Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ingeniería de Transporte y Logística

Sistemas urbanos inteligentes

Neural Networks and the Multinomial Logit for Brand Choice Modelling: a Hybrid Approach

YVES BENTZ Y DWIGHT MERUNKA (2000)

Guillermo Otárola, Diego Guamán

1. INTRODUCCION

Las elecciones discretas las realizamos todos los días, incluso sin tener conciencia de ello.

1. INTRODUCCION

En marketing existen muchas variables que impactan en el comportamiento del consumidor (elección de un producto)

Características del producto

- Precio
- Promociones
- Publicidad
- Calidad

Características del consumidor

- Edad
- Sexo
- Ingreso

ESTRUCTURA DEL PAPER

“Redes neuronales y el logit multinomial para el modelado de elección de marca: un enfoque híbrido”

Journal of Forescasting
J. Forescast .19,177±200 (2000)

YVES BENTZ and DWIGHT MERUNKA
1London Business School, UK
2IAE Aix-en-Provence, France

- Comparar el modelo multinomial logit (MNL) con una red neuronal feed forward con función de salida softmax.
- Utiliza el modelado de elección de marca donde por muchos años se ha realizado a través de MNL.
- Propone un enfoque hibrido y complementario entre modelos para mejorar el rendimiento predictivo.

CASO 1

“Elección de Marcas de Chocolate”

3 Marcas de chocolate y 2 variables de decisión, V1,V2 (usando datos sintéticos)

CASO 2

“Pronosticar las compras individuales de café instantáneo”

Datos basado en análisis de compra en tiendas de Australia en diferentes alternativas de café.

Key words Redes neuronales; modelo logit multinomial; modelos de elección; elección de marca

1. INTRODUCCION

Los modelos de elección discreta, como el Logit Multinomial (MNL) se basan en la teoría de utilidad aleatoria, desarrollada en la década de 1970 (MacFadden, 1974).

- El MNL es atractivo ya que es estocástico y utiliza variables de decisión (parte determinística).
- El MNL utiliza una función de utilidad que en su forma más simple es lineal.
- El MNL puede considerar funciones no lineales, pero la especificación de estas resulta complejo y hay un alto riesgo de introducir un sesgo de especificación

Dadas las limitaciones del MNL puede resultar mejor utilizar modelos más generales (regresión no lineal, no paramétricos, semi paramétricos). Las redes neuronales pertenecen a esta categoría.

- Son parecidas a el MNL
- Se describen gráficamente
- Son de uso fácil

Las redes neuronales se han aplicado ampliamente en la investigación de mercados y pueden verse como generalizaciones no lineales de regresiones lineales o logísticas

2. MODELOS LOGIT

Permiten modelar el comportamiento de las personas al escoger una alternativa (Y). (Y es una variable categórica).

Tiene un gran poder explicativo y es de fácil aplicación

- El MNL es atractivo ya que es estocástico y utiliza variables de decisión (parte determinística).

- El MNL utiliza una función de utilidad que en su forma más simple es lineal.

- El MNL puede considerar funciones no lineales, pero la especificación de estas resulta complejo y hay un alto riesgo de introducir un sesgo de especificación

