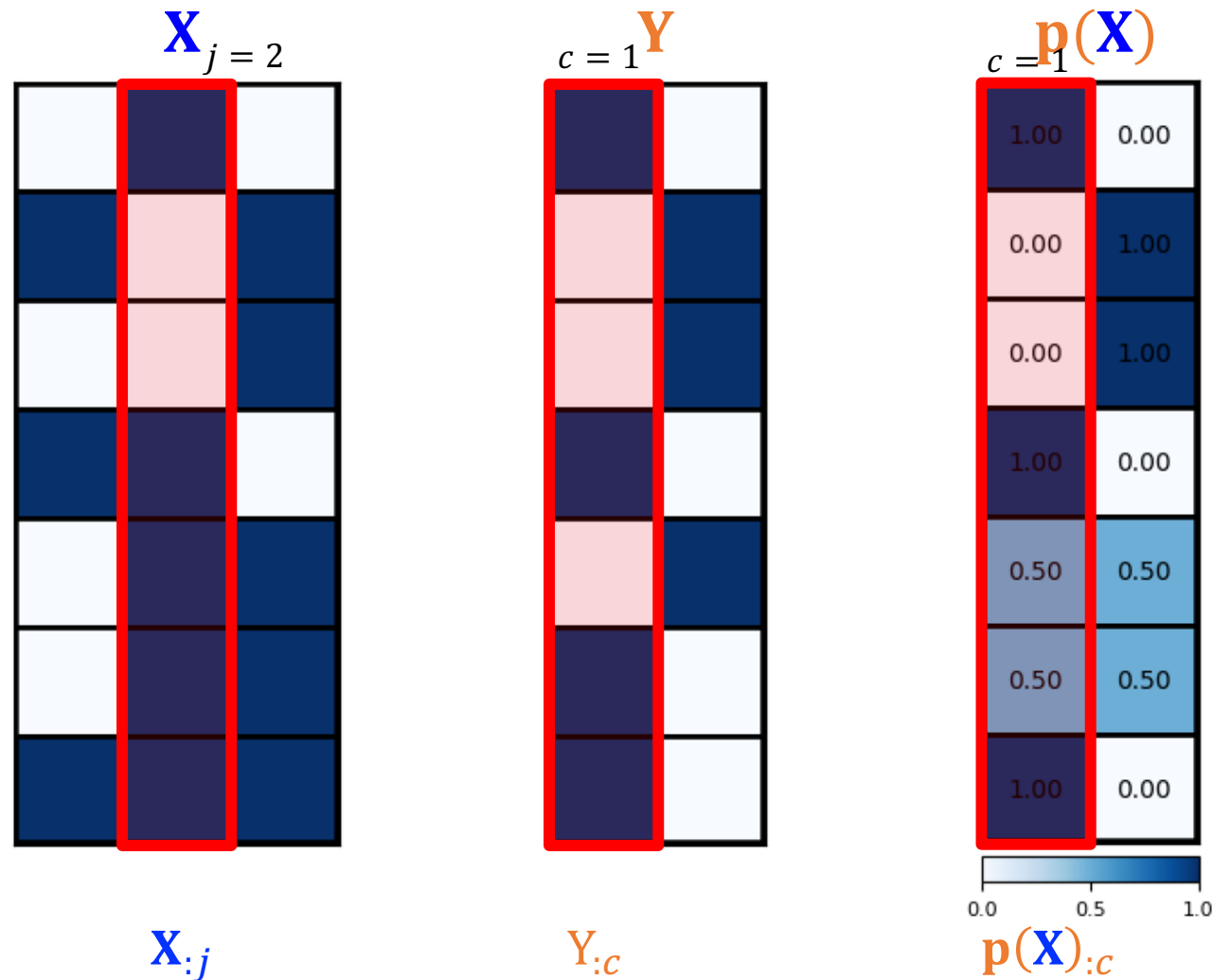
**p(X)**

**p(X)**

**p(X)**



# Balance Equations of Logistic Regression

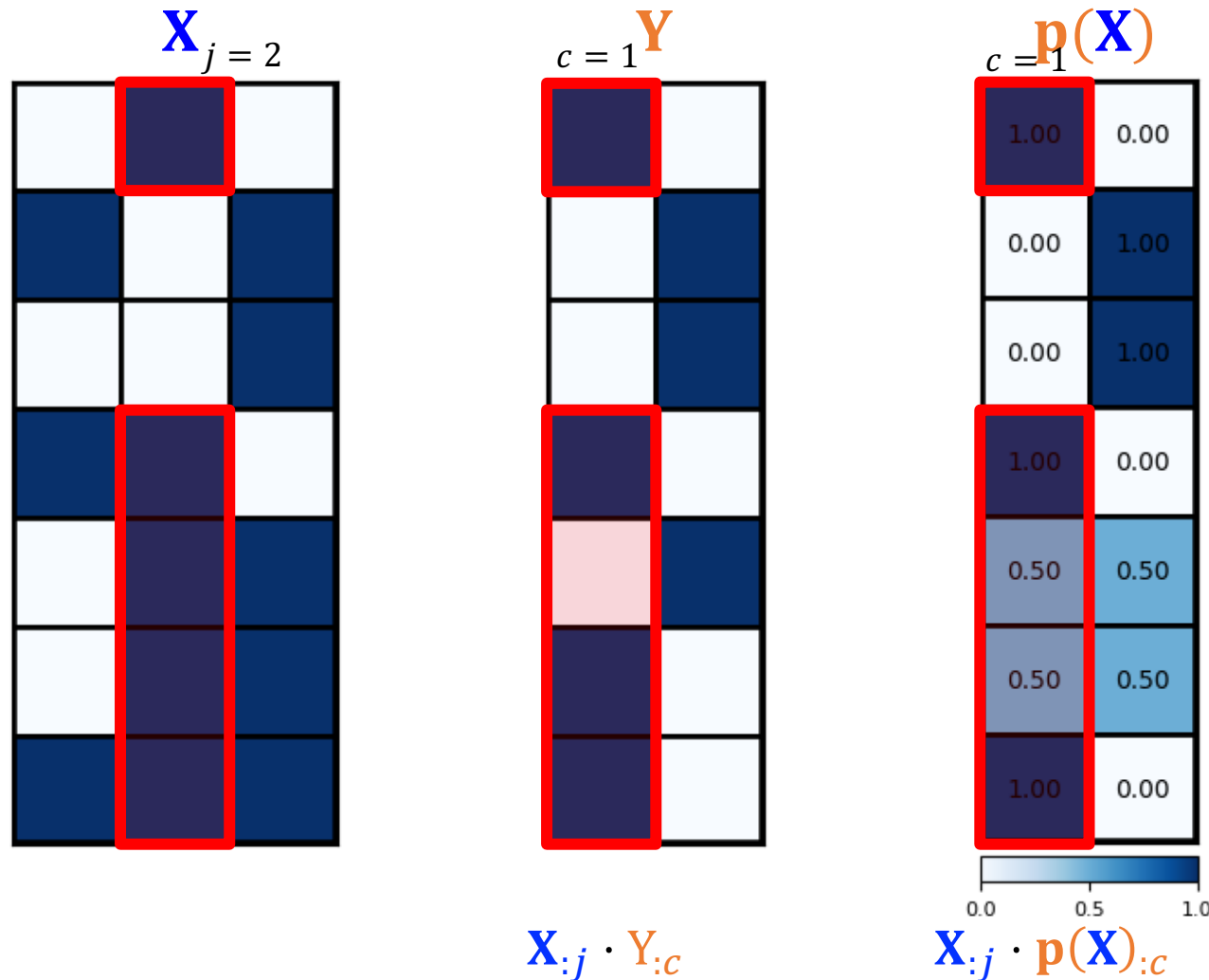


Terminology of "Balance Equations" from "John Mount. The equivalence of logistic regression and maximum entropy models, 2011: <https://github.com/WinVector/Examples/blob/main/dfiles/LogisticRegressionMaxEnt.pdf>

Python file 212, choice 3: <https://github.com/northeastern-datalab/cs7840-activities/tree/main/notebooks>

Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>

# Balance Equations of Logistic Regression



For the data points in the training data with a positive attribute (here coordinate  $X_2=1$ ):

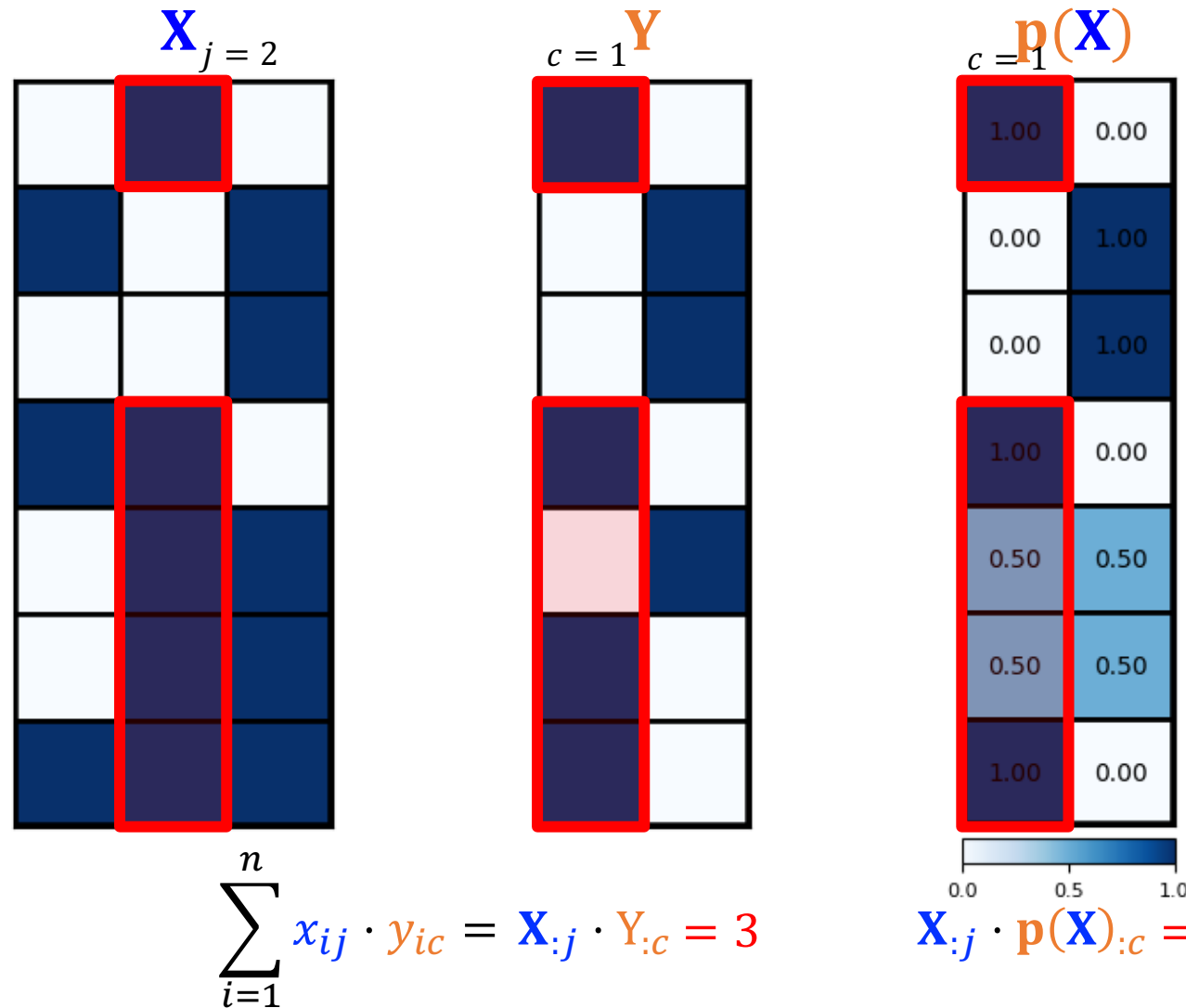
- the count of observations with positive responses equals
- the summed probability mass of the fitted estimate function

Idea: "Summaries of the training data are preserved by the model."