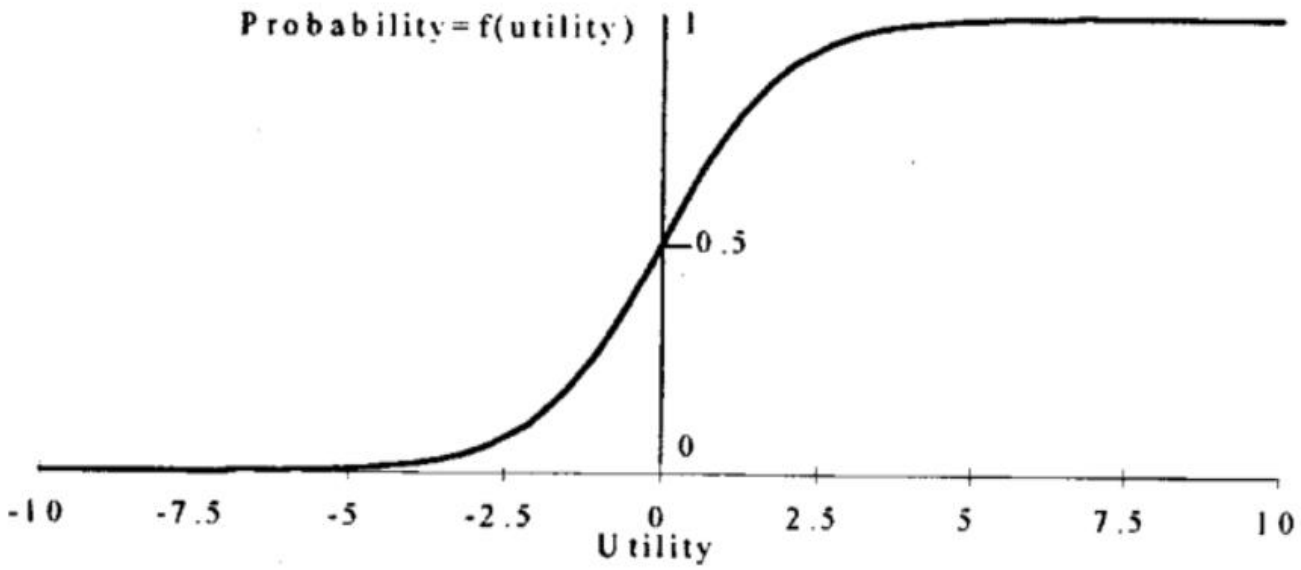
$$P_i(Y = 1) = f(\text{Utility})$$

$$\text{Utility} = a + \sum_{j \in T} b_j \times x_{ij} + \varepsilon$$

2. MODELOS BINOMIAL

$$P_i(Y = 1) = f(\text{Utility}) \qquad \text{Utility} = a + \sum_{j \in T} b_j \times x_{ij} + \varepsilon$$

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



Fuente: Yves Bentz, Dwight Merunka (2000)

No linealidad de la probabilidad de elección respecto a cada variable de entrada

Las variables interactúan de forma lineal en la función de utilidad

2. LOGIT MULTINOMIAL

La probabilidad de elección de una alternativa es función de todas las alternativas

$$P_i(Y = j) = \frac{e^{v_{ij}}}{\sum_{j=1}^{J} e^{v_{ij}}} \qquad v_{ij} = \sum_{k \in T} b_k \times x_{ijk}$$

Los parámetros de la utilidad sistematica (Vij) y un termino independiente se estiman mediante máxima verosimilitud

Entre todos los valores que pueden adquirir los parámetros, se escogerá como el mejor valor aquel que hace más probable la elección observada

$$L(Y|X,\ B) = \prod_{i=1}^{N} P_i^{Y_i}(1 - P_i)^{1 - Y_i} = \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\frac{e^{v_{ij}(x_{ij}, b)}}{\sum_{h=1}^{J} e^{v_{ih}(x_{ih}, b)}} \right)^{Y_{ij}}$$

2. LOGIT MULTINOMIAL

La probabilidad de elección de una alternativa es función de todas las alternativas

$$P_i(Y = j) = \frac{e^{v_{ij}}}{\sum_{j=1}^{J} e^{v_{ij}}} \qquad v_{ij} = \sum_{k \in T} b_k \times x_{ijk}$$

Los parámetros de la utilidad sistemática (Vij) y un termino independiente se estiman mediante máxima verosimilitud

Entre todos los valores que pueden adquirir los parámetros, se escogerá como el mejor valor aquel que hace más probable la elección observada

$$\log L = \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \ln(P_i(Y = j/x,\ B)) \quad \text{with } P_i(Y = j/x,\ B) = \frac{e^{v_{ij}(x_{ij},b)}}{\sum_{h=1}^{J} e^{v_{ih}(x_{ih},b)}}$$

2. LOGIT MULTINOMIAL

La probabilidad de elección de una alternativa es función de todas las alternativas

$$P_i(Y=j) = \frac{e^{v_{ij}}}{\sum_{j=1}^{J} e^{v_{ij}}} \qquad v_{ij} = \sum_{k \in T} b_k \times x_{ijk}$$

$$P_i(Y=j) = \frac{1}{\sum_{k \in T} e^{(v_{ik} - v_{ik})}}$$

La probabilidad de elección es función de la diferencia de utilidades

2. LOGIT MULTINOMIAL

La probabilidad de elección de una alternativa es función de todas las alternativas

$$P_i(Y = j) = \frac{e^{v_{ij}}}{\sum_{j=1}^{J} e^{v_{ij}}}$$

$$P_i(Y = 1) = \frac{e^{(v_{i1})}}{\sum_{k \in S} e^{(v_{ik})}} = \frac{e^{(v_{i1})}}{e^0 + e^{(v_{i1})}} = \frac{e^{(v_{i1})}}{1 + e^{(v_{i1})}}$$

$$P_i(Y = j) = \frac{1}{\sum_{k \in T} e^{(v_{ik} - v_{ik})}}$$

Se recupera la regresión logística

2. MODELOS LOGIT

Modelos simplificados:	Alto sesgo baja varianza	Datos con ruido
Modelos complejos:	Alta varianza bajo sesgo	Datos de buena calidad y cantidad

Los modelos logit se basan en supuesto simplificadores para limitar la complejidad del modelo.

- Independencia de alternativas irrelevantes (bus rojo – bus azul)
- Todas las personas tienen disponibles todas las alternativas
- Efectos de las interacciones entre variables explicativas

Los individuos poseen información perfecta
Los individuos son racionales
Los individuos maximizan su nivel de utilidad
Los individuos actúan de forma determinística
Los individuos poseen un conjunto de alternativas

2. MODELOS LOGIT

Modelos simplificados:	Alto sesgo baja varianza	Datos con ruido
Modelos complejos:	Alta varianza bajo sesgo	Datos de buena calidad y cantidad

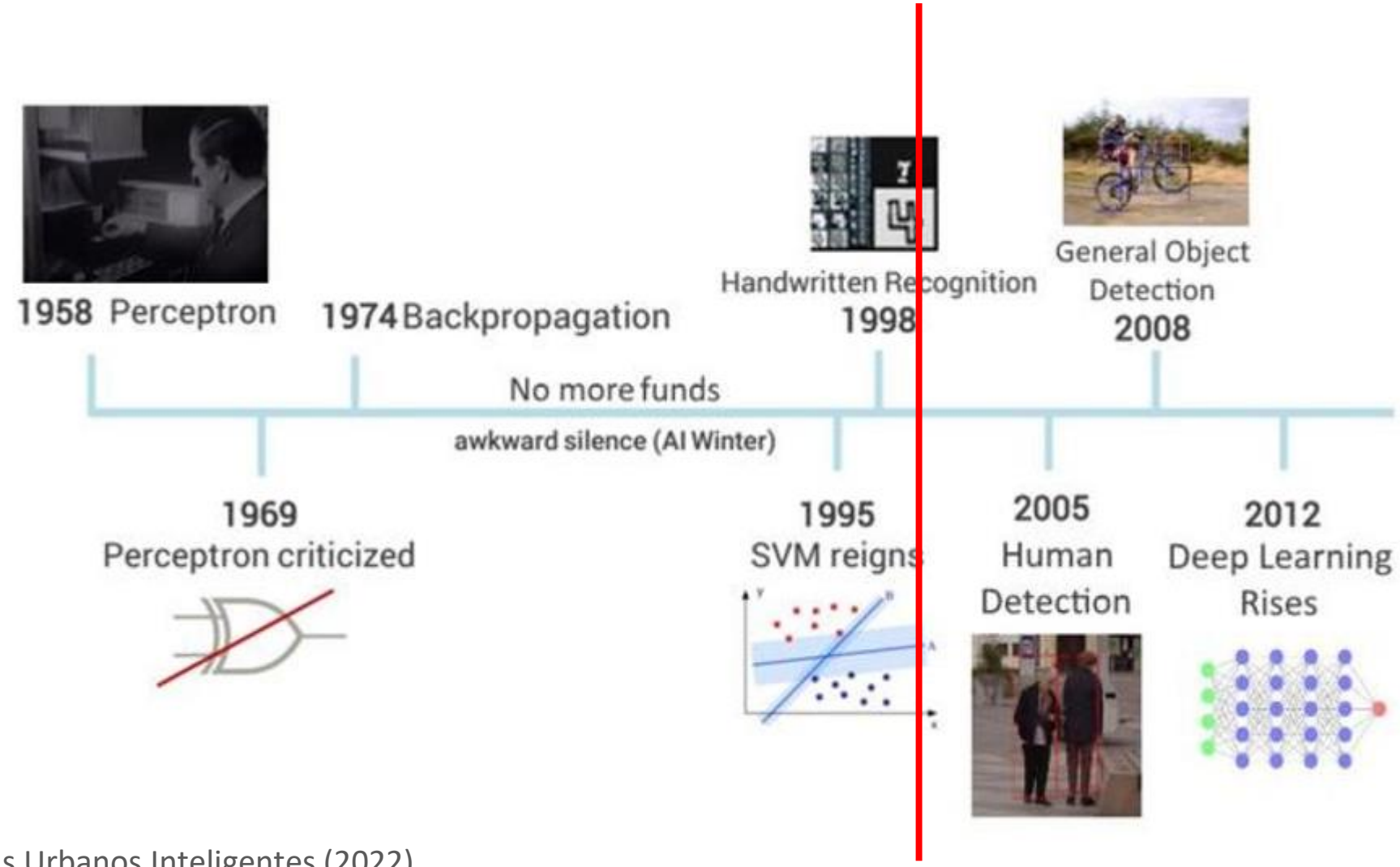
Los modelos logit se basan en supuesto simplificadores para limitar la complejidad del modelo.