The dot product is usually a measure of how much two vectors are aligned



# Balancing Equations of Logistic Regression

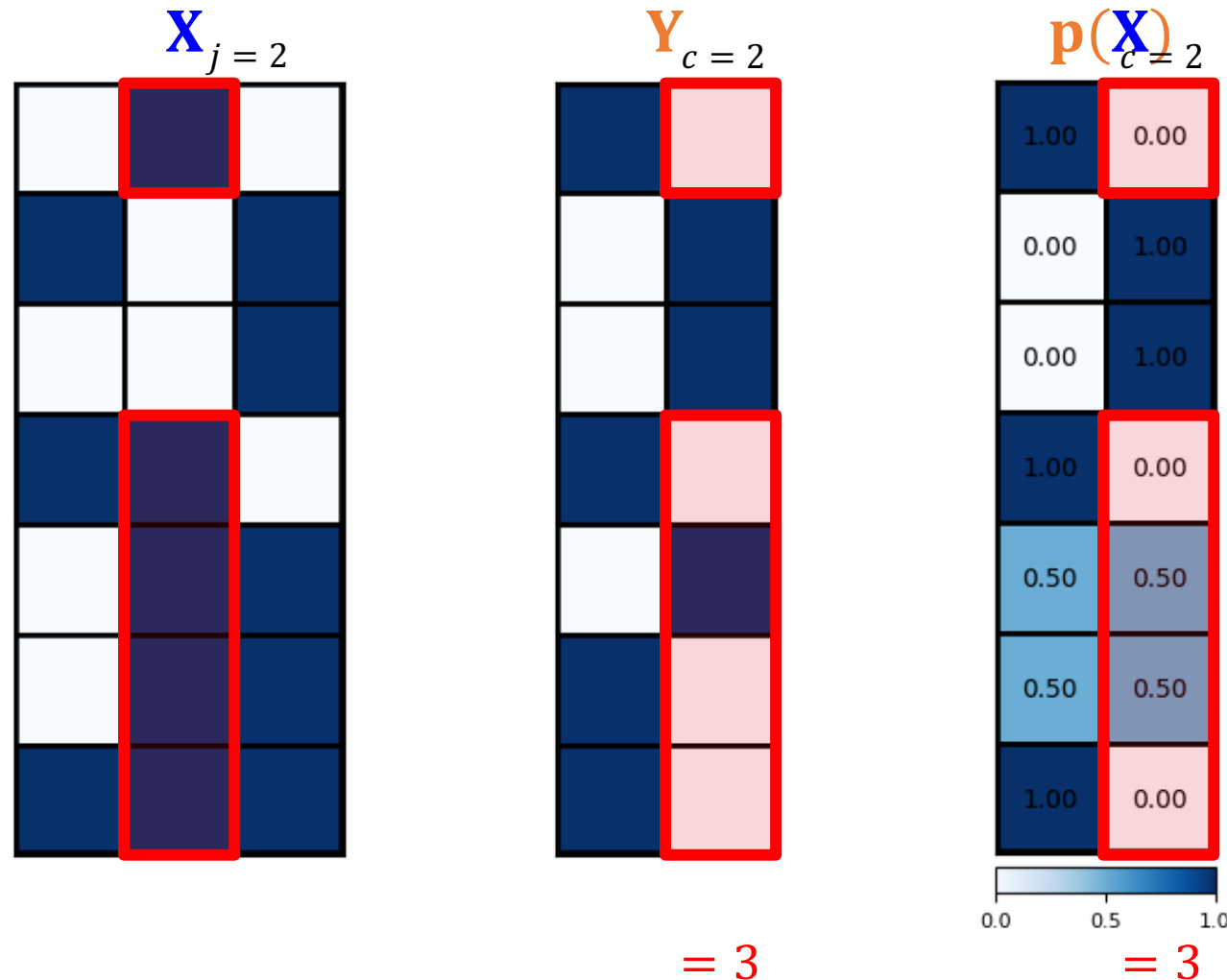


For the data points in the training data with a positive attribute (here coordinate  $X_2=1$ ):

- the count of observations with positive responses equals
- the summed probability mass of the fitted estimate function

Idea: "Summaries of the training data are preserved by the model."

# Balancing Equations of Logistic Regression

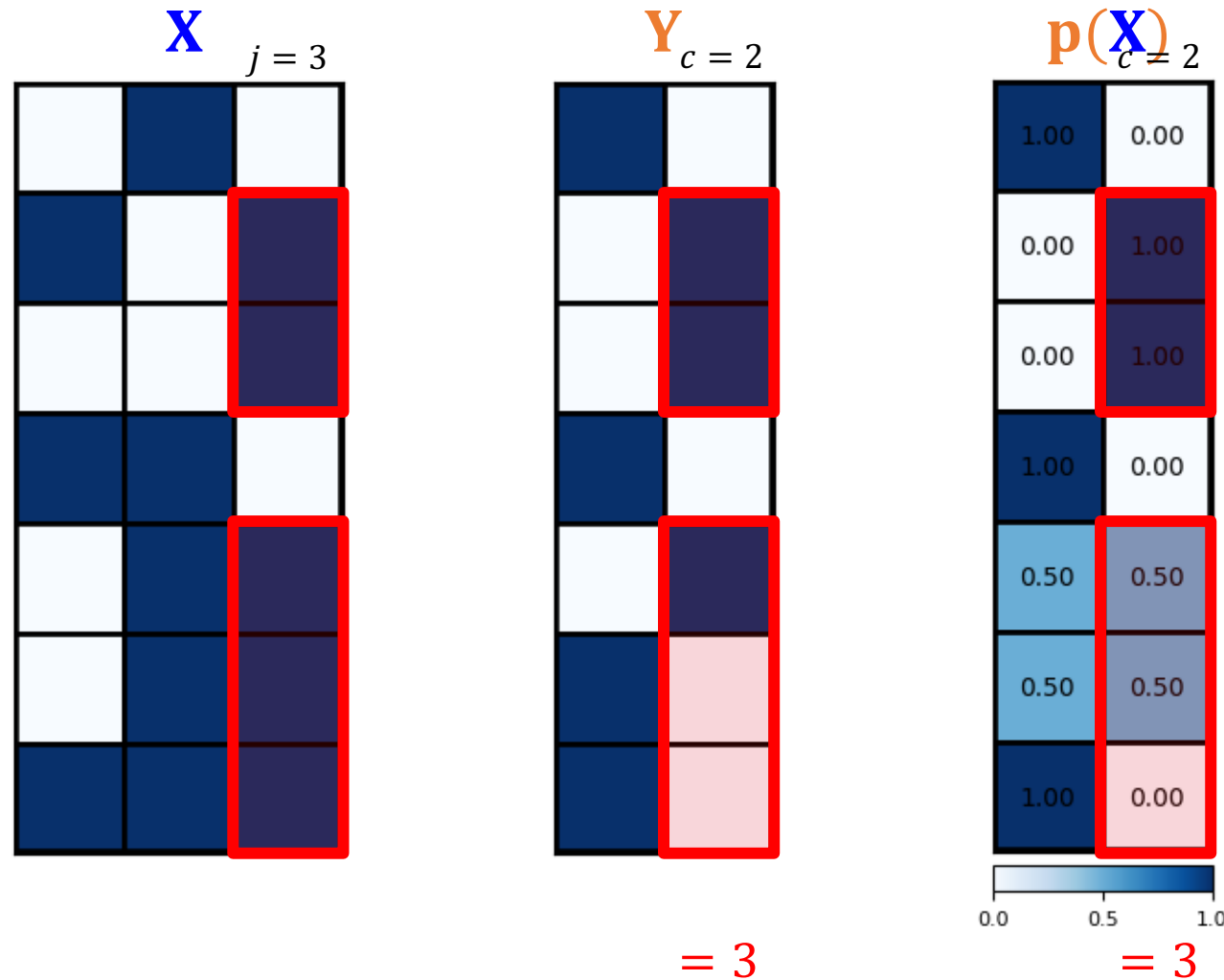


For the data points in the training data with a positive attribute (here coordinate  $X_2=1$ ):

- the count of observations with positive responses equals
- the summed probability mass of the fitted estimate function

Idea: "Summaries of the training data are preserved by the model."

# Balancing Equations of Logistic Regression



For the data points in the training data with a positive attribute (here coordinate  $X_2=1$ ):

- the count of observations with positive responses equals
- the summed probability mass of the fitted estimate function

Idea: "Summaries of the training data are preserved by the model."

# Balancing Equations of Logistic Regression

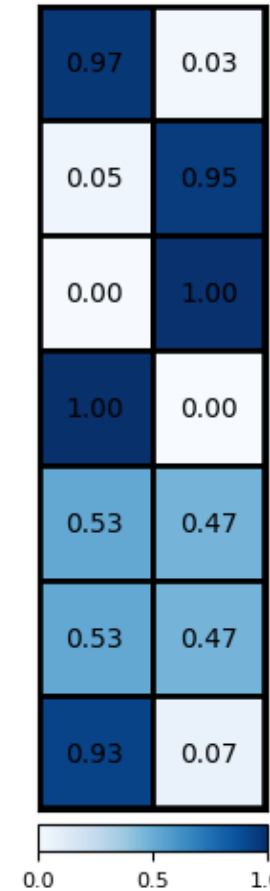
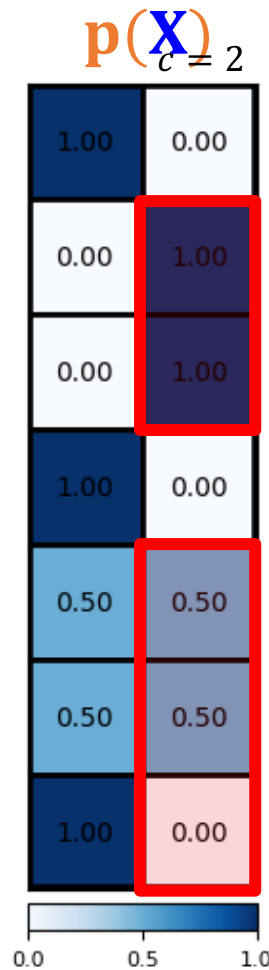
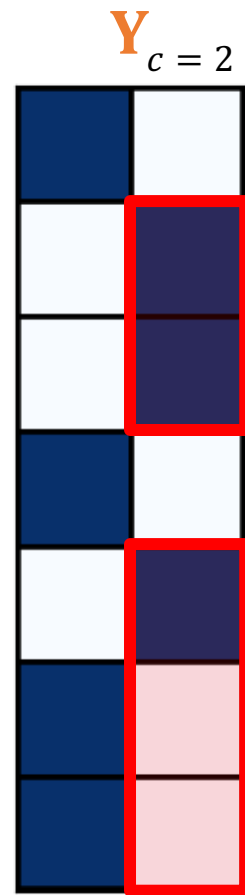
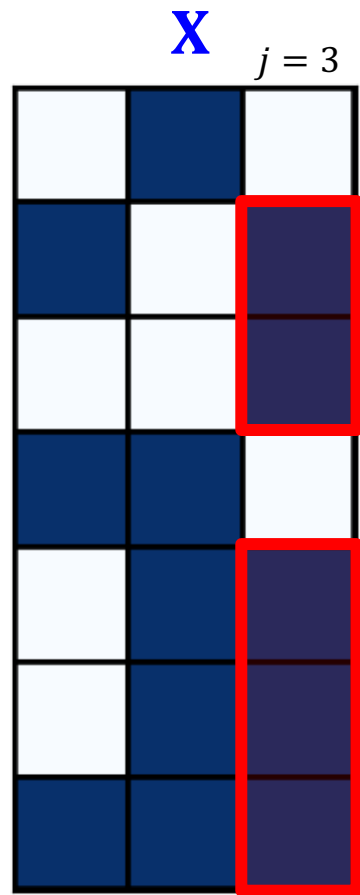
Almost no regularization

$C = 1e5$

$C = 100$

Default regularization

$C = 1$



= 3

= 3

# Balancing Equations apply only to LogReg w/o regularization!

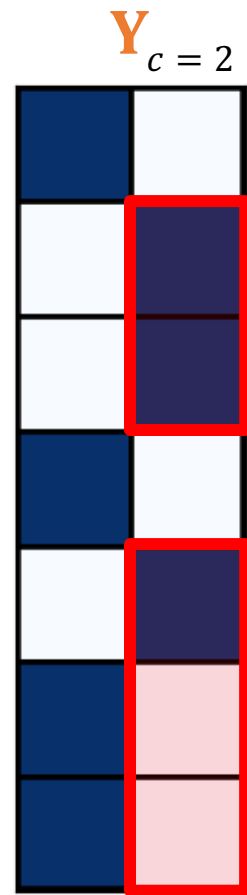
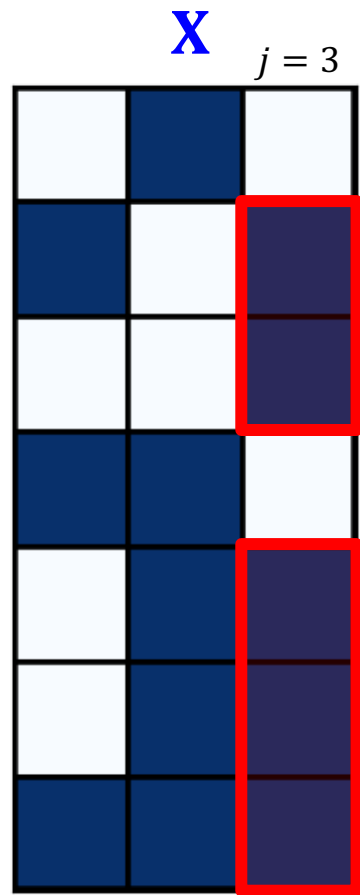
Almost no regularization