- Independencia de alternativas irrelevantes (bus rojo – bus azul)
- Todas las personas tienen disponibles todas las alternativas
- **Efectos de las interacciones entre variables explicativas**

Es necesario especificar de forma adecuada las funciones de utilidad.

Proceso largo y tedioso que requiere conocimiento del dominio

3. REDES NEURONALES FEEDFORWARD



Fuente: Sistemas Urbanos Inteligentes (2022)

3. REDES NEURONALES FEEDFORWARD

Las redes neuronales se habían aplicado en varios problemas:

- Modelado de mercados financieros (Trippi y Turban, 1992)
- Análisis de riesgo crediticio (Burgess, 1995)
- Respuesta del mercado ante estrategias de marketing (Hruschka, 1993)

Las base de datos de para la elección de marca tenían una gran cantidad de ejemplos y en este problema existe interacción entre algunas variables (Ejemplo: precio - calidad)

Redes neuronales

Feedforward

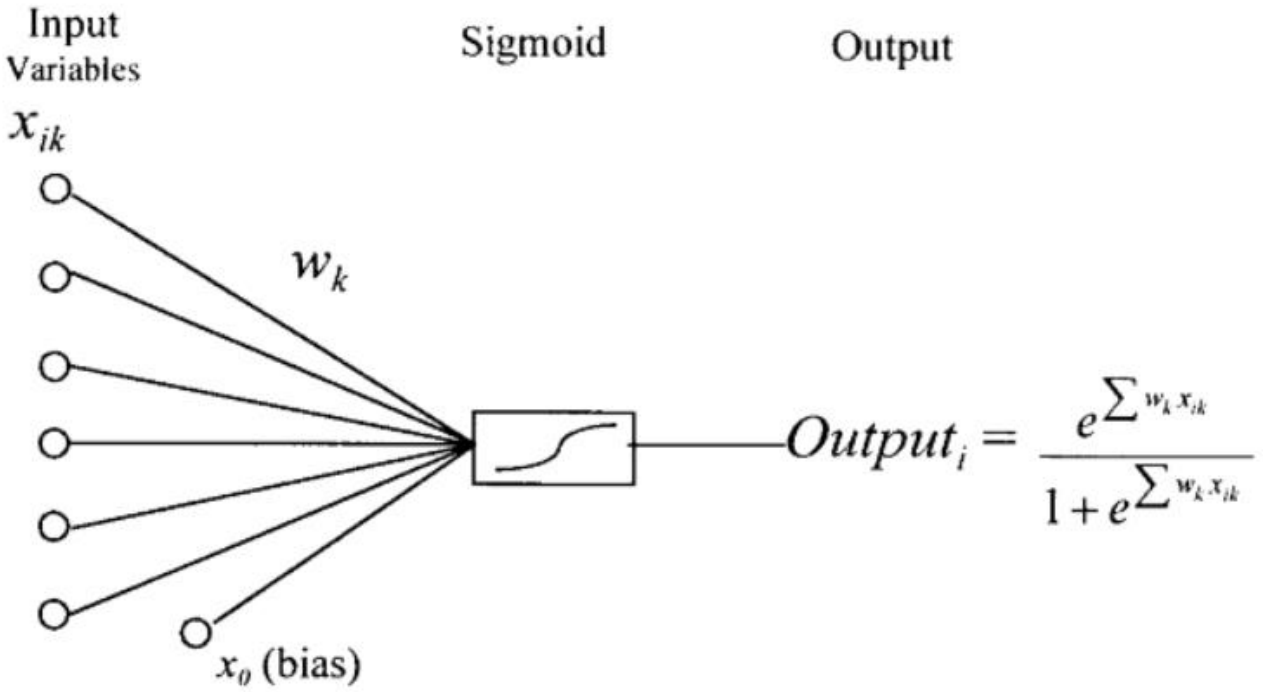
Una capa oculta

Función de activación logística o Softmax

Backpropagation

3. REDES NEURONALES FEEDFORWARD

ARQUITECTURA DE LA RED – LOGIT BINOMIAL – RELACION LINEAL

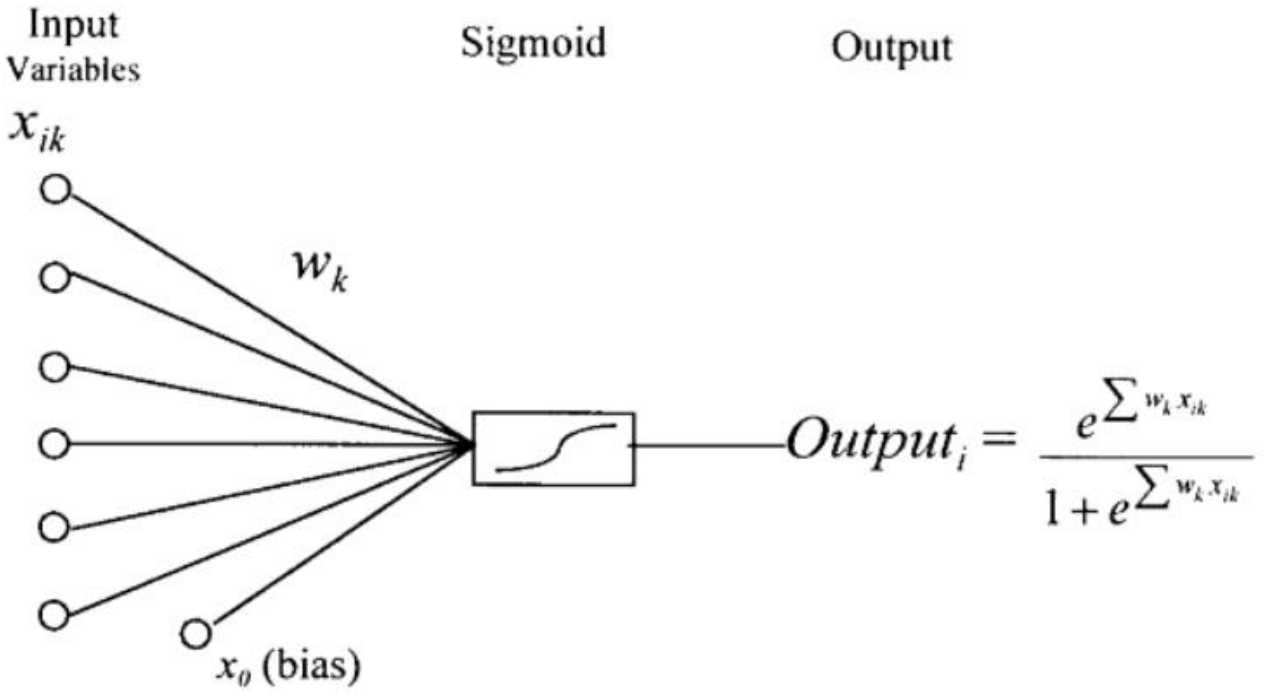


$$P_i(Y=1) = f(\text{Utility}) \qquad \text{Utility} = a + \sum_{j \in T} b_j \times x_{ij} + \varepsilon \qquad f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Fuente: Yves Bentz, Dwight Merunka (2000)