$C = 1e5$

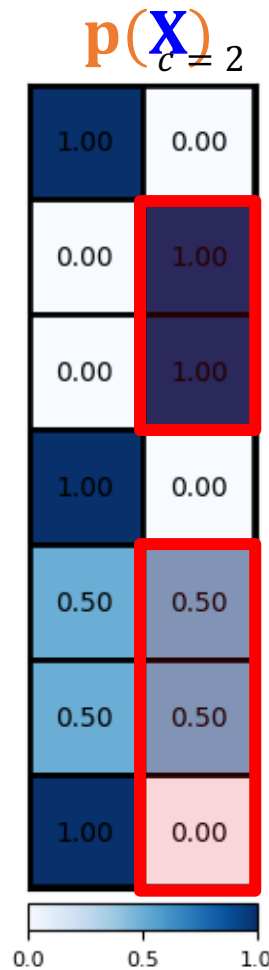
$C = 100$

Default regularization

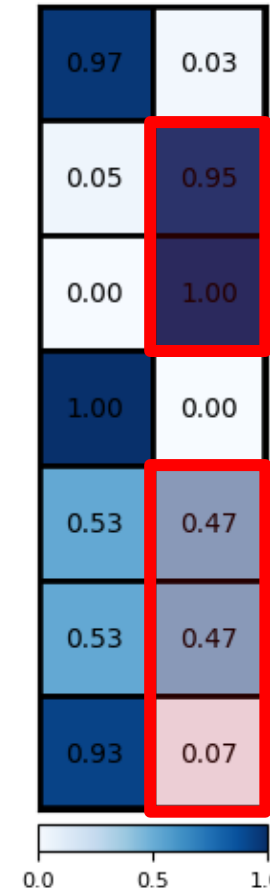
$C = 1$



$= 3$



$= 3$

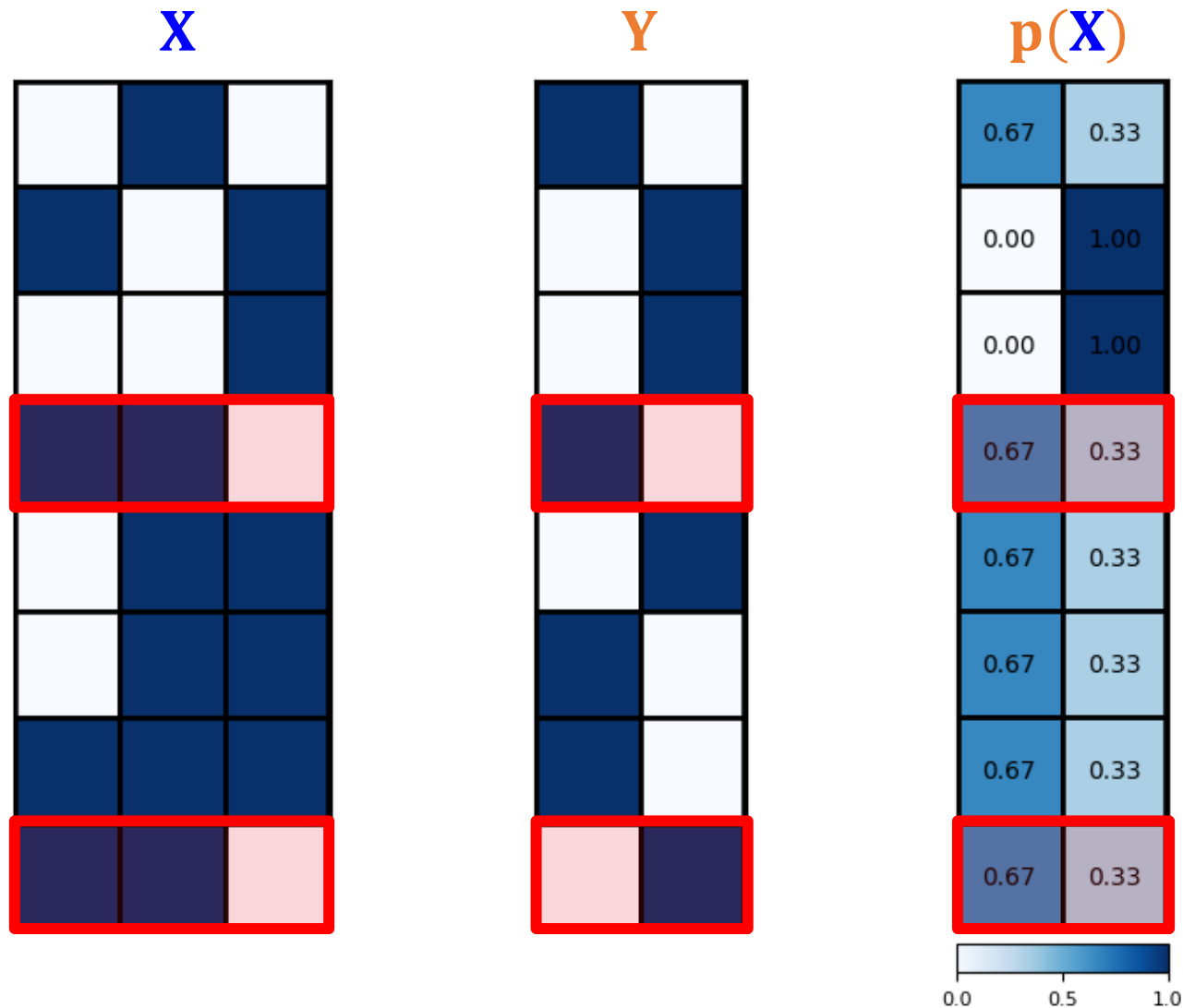


$= 2.96$

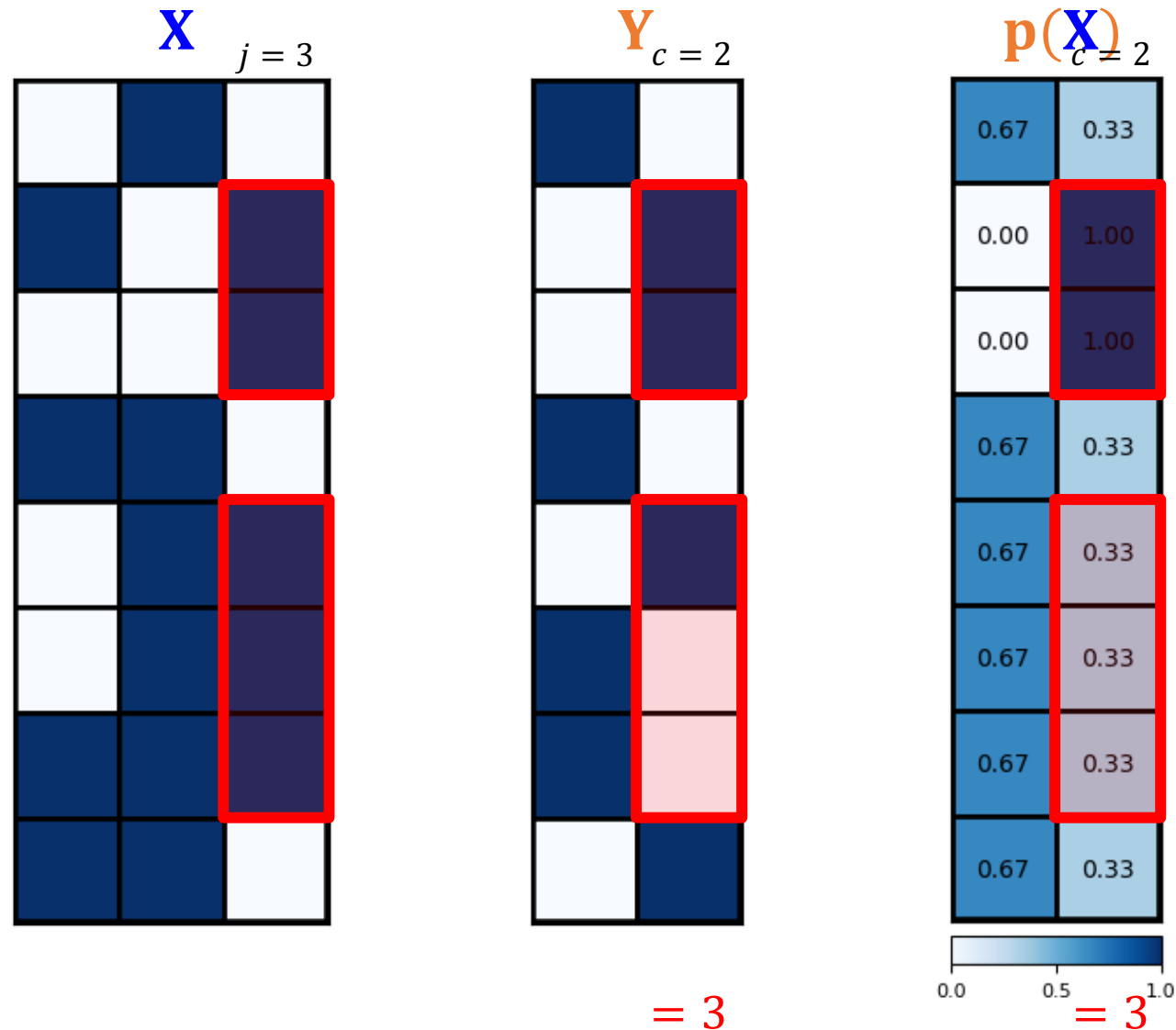


$= 2.45$

# Balancing Equations

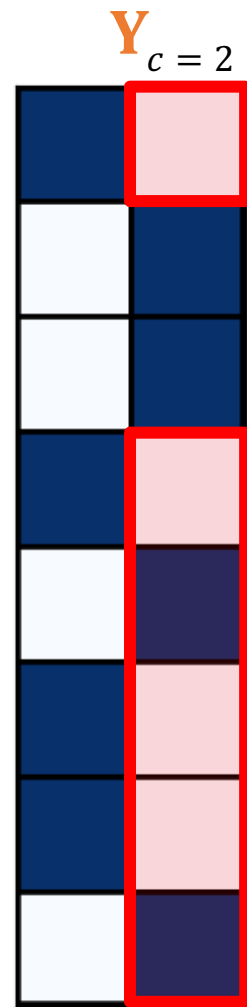
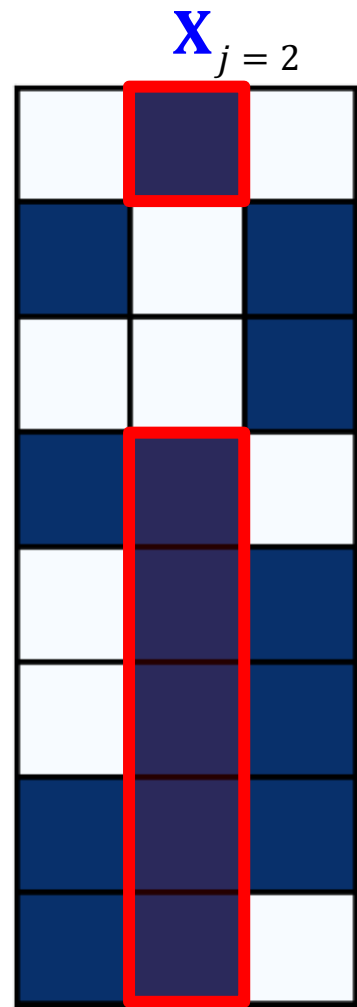


# Balancing Equations

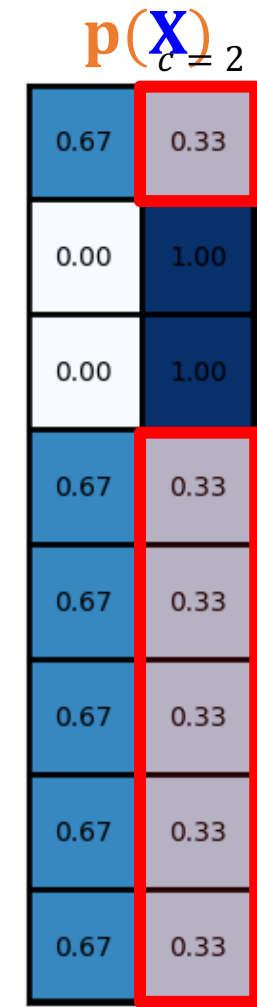


# Balancing Equations as Desiderata

We want to find an estimate function with  $\mathbf{p}(\mathbf{x}_i)_c \approx y_{i,c}$



$$\mathbf{X}_{:j} \cdot \mathbf{Y}_{:c} = 2$$



$$\mathbf{p}(\mathbf{X})_{:c} \cdot \mathbf{X}_{:j} = 2$$

We observe the following "Balancing Equations" of logistic regression:

Fix a particular attribute (coordinate) (here  $j = 2$ ) and a category (here  $c = 2$ )

The sum of category  $c$  (number of times category  $c$  is true) in the training data when attribute  $j$  is true

$$\mathbf{X}_{:j} \cdot \mathbf{Y}_{:c} = \sum_{i=1}^n x_{ij} \cdot y_{ic}$$

is equal the sum of probability mass the model places on that category  $c$  summed across all data when attribute  $j$  is true

$$\mathbf{X}_{:j} \cdot \mathbf{p}(\mathbf{X})_{:c} = \sum_{i=1}^n x_{ij} \cdot \mathbf{p}(\mathbf{x}_i)_c$$

"Summaries of the training data are preserved by the model."



# Deriving multinomial logistic regression with Max Entropy

We want to find a probability distribution  $\mathbf{p}(\mathbf{x})$

$$\mathbf{p}(\mathbf{x})_c \geq 0 \quad \text{for any } \mathbf{x} \in \mathbb{R}^k, c \in [k] \quad \textcircled{1}$$

$$\sum_{c=1}^k \mathbf{p}(\mathbf{x})_c = 1 \quad \text{for any } \mathbf{x} \in \mathbb{R}^k \quad \textcircled{2}$$

*This inequality turns out to be a bit hard: Because of the inequality we would have to use KKT (Karush-Kuhn-Tucker) instead of simpler Lagrange. We ignore this for now and later check if the solution also satisfies it*

For which the balancing equations hold over the training set:

$$\sum_{i=1}^n \mathbf{p}(\mathbf{x}_i)_c \cdot x_{ij} = \sum_{i=1}^n y_{ic} \cdot x_{ij} \quad \text{for any } i \in [n] \text{ and } c \in [k] \quad \textcircled{3}$$

We make no other assumptions. Well, we want to maximize the entropy of that distribution over our training set

$$\max \left[ \sum_{i=1}^n H(\mathbf{p}(\mathbf{x}_i)) \right] = \max \left[ - \sum_{i=1}^n \sum_{c=1}^k \mathbf{p}(\mathbf{x}_i)_c \cdot \lg(\mathbf{p}(\mathbf{x}_i)_c) \right] \quad \textcircled{4}$$

# Deriving multinomial logistic regression with Max Entropy

$$L = - \sum_{i=1}^n \sum_{c=1}^k \mathbf{p}(\mathbf{x}_i)_c \cdot \lg(\mathbf{p}(\mathbf{x}_i)_c) \quad \textcircled{4}$$

$$+ \sum_{i=1}^n \beta_i \left( \left( \sum_{c=1}^k \mathbf{p}(\mathbf{x}_i)_c \right) - 1 \right) \quad \textcircled{2}$$

What next ?

$$+ \sum_{j=1}^m \sum_{c=1}^k \lambda_{j,c} \left( \sum_{i=1}^n \mathbf{p}(\mathbf{x}_i)_c \cdot x_{ij} - \sum_{i=1}^n y_{ic} \cdot x_{ij} \right) \quad \textcircled{3}$$

for any  $i \in [n]$  and  $c \in [k]$

# Deriving multinomial logistic regression with Max Entropy

$$L = - \sum_{i=1}^n \sum_{c=1}^k \mathbf{p}(\mathbf{x}_i)_c \cdot \lg(\mathbf{p}(\mathbf{x}_i)_c) + \sum_{i=1}^n \beta_i \left( \left( \sum_{c=1}^k \mathbf{p}(\mathbf{x}_i)_c \right) - 1 \right) + \sum_{j=1}^m \sum_{c=1}^k \lambda_{j,c} \left( \sum_{i=1}^n \mathbf{p}(\mathbf{x}_i)_c \cdot x_{ij} - \sum_{i=1}^n y_{ic} \cdot x_{ij} \right)$$

for any  $i \in [n]$  and  $c \in [k]$

$$\frac{\partial}{\partial \mathbf{p}(\mathbf{x}_i)_c} L = -\lg(\mathbf{p}(\mathbf{x}_i)_c) - \frac{1}{\ln(2)} + \beta_i + \underbrace{\sum_{j=1}^m \lambda_{j,c} \cdot x_{ij}}_{\boldsymbol{\lambda}_c \cdot \mathbf{x}_i} = 0$$

for any  $i \in [n]$  and  $c \in [k]$

$$(x \cdot \lg(x))' = \cancel{x} \frac{1}{\cancel{x} \ln(2)} + \lg(x)$$

# Let's derive multinomial logistic regression based on Max Entropy

$$-\lg(\mathbf{p}(\mathbf{x}_i)_c) - \frac{1}{\ln(2)} + \beta_i + \lambda_c \cdot \mathbf{x}_i = 0$$

for any  $i \in [n]$  and  $c \in [k]$

$$\lg(\mathbf{p}(\mathbf{x}_i)_c) = \lambda_c \cdot \mathbf{x}_i + \beta_i - \frac{1}{\ln(2)}$$

$$\mathbf{p}(\mathbf{x}_i)_c = e^{\lambda_c \cdot \mathbf{x}_i + \beta_i - \frac{1}{\ln(2)}} = C \cdot e^{\beta_i} \cdot e^{\lambda_c \cdot \mathbf{x}_i}$$

$e^{-\frac{1}{\ln(2)}} \approx 0.236 \dots$  some constant

now we also see  $\mathbf{p}(\dots)_c \geq 0$  because of exponential form

①

Now we need to solve for some of the free variables.

What do we do ?

$\mathbf{p}(\mathbf{x}_i)_c = \mathbf{p}(\mathbf{X}_i)_c$

	$\mathbf{p}(\mathbf{X})$	
	$c$	
$i$	1.00	0.00
	0.00	1.00
	0.00	1.00
	1.00	0.00
	0.50	0.50
	0.50	0.50
	1.00	0.00
	1.00	0.00

# Let's derive multinomial logistic regression based on Max Entropy

$$-\lg(\mathbf{p}(\mathbf{x}_i)_c) - \frac{1}{\ln(2)} + \beta_i + \lambda_c \cdot \mathbf{x}_i = 0$$

for any  $i \in [n]$  and  $c \in [k]$

$$\lg(\mathbf{p}(\mathbf{x}_i)_c) = \lambda_c \cdot \mathbf{x}_i + \beta_i - \frac{1}{\ln(2)}$$

$$\mathbf{p}(\mathbf{x}_i)_c = e^{\lambda_c \cdot \mathbf{x}_i + \beta_i - \frac{1}{\ln(2)}} = C \cdot e^{\beta_i} \cdot e^{\lambda_c \cdot \mathbf{x}_i}$$

now we also see  $\mathbf{p}(\dots)_c \geq 0$   
because of exponential form

$e^{-\frac{1}{\ln(2)}} \approx 0.236 \dots$  some constant

Now we need to solve for some of the free variables.

Luckily we have a normalizing constraint by summing over the  $c \in [k]$

$$\sum_{c=1}^k \mathbf{p}(\mathbf{x})_c = 1 \quad \Rightarrow \quad C \cdot \sum_{c=1}^k e^{\lambda_c \cdot \mathbf{x}_i + \beta_i} = 1 \quad \Rightarrow \quad \mathbf{p}(\mathbf{x}_i)_c = \frac{e^{\lambda_c \cdot \mathbf{x}_i}}{\sum_{i=1}^k e^{\lambda_i \cdot \mathbf{x}_i}}$$

$$C \cdot e^{\beta_i} \sum_{c=1}^k e^{\lambda_c \cdot \mathbf{x}_i} = 1$$

$$e^{\beta_i} = \frac{1}{C \cdot \sum_{c=1}^k e^{\lambda_c \cdot \mathbf{x}_i}}$$

$$\mathbf{p}(\mathbf{x})_c = \frac{\exp(\lambda_c \cdot \mathbf{x})}{\sum_{i=1}^k \exp(\lambda_i \cdot \mathbf{x})}$$

We still have to solve for the free parameters (using numerical methods), but we now know what the function is: the softmax!

# Normal distribution (probit) vs. Logistic distribution (logit)

## 2 The origins of the logistic function

The logistic function was invented in the 19th century for the description of the growth of populations and the course of autocatalytic chemical reactions, or *chain reactions*. In either case we consider the time path of a quantity  $W(t)$  and its growth rate

$$\dot{W}(t) = dW(t)/dt. \quad (4)$$

...

Like Quetelet, Verhulst approached the problem by adding an extra term to (5) to represent the increasing resistance to further growth, as in

$$\dot{W}(t) = \beta W(t) - \phi(W(t)). \quad (6)$$

and then experimenting with various forms of  $\phi$ . The logistic appears when this is a simple quadratic, for in that case we may rewrite (6) as

$$\dot{W}(t) = \beta W(t)(\Omega - W(t)) \quad (7)$$

where  $\Omega$  denotes the upper limit or *saturation level* of  $W$ , its asymptote as  $t \rightarrow \infty$ . Growth is now proportional both to the population already attained  $W(t)$  and to the remaining room for further expansion  $\Omega - W(t)$ . If we express  $W(t)$  as a proportion  $P(t) = W(t)/\Omega$  this gives

$$P(t) = \beta P(t)\{1 - P(t)\}, \quad (8)$$

and the solution of this differential equation is

$$P(t) = \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}, \quad (9)$$

which Verhulst named the *logistic* function. The population  $W(t)$  then follows

$$W(t) = \Omega \frac{\exp(\alpha + \beta t)}{1 + \exp(\alpha + \beta t)}. \quad (10)$$

Verhulst published his suggestions between 1838 and 1847 in three papers.

# Normal distribution (probit) vs. Logistic distribution (logit)

Based on the CDF of the Normal distribution

Bliss published two brief notes in *Science*, introducing the term *probit*; he followed this up with a series of articles setting out the maximum likelihood estimation of the probit curve, in one instance with assistance from R.A. Fisher, Bliss (1935).

...

The acceptance of the probit method was aided by the articles of Bliss, who published regularly in this field until the 1950's, and by Finney and others (Gaddum returned to pharmacology). The full flowering of this school probably coincides with the first edition of Finney's monograph in 1947.

In the practical aspect of ease of computation the logit had a clear advantage over the probit, even with maximum likelihood estimation. To quote Cochran (from his comments on Fisher (1954), p.147) "*.. the speed with which a new technique becomes widely used is considerably influenced by the simplicity or otherwise of the calculations that it requires. Next door to the lecture room in which the probit method is expounded one may still find the laboratory in which the workers compute their LD 50s by the [much less sophisticated] Behrens (Reed-Muench) method ..*". On this count the logit spread much more quickly in workfloor practice than in the academic discourse. Until the advent of the computer and the pocket calculator, some twenty years later, all numerical work was done by hand, that is with pencil and paper, sometimes aided by graphical inspection of 'freehand curves', 'fitted by eye'.

As far as I can see the introduction of the logistic as an alternative to the normal probability function is the work of a single person, namely Joseph Berkson (1899–1982), Reed's co-author of the paper on autocatalytic func-

...

Berkson's suggestion was not well received by the biometric establishment. In the first place, the logit was regarded as somewhat inferior and disreputable because unlike the probit it can not be related to an underlying (normal) distribution of tolerance levels. Aitchison and Brown (1957) dismiss the logit in a single sentence, because it "lacks a well-recognized and manageable frequency distribution of tolerances which the probit curve does possess in a natural way" (p.72). Berkson was aware of this defect and tried to remedy it by adapting the autocatalytic argument, in Berkson (1951), but this did not convince as this argument essentially deals with a process over time. In retrospect it is surprising that so much importance was attached to these somewhat ideological points of interpretation. At the time no one (not even Berkson) seems to have recognized the formidable power of the logistic's analytical properties. In the second place, Berkson's case for the logit was not helped by his simultaneous attacks on the established wisdom of maximum likelihood estimation and his advocacy of minimum chi-squared. The unpleasant atmosphere in which this discussion was conducted can be gauged from the acrimonious exchanges between R.A. Fisher and Berkson in Fisher (1954).

# Probit model (Normal distr.) vs. Logistic regression (logit model)

## Probit model

The probit model uses  $\Phi$ , the CDF (Cumulative Distribution Function) of the Normal Distribution and looks similar to the logistic function

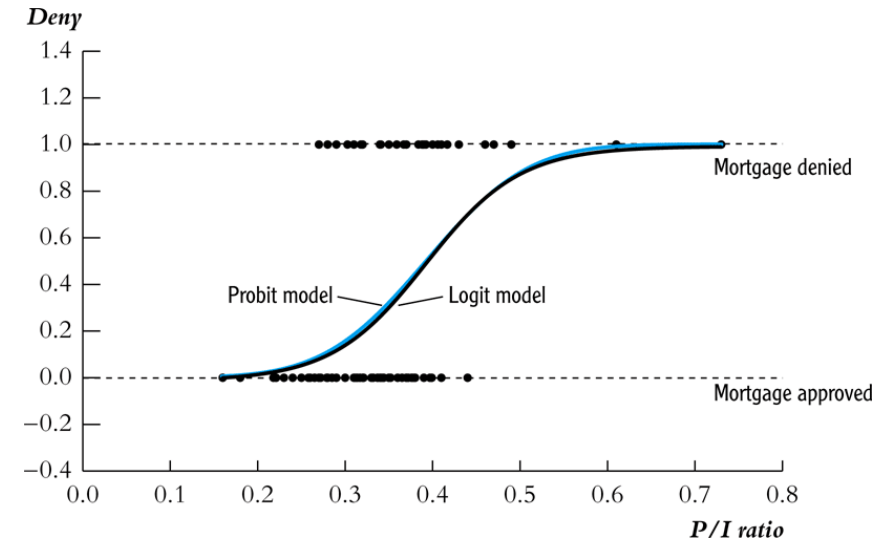
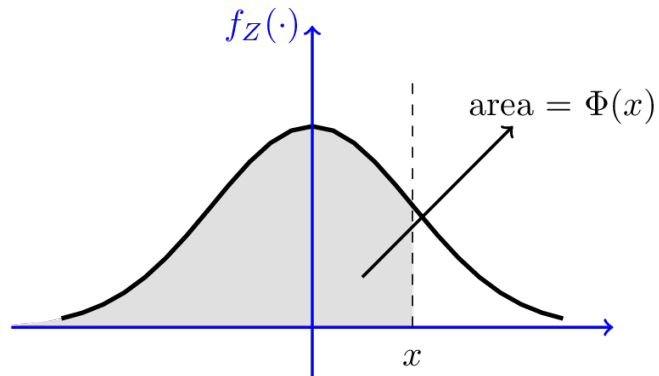


Figure from: [https://bookdown.org/cuborican/RE\\_STAT/linear-probability-probit-logit.html](https://bookdown.org/cuborican/RE_STAT/linear-probability-probit-logit.html) , <https://bookdown.org/machar1991/ITER/11-2-palr.html> , [https://www.probabilitycourse.com/chapter4/4\\_2\\_3\\_normal.php](https://www.probabilitycourse.com/chapter4/4_2_3_normal.php)

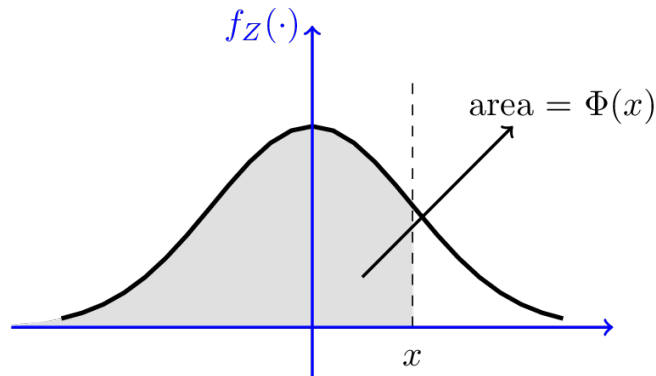
Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>



# Probit model (Normal distr.) vs. Logistic regression (logit model)

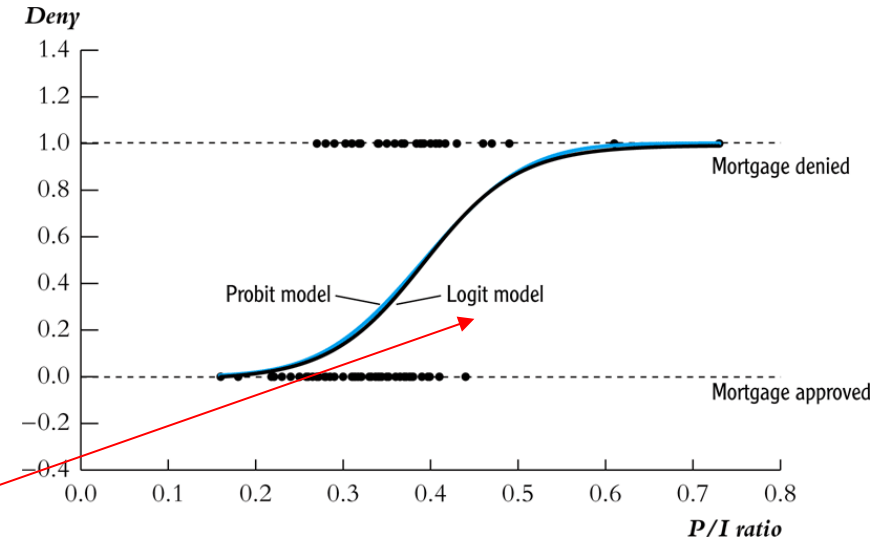
## Probit model

The probit model uses  $\Phi$ , the CDF (Cumulative Distribution Function) of the Normal Distribution and looks similar to the logistic function