3. REDES NEURONALES FEEDFORWARD

ARQUITECTURA DE LA RED – LOGIT BINOMIAL – RELACION LINEAL



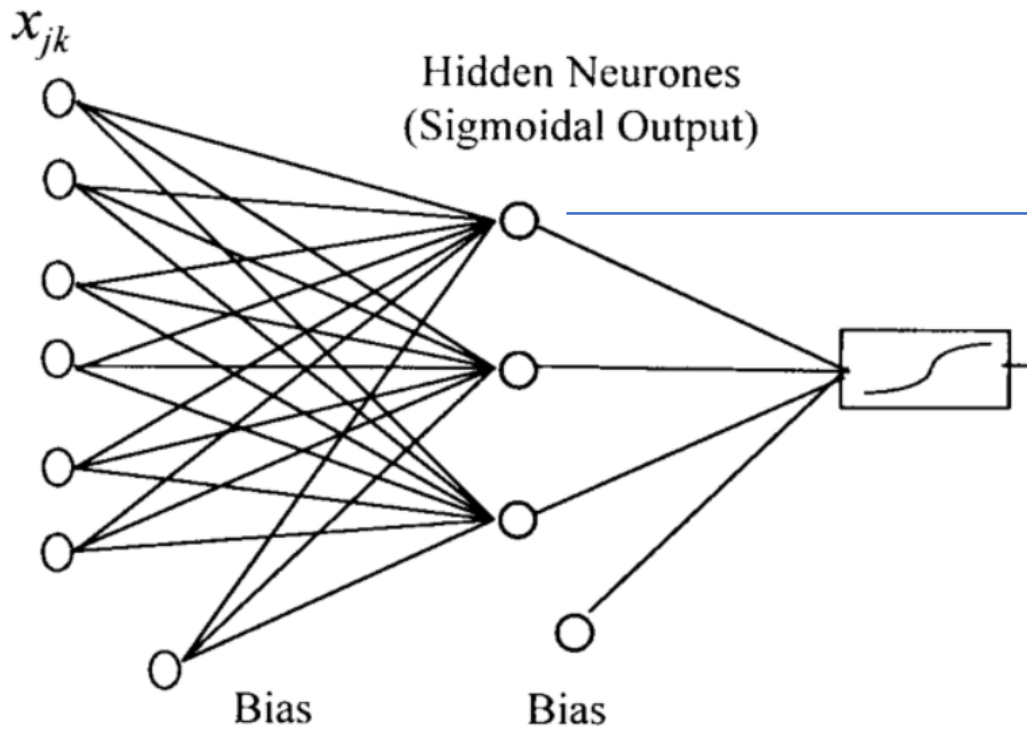
$$E = -\sum_{i=1}^{M}(Y_i \ln(S_i) + (1 - Y_i)\ln(1 - S_i))$$

La misma función que log verosimilitud para dos alternativas (en negativo)

Fuente: Yves Bentz, Dwight Merunka (2000)