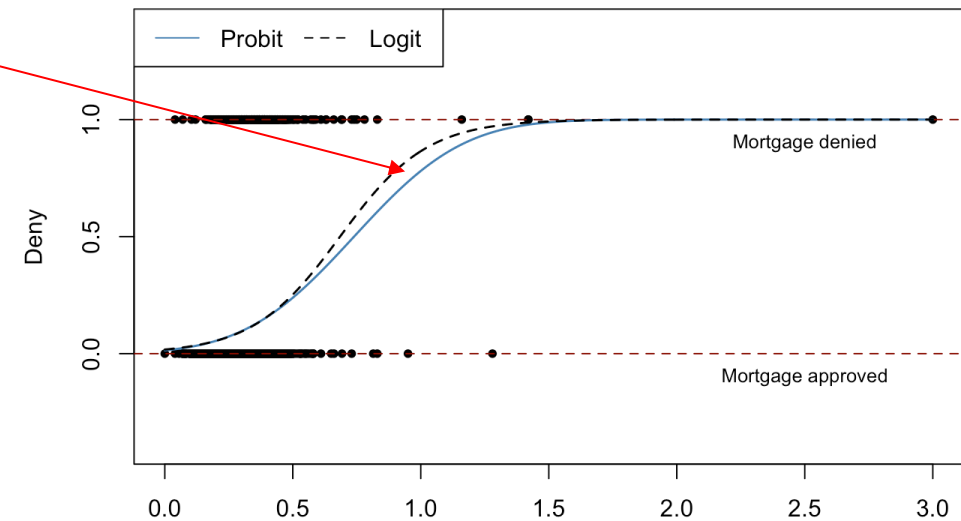


Logistic function has slightly fatter tails

Also, the logistic model is a mathematically simpler model (CDF has a simple closed form)



Probit and Logit Models Model of the Probability of Denial, Given P/I Ratio



# Normal distribution (probit) vs. Logistic distribution (logit)

## Normal distribution

PDF

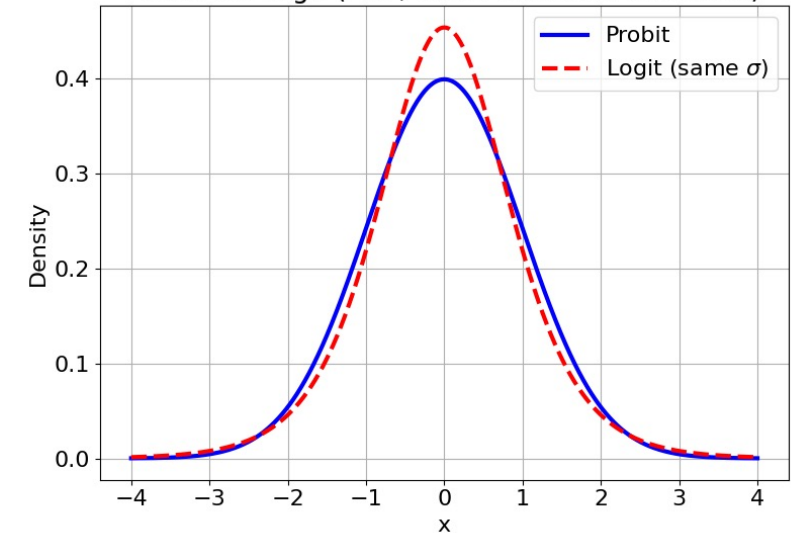
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## Logistic distribution

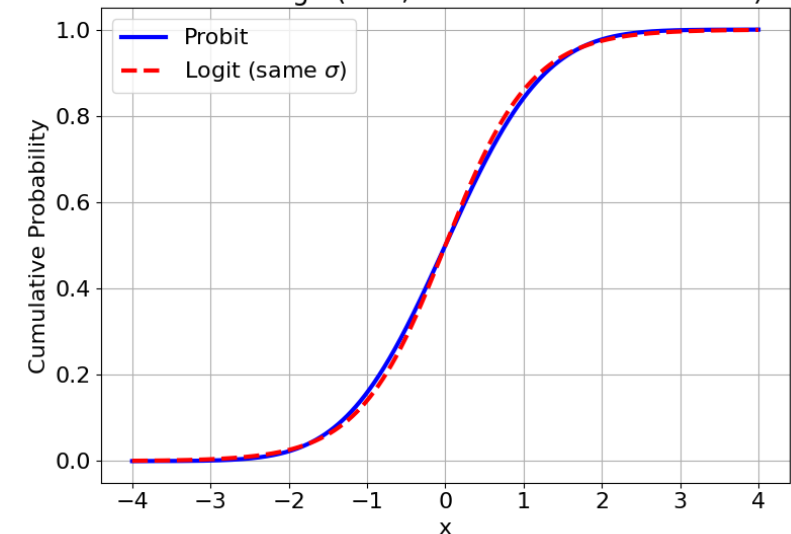
CDF

$$F(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

Probit vs. Logit (PDF, same standard deviation)



Probit vs. Logit (CDF, same standard deviation)



# Normal distribution (probit) vs. Logistic distribution (logit)

## Normal distribution

PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard normal

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

CDF

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right]$$

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt$$

## Logistic distribution

$$f(x) = \frac{e^{-\frac{x-\mu}{s}}}{s \left( 1 + e^{-\frac{x-\mu}{s}} \right)^2}$$

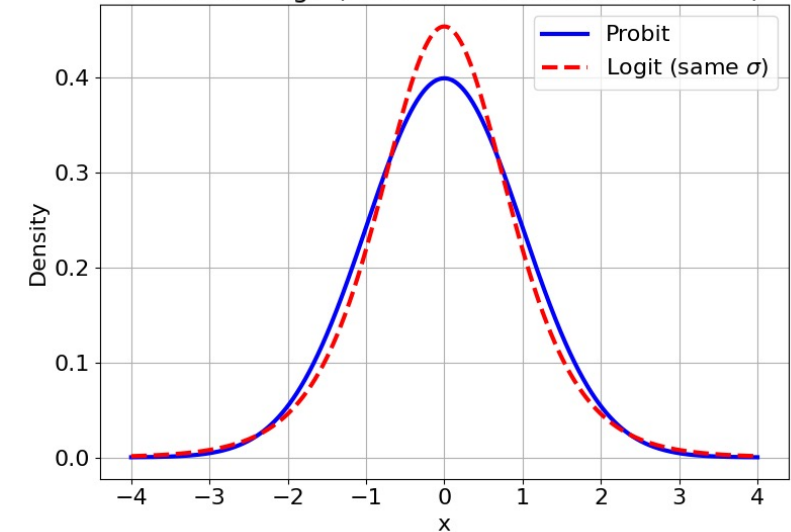
Variance

$$\sigma^2 = \frac{s^2 \pi^2}{3}$$

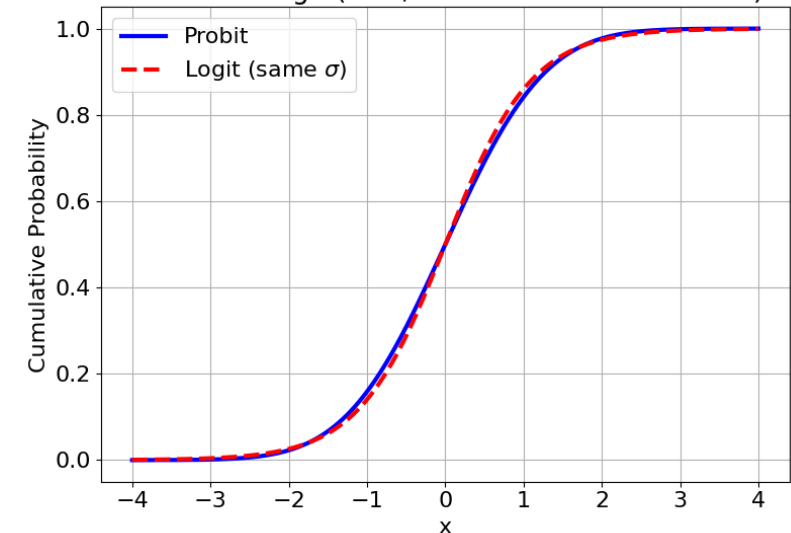
scale parameter  $s = \frac{\sqrt{3}}{\pi}$  for  $\sigma^2 = 1$

$$F(x) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}$$

Probit vs. Logit (PDF, same standard deviation)



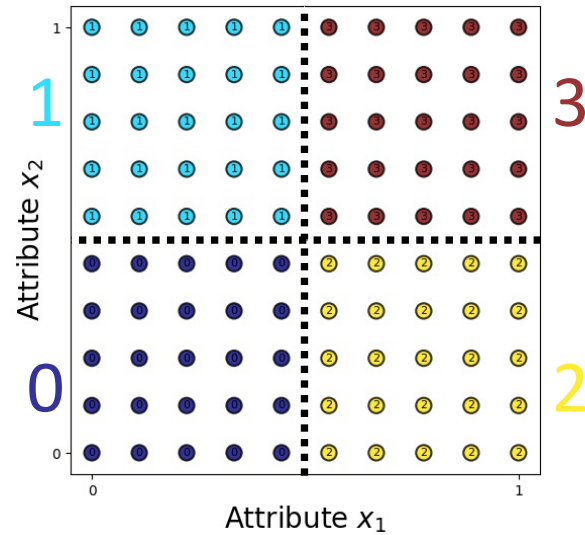
Probit vs. Logit (CDF, same standard deviation)



# Playing with multinomial logistic regression

# Multinomial logistic regression

$m = 2$  attributes  
 $k = 4$  classes

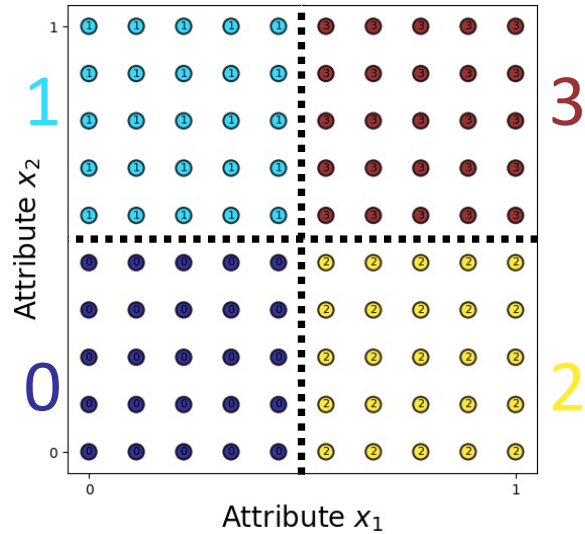


$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

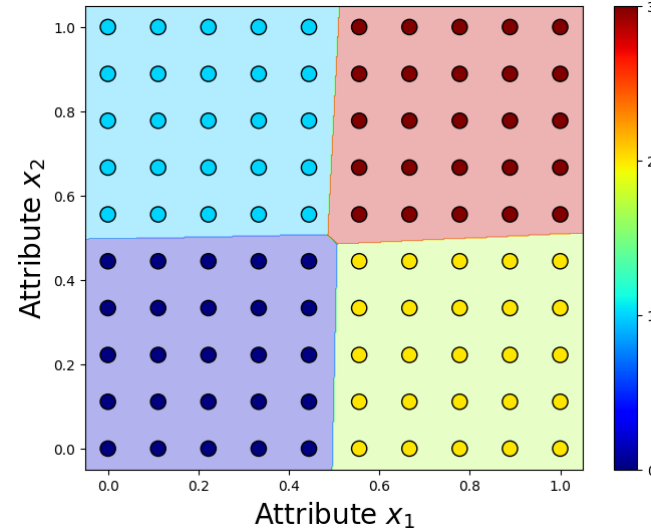
Can multinomial logistic regression fit these data ?