3. REDES NEURONALES FEEDFORWARD

ARQUITECTURA DE LA RED – LOGIT BINOMIAL – RELACION NO LINEAL



Relación lineal que se activa o desactiva

$$Output_i = \frac{e^{g(x_{i1},\ldots,x_{ik})}}{1+e^{g(x_{i1},\ldots,x_{ik})}}$$

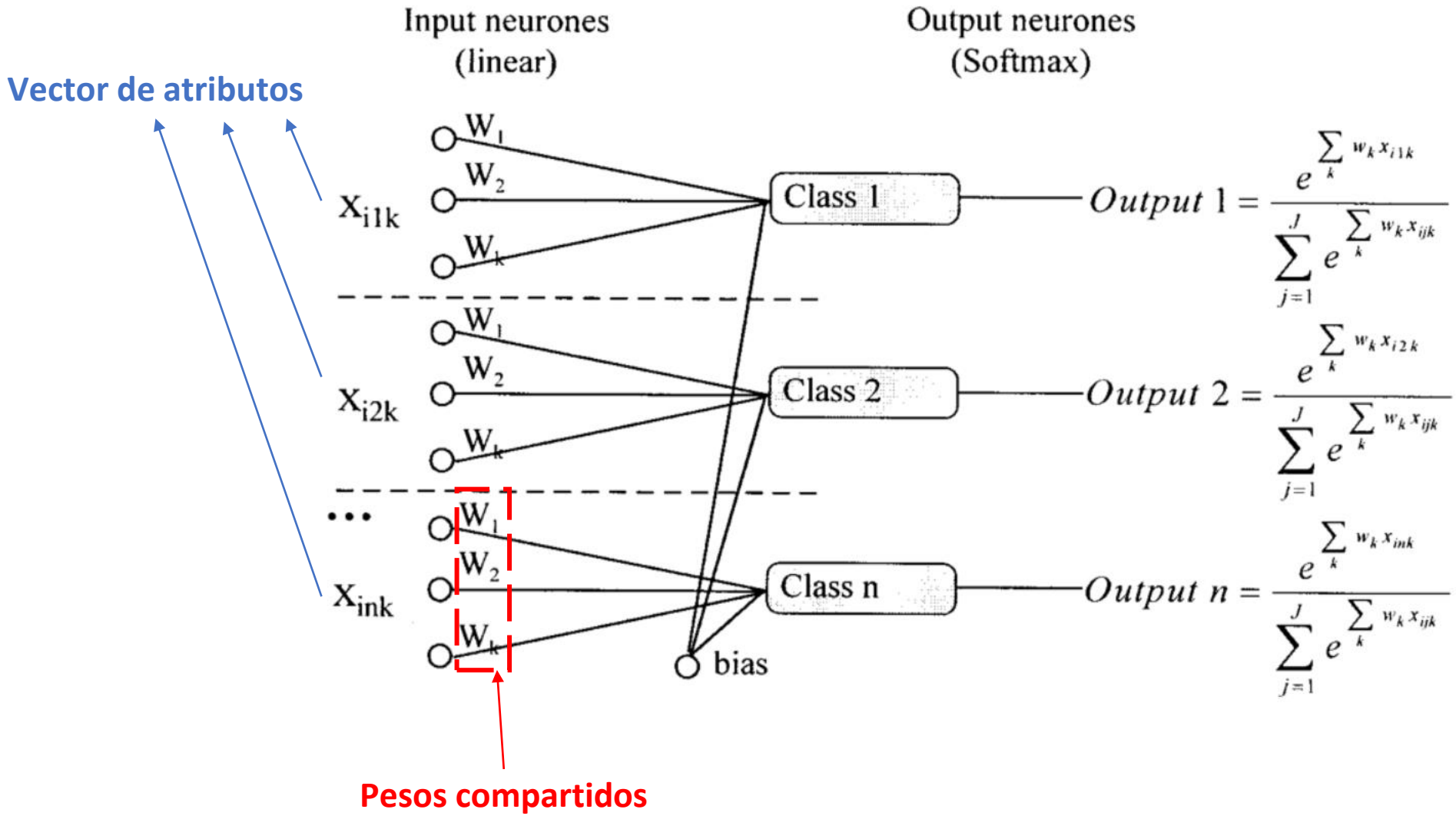
$$P_i(Y=1) = \frac{1}{1+e^{-g(\vec{x}_i)}}$$

Aproxima cualquier función, incluido el ruido

Fuente: Yves Bentz, Dwight Merunka (2000)

3. REDES NEURONALES FEEDFORWARD

ARQUITECTURA DE LA RED – LOGIT MULTINOMIAL – RELACION NO LINEAL



$$E = -\sum_{i=1}^{M}\sum_{j=1}^{J} Y_{ij} \ln(S_{ij})$$

Fuente: Yves Bentz, Dwight Merunka (2000)

3. REDES NEURONALES FEEDFORWARD

Relaciones lineales	→	**Sin capas ocultas**
Logit binomial	→	Un perceptrón con salidas logarítmicas
Logit multinomial	→	Tantos perceptrones como alternativas
Relaciones no lineales	→	**Una capa oculta**
Logit binomial	→	Un perceptrón con salidas logarítmicas
Logit multinomial	→	Tantos perceptrones como alternativas

3. REDES NEURONALES FEEDFORWARD

¿ Qué es mejor ?

Pueden incurrir en sesgo de especificación	Red neuronal no tiene sesgo de especificación
Requiere criterio experto	Aprende relaciones complejas sin el criterio experto
Se obtienen coeficientes interpretables	Son de difícil interpretación
Tiene estadísticas de significancia	Carecen de indicadores estadísticos

Si la función ajustada no es demasiado compleja, es posible visualizar efectos no lineales de las variables de entrada.

Proyectando la función en subconjuntos del espacio de entrada

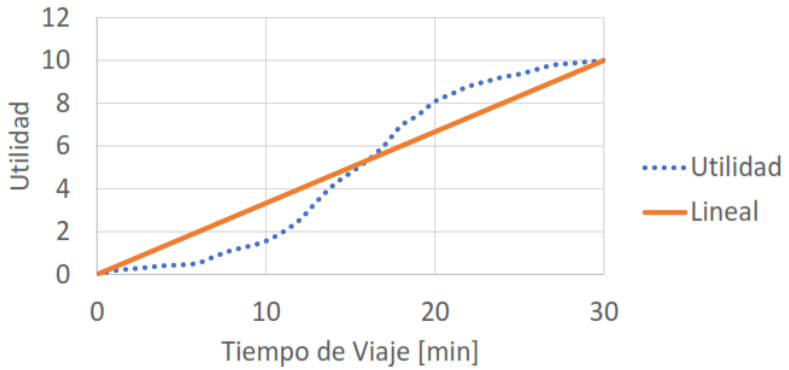
3. REDES NEURONALES FEEDFORWARD

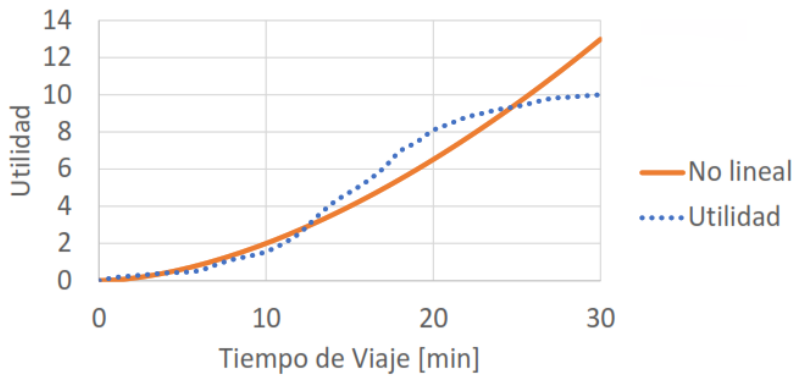
¿ Qué es mejor ?