# Multinomial logistic regression

$m = 2$  attributes  
 $k = 4$  classes



$C = 10^{10}$

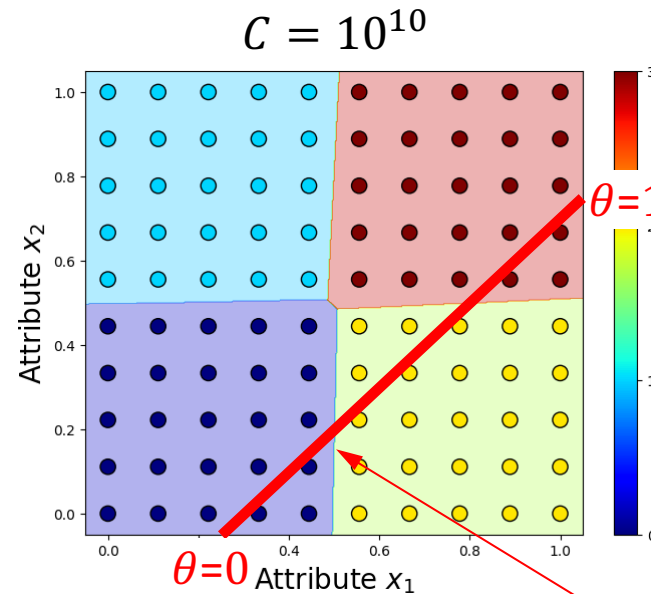
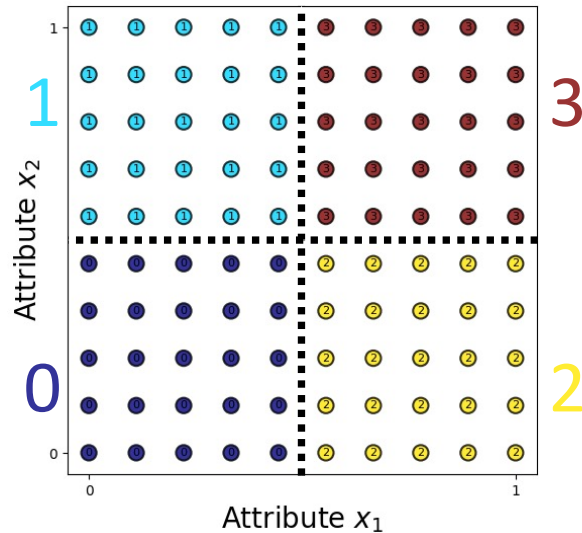


$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

# Multinomial logistic regression

$m = 2$  attributes  
 $k = 4$  classes



c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

Let's make a "cut" and show the class distribution along that cut:

$$\theta \in [0,1]$$

$$x_1(\theta) = 0.25 + 0.75 \cdot \theta$$

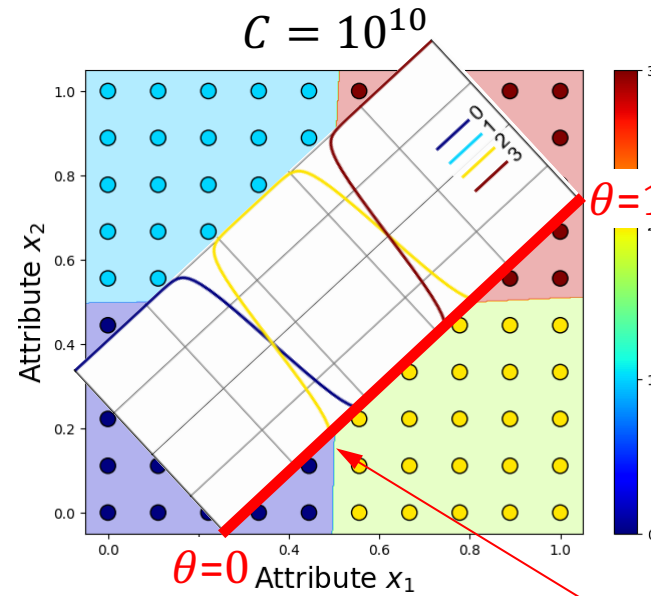
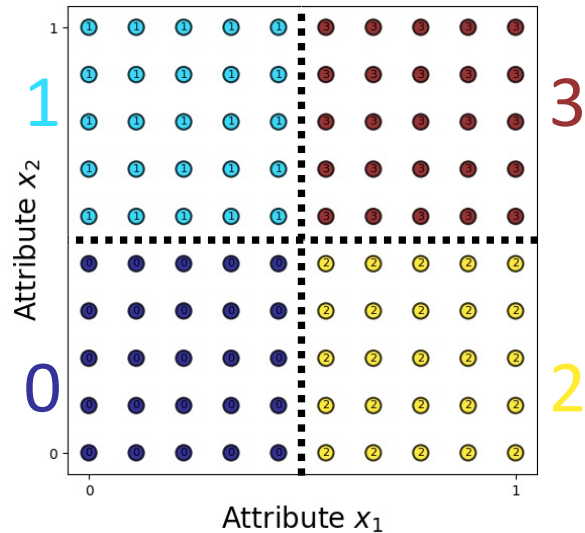
$$x_2(\theta) = 0.75 \cdot \theta$$



$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

# Multinomial logistic regression

$m = 2$  attributes  
 $k = 4$  classes



$C = 1$

What happens if we use the default regularizer ?

$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

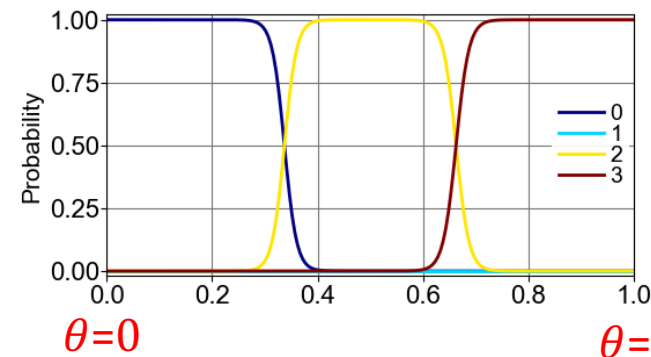
c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

Let's make a "cut" and show the class distribution along that cut:

$$\theta \in [0,1]$$

$$x_1(\theta) = 0.25 + 0.75 \cdot \theta$$

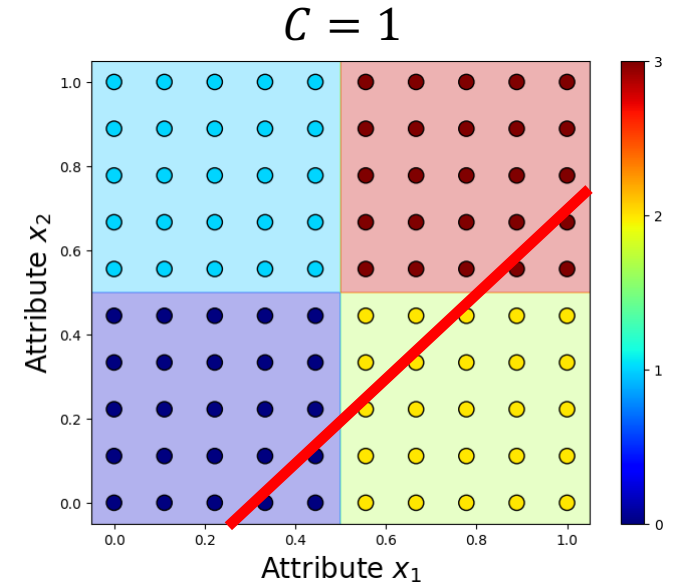
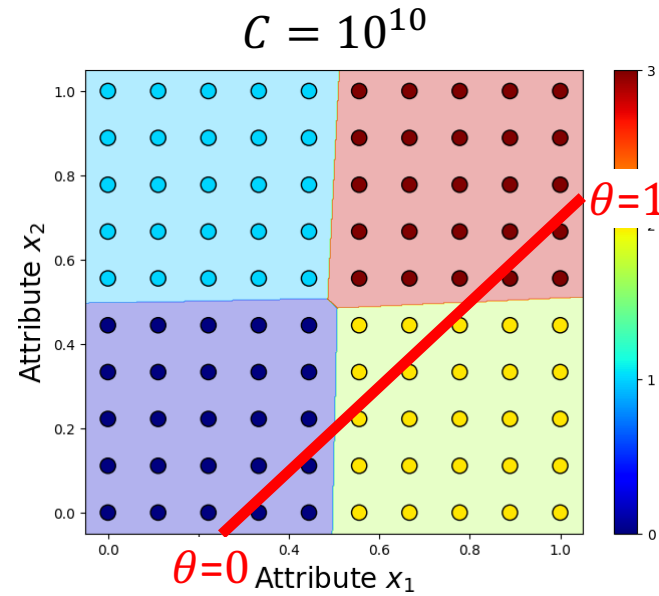
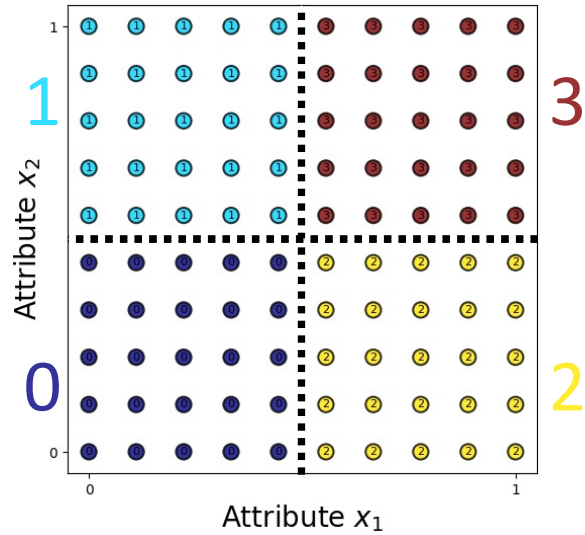
$$x_2(\theta) = 0.75 \cdot \theta$$





# Multinomial logistic regression

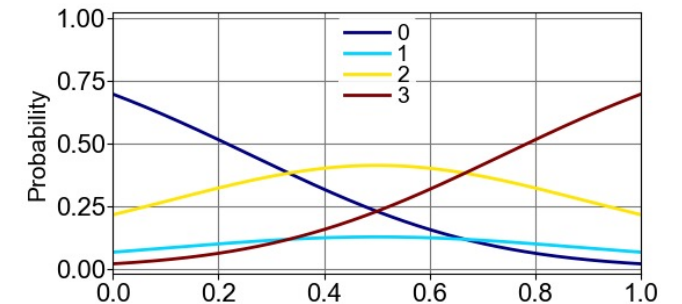
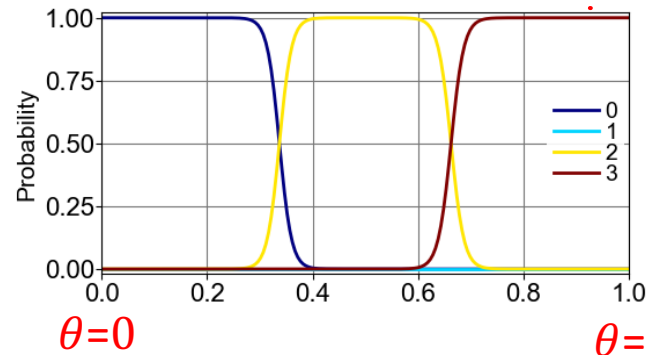
$m = 2$  attributes  
 $k = 4$  classes



$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

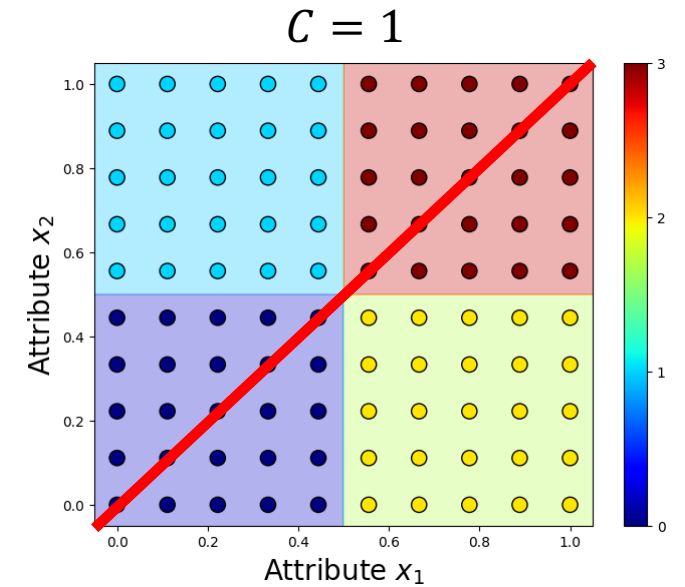
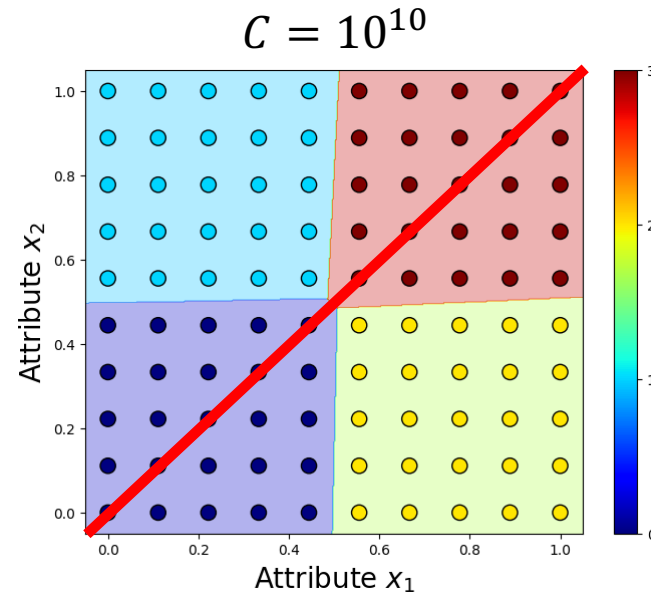
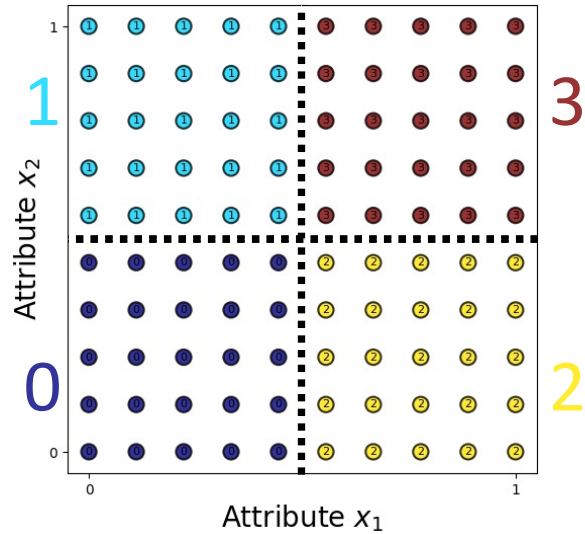
c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	2.34	-2.34	-2.34
1	0.00	-2.34	2.34
2	0.00	2.34	-2.34
3	2.34	2.34	2.34



# Multinomial logistic regression

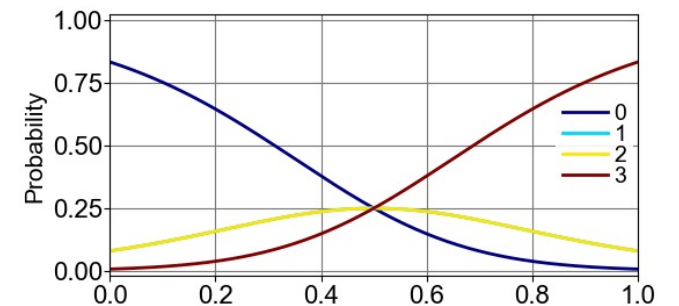
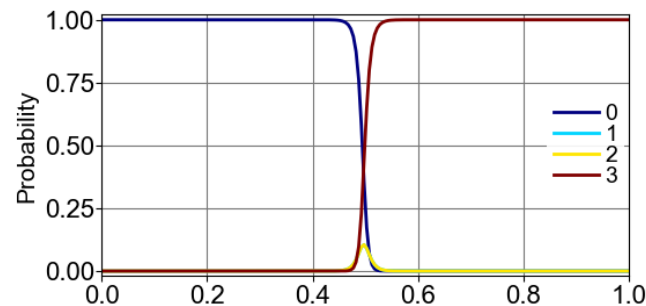
$m = 2$  attributes  
 $k = 4$  classes



$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

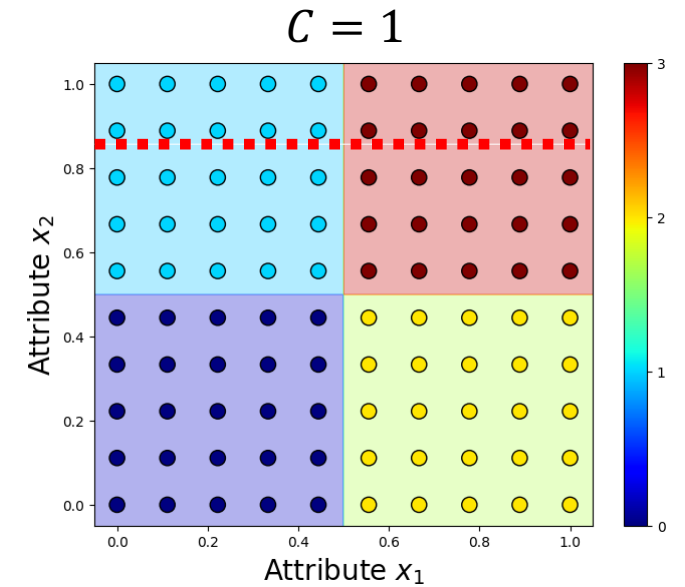
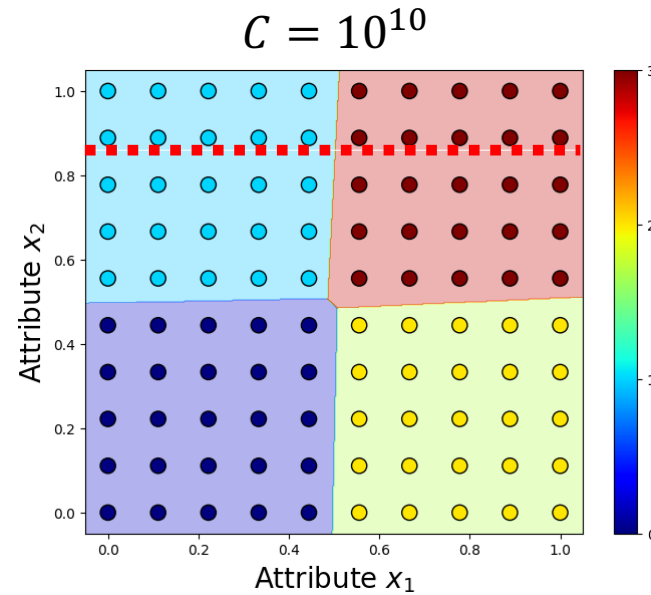
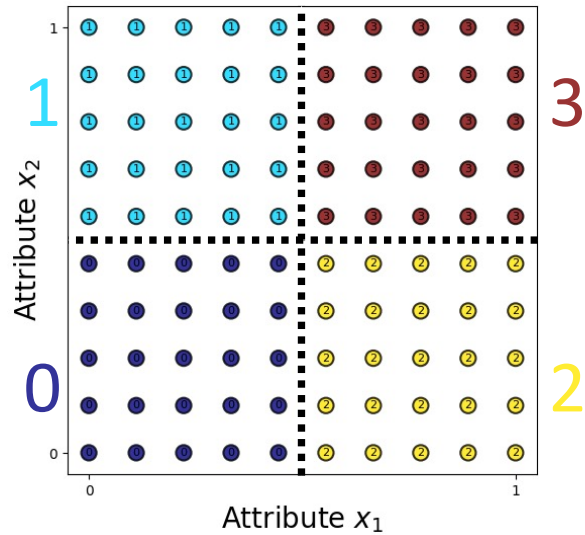
c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	2.34	-2.34	-2.34
1	0.00	-2.34	2.34
2	0.00	2.34	-2.34
3	2.34	2.34	2.34



# Multinomial logistic regression

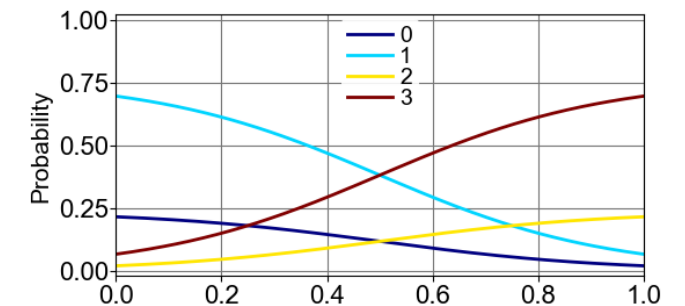
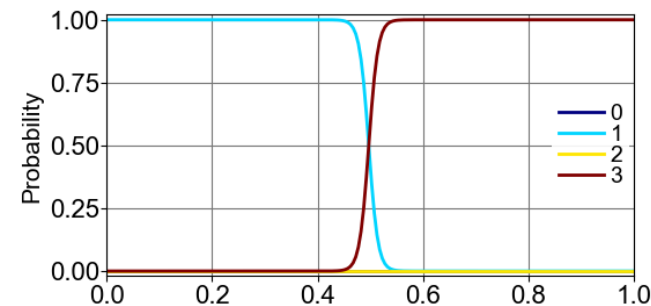
$m = 2$  attributes  
 $k = 4$  classes



$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	2.34	-2.34	-2.34
1	0.00	-2.34	2.34
2	0.00	2.34	-2.34
3	2.34	2.34	2.34

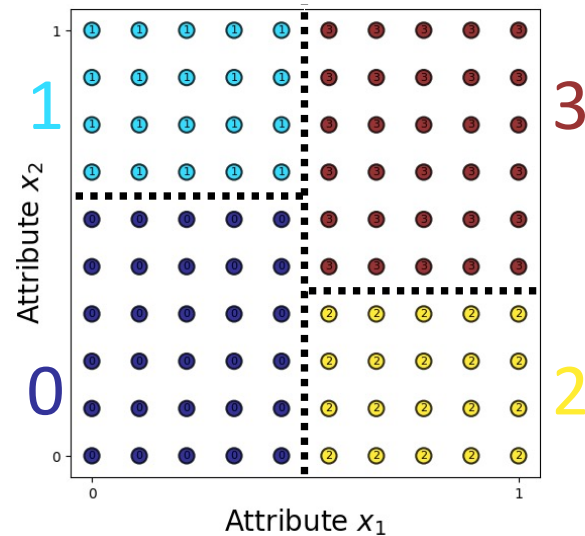


# Multinomial logistic regression

$C = 10^{10}$

$C = 1$

$m = 2$  attributes  
 $k = 4$  classes



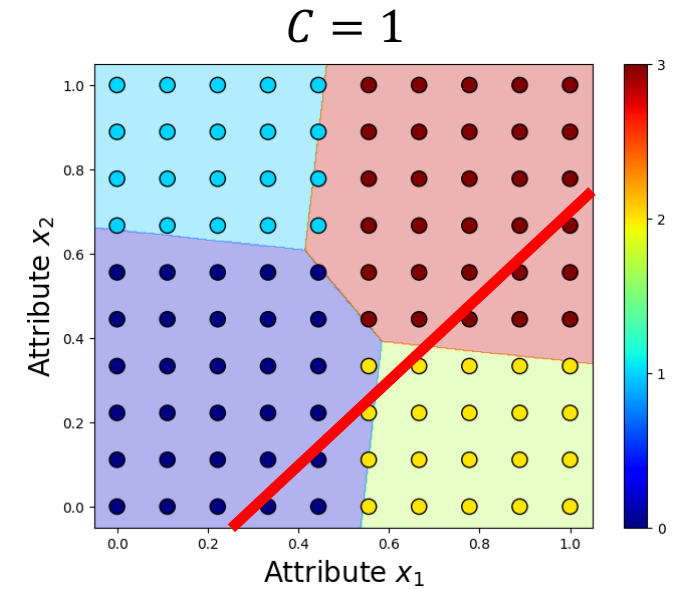
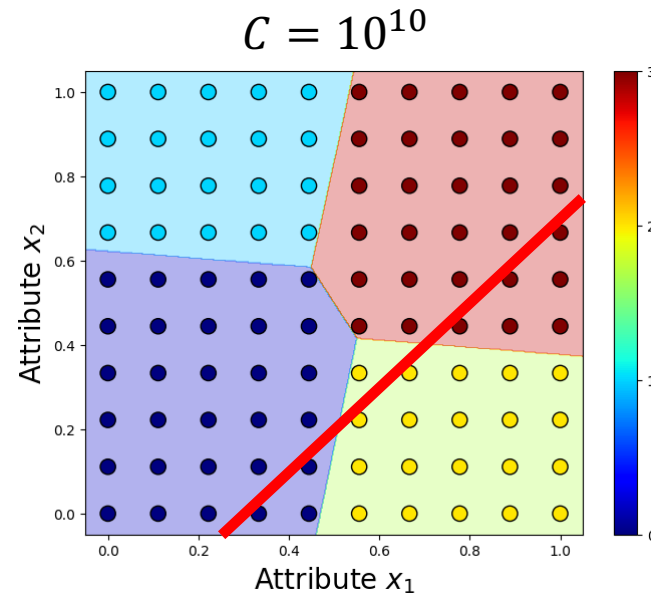
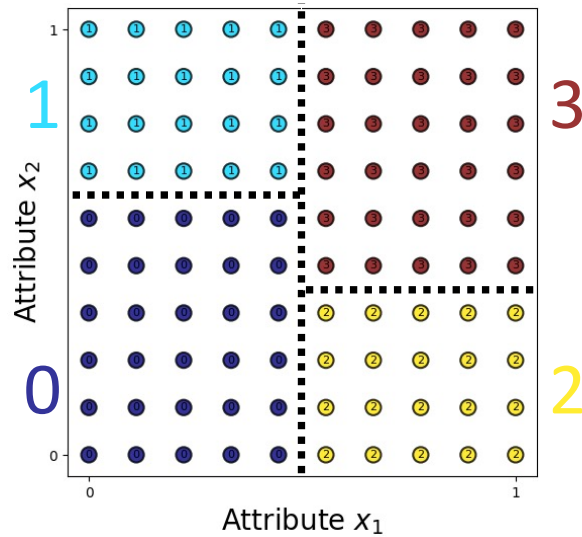
$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

How well can multinomial logistic regression fit these data that we had seen in our lecture on decision trees



# Multinomial logistic regression

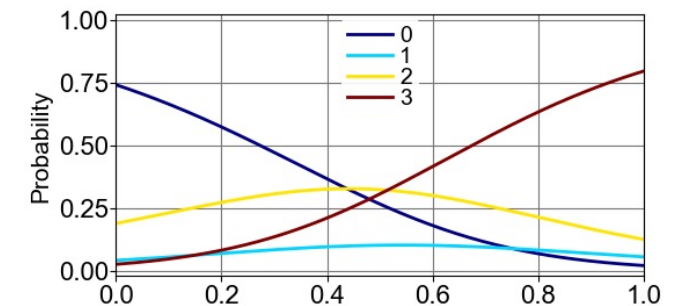
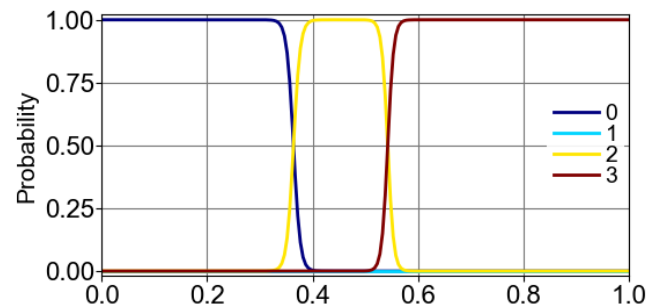
$m = 2$  attributes  
 $k = 4$  classes



$$\mathbb{P}[y = c] = \frac{\exp(\lambda_{c,0} + \lambda_{c,1} \cdot x_1 + \lambda_{c,2} \cdot x_2)}{\sum_{i=1}^k \exp(\lambda_{i,:} \cdot \mathbf{x})}$$

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	57.03	-56.75	-56.75
1	-1.39	-59.20	60.67
2	-1.39	60.67	-59.20
3	-54.24	55.28	55.28

c	intercept	coefficients	
	$\lambda_{:,0}$	$\lambda_{:,1}$	$\lambda_{:,2}$
0	2.34	-2.34	-2.34
1	0.00	-2.34	2.34
2	0.00	2.34	-2.34
3	2.34	2.34	2.34



## Part 3: Applications

### L19: Logistic regression (2/2)

[Connections (multinomial) logistic regression, Luce's choice axiom, Bradley-Terry(-Luce) model]

Wolfgang Gatterbauer, Javed Aslam

cs7840 Foundations and Applications of Information Theory (fa24)

<https://northeastern-datalab.github.io/cs7840/fa24/>

11/13/2024

# Pre-class conversations

- Questions: exp/e, python file (please submit yours too w/ scribes)
  
  
  
  
  
  
  
  
  
  
  
  
  
- Today:
  - Choice theorem
  - Occam, Max Entropy
  - MDL

Luce's choice axiom &  
Bradley-Terry model  
(another justification for  
the softMax function)



# "Irrelevant Alternatives" in Rational Choice Theory

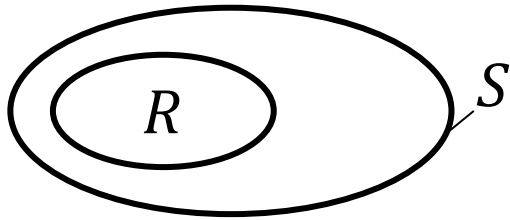
A motivating story:

Philosopher Sidney Morgenbesser, ordering dessert, is told by a waitress that he can choose between **A**pple pie or **B**lueberry. He orders **A**pple. Soon the waitress comes back and explains **C**herry pie is also an option. Morgenbesser replies "In that case, I'll have **B**lueberry."

'Spoiler effect':

Suppose **C**harlie (an irrelevant alternative) enters a race between **A**lice and **B**ob, where **A**lice (leader) is set to defeat **B**ob. **Independence of Irrelevant Alternatives (IIA)** says that in a rational system of voting, if **C**harlie joins the race but loses, **A**lice should not suddenly lose the election to **B**ob. In this context, violating IIA is commonly referred to as the 'spoiler effect': support for **C**harlie 'spoils' the election for **A**lice.

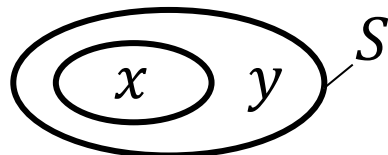
# Luce's Choice Axiom



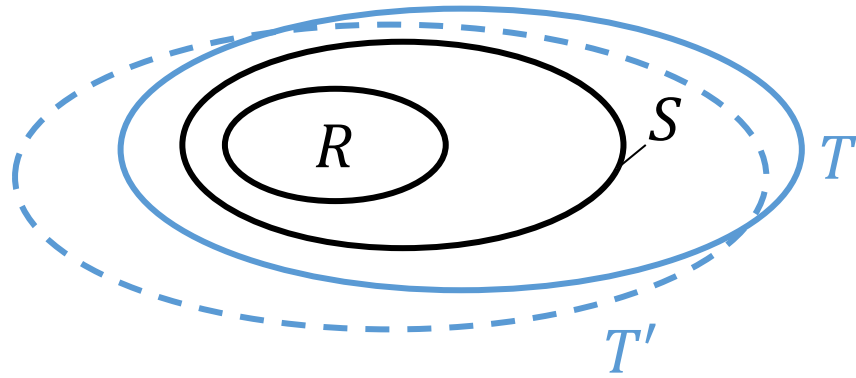
$S$  contains a set of alternatives,  
with  $R \subset S$  being a strict subset

Let  $\mathbb{P}_S(R)$  be the probability (of an  
individual or a group) of choosing  
from subset  $R$ , if given  $S$  as  
alternatives.

E.g.,  $\mathbb{P}_{\{x,y\}}(x)$  is probability  
of choosing  $x$  over  $y$ .



# Luce's Choice Axiom



$S$  contains a set of alternatives, with  $R \subset S$  being a strict subset

Let  $\mathbb{P}_S(R)$  be the probability (of an individual or a group) of choosing from subset  $R$ , if given  $S$  as alternatives.

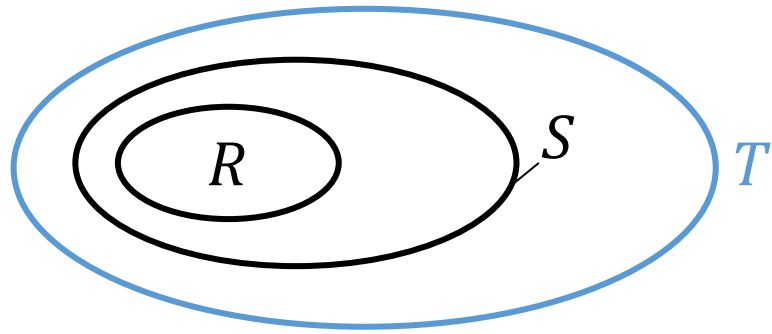
E.g.,  $\mathbb{P}_{\{x,y\}}(x)$  is probability of choosing  $x$  over  $y$ .

LUCE'S CHOICE AXIOM (original formulation, but my own interpretation): For all  $T \supset S$ ,  $\mathbb{P}_T(R) = \mathbb{P}_T(S) \cdot \mathbb{P}_S(R)$

Also known as **INDEPENDENCE FROM IRRELEVANT ALTERNATIVES (IIA)**: the choice ratio between any two items is unaffected by the presence of other items in the set.

$$\frac{\mathbb{P}_T(x)}{\mathbb{P}_T(y)} = \frac{\cancel{\mathbb{P}_T(S)} \cdot \mathbb{P}_S(x)}{\cancel{\mathbb{P}_T(S)} \cdot \mathbb{P}_S(y)} = \frac{\mathbb{P}_S(x)}{\mathbb{P}_S(y)}$$

# Luce's Choice Axiom



$S$  contains a set of alternatives, with  $R \subset S$  being a strict subset

Let  $\mathbb{P}_S(R)$  be the probability (of an individual or a group) of choosing from subset  $R$ , if given  $S$  as alternatives.

E.g.,  $\mathbb{P}_{\{x,y\}}(x)$  is probability of choosing  $x$  over  $y$ .

LUCE'S CHOICE AXIOM (original formulation, but my own interpretation): For all  $T \supset S$ ,  $\mathbb{P}_T(R) = \mathbb{P}_T(S) \cdot \mathbb{P}_S(R)$

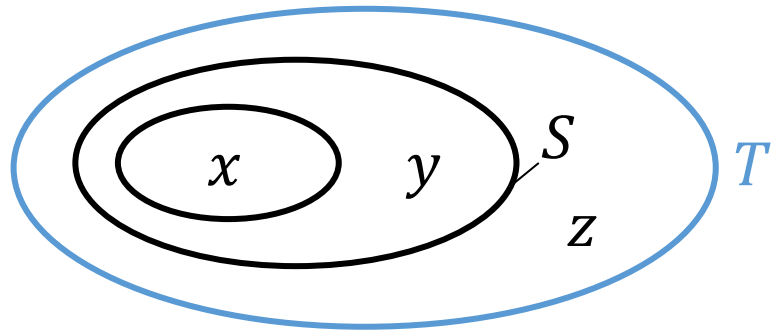
Also known as **INDEPENDENCE FROM IRRELEVANT ALTERNATIVES (IIA)**: the choice ratio between any two items is unaffected by the presence of other items in the set.

$$\frac{\mathbb{P}_T(x)}{\mathbb{P}_T(y)} = \frac{\cancel{\mathbb{P}_T(S)} \cdot \mathbb{P}_S(x)}{\cancel{\mathbb{P}_T(S)} \cdot \mathbb{P}_S(y)} = \frac{\mathbb{P}_S(x)}{\mathbb{P}_S(y)}$$

**Ratio scale representation:** There is a positive **ratio scale**  $u$  (i.e., a numerical value representing the **utility** or strength or preference for that alternative, unique up to multiplication by a positive constant) on  $T$  such that

$$\mathbb{P}_S(x) = \frac{u(x)}{\sum_{y \in S} u(y)} = \frac{u(x)}{u(S)}$$

# Luce's Choice Axiom (a philosophical side-topic)



Assume Alice is indifferent to using a bike or a bus for going to school. There are 3 choices in total:

$x$ : a bicycle

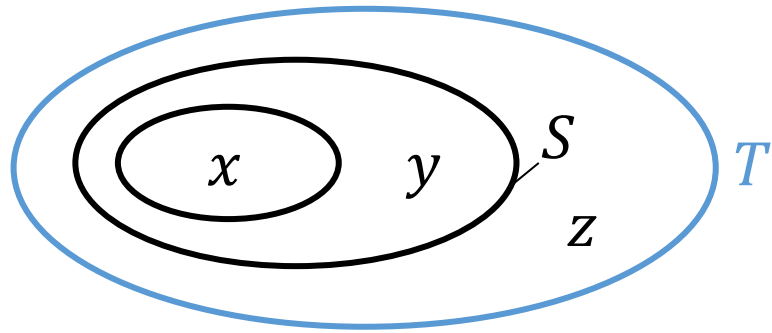
$y$ : a red bus

$z$ : a blue bus

If presented with options  $\{x, y\}$  as only choices, **what are her relative preferences:**

?

# Luce's Choice Axiom (a philosophical side-topic)



Assume Alice is indifferent to using a bike or a bus for going to school. There are 3 choices in total:

$x$ : a bicycle

$y$ : a red bus

$z$ : a blue bus

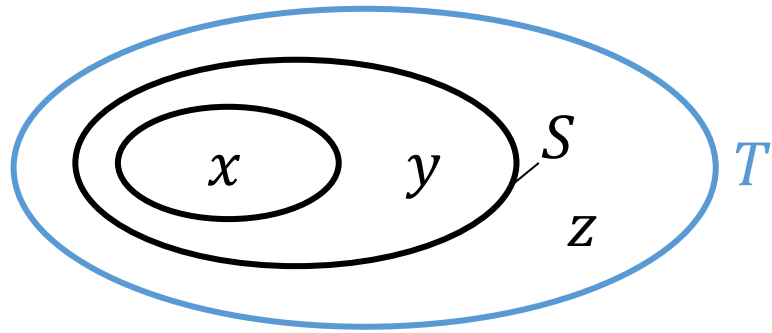
If presented with options  $\{x, y\}$  as only choices, her relative preferences are:

$$\frac{1}{2}, \frac{1}{2}$$

When adding  $z$  (her options are now  $\{x, y, z\}$ ), what are her preferences



# Luce's Choice Axiom (a philosophical side-topic)



Assume Alice is indifferent to using a bike or a bus for going to school. There are 3 choices in total:

$x$ : a bicycle

$y$ : a red bus

$z$ : a blue bus

If presented with options  $\{x, y\}$  as only choices, her relative preferences are:

$$\frac{1}{2}, \frac{1}{2}$$

When adding  $z$  (her options are now  $\{x, y, z\}$ ), what are her preferences

$$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$$

$$\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$$



# Bradley-Terry-Luce model (BTL)

Model for the outcome of **pairwise comparison** between items (teams, or objects)

Item  $i$  has a positive latent score (**utility**, skill)  $u_i$ .

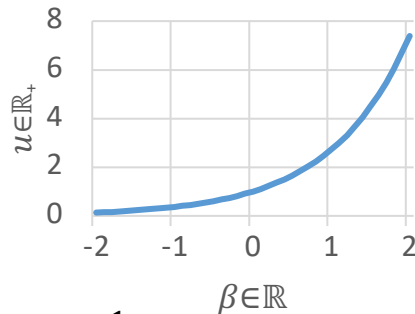
Goal: based on a set of pairwise comparisons, derive a full ranking of all items (participants).

Key assumption: The probability that  $i$  is chosen over (wins against)  $j$  is:  $\mathbb{P}(i > j) = \frac{u_i}{u_i + u_j}$

Most commonly used with an exponential scoring function

$$u_i = \exp(\beta_i)$$

$$\mathbb{P}(i > j) = \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)} = \frac{1}{1 + \exp(-(\beta_i - \beta_j))}$$





# Bradley-Terry-Luce model (BTL)

Model for the outcome of **pairwise comparison** between items (teams, or objects)

Item  $i$  has a positive latent score (**utility**, skill)  $u_i$ .

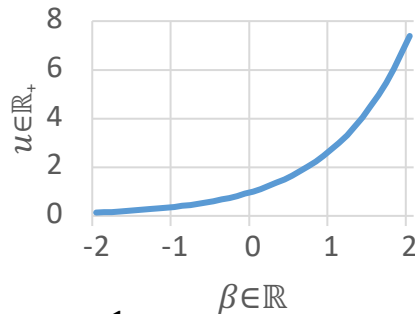
Goal: based on a set of pairwise comparisons, derive a full ranking of all items (participants).

Key assumption: The probability that  $i$  is chosen over (wins against)  $j$  is:  $\mathbb{P}(i > j) = \frac{u_i}{u_i + u_j}$

Most commonly used with an exponential scoring function

$$u_i = \exp(\beta_i)$$

$$\mathbb{P}(i > j) = \frac{\exp(\beta_i)}{\exp(\beta_i) + \exp(\beta_j)} = \frac{1}{1 + \exp(-(\beta_i - \beta_j))}$$



Model is overparameterized (scalar multiples of  $u$ 's / units don't matter):

$$\frac{u_i}{u_i + u_j} = \frac{\alpha \cdot u_i}{\alpha \cdot u_i + \alpha \cdot u_j}$$

Same as additive constants on the logits:

$$\begin{aligned} \text{logit}(\mathbb{P}(i > j)) &= \log\left(\frac{\mathbb{P}(i > j)}{1 - \mathbb{P}(i > j)}\right) = \log\left(\frac{\mathbb{P}(i > j)}{\mathbb{P}(j > i)}\right) \\ &= \beta_i - \beta_j = (\beta_i + c) - (\beta_j + c) \end{aligned}$$

To make it uniquely identifiable, add another constraint, e.g.  $\beta_j = 0$ . Then:

$$\mathbb{P}(i > j) = \frac{1}{1 + \exp(-\beta_i)}$$

BT model named after [Bradley,Terry'52] ("Rank analysis of incomplete block designs", Biometrika 1952, <https://doi.org/10.2307/2334029>), though [Zermelo'29] ("Die Berechnung der Turnierergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung", Mathematische Zeitschrift. 1929, <https://doi.org/10.1007/BF01180541>) had studied it already earlier. Then [Luce'59] ("Individual choice behavior: A theoretical analysis", 1959. 2005 ed: <https://doi.org/10.1037/14396-000>), provided an axiomatic basis for it.

Gatterbauer, Aslam. Foundations and Applications of Information Theory: <https://northeastern-datalab.github.io/cs7840/>