Si la función ajustada no es demasiado compleja, es posible visualizar efectos no lineales de las variables de entrada.

Proyectando la función en subconjuntos del espacio de entrada

$$V_i = ASC_i + \theta_{TV} \cdot TV_i^{\gamma_{TV}} + \theta_{Costo} \cdot Costo_i$$





Fuente: Modelos de Demanda de Transporte (2020)

3. REDES NEURONALES FEEDFORWARD

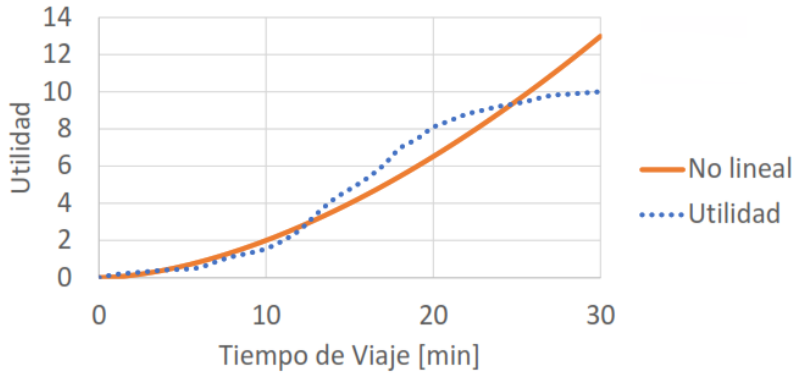
¿ Qué es mejor ?

Si la función ajustada no es demasiado compleja, es posible visualizar efectos no lineales de las variables de entrada.

Proyectando la función en subconjuntos del espacio de entrada

Mejor rendimiento predictivo de MNL

$$V_i = ASC_i + \theta_{TV} \cdot TV_i^{\gamma_{TV}} + \theta_{Costo} \cdot Costo_i$$



Fuente: Modelos de Demanda de Transporte (2020)

CASO 1: Elección de marca de chocolate con datos sintéticos

Variable dependiente:
Consumo propio, precio mayor
Para un hijo, precio menor

Elección de marca de chocolate:

Datos sintéticos

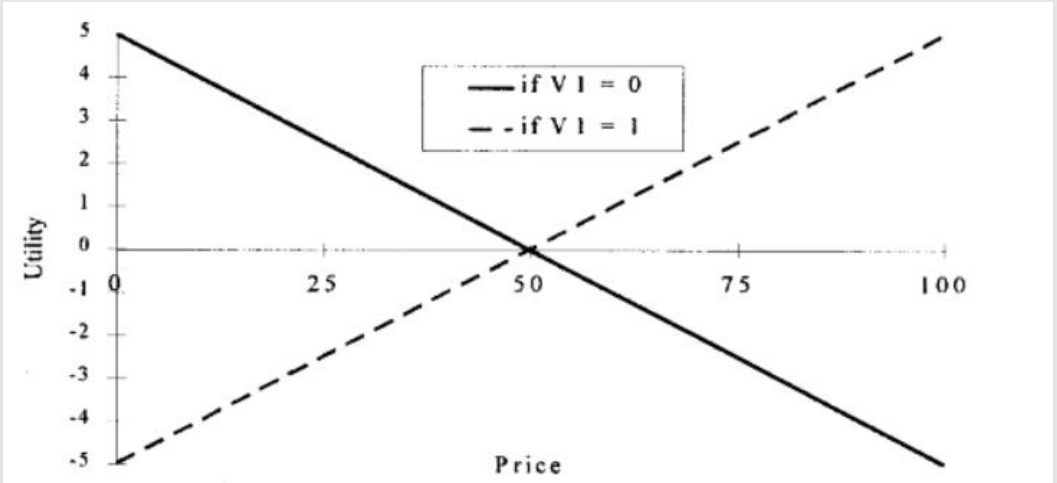
Preferencias conocidas

Tres marcas de chocolate

Dos variables independientes:

V1 (dummy)
La compra es para consumo propio
La compra es para un hijo

V2
Precio por cantidad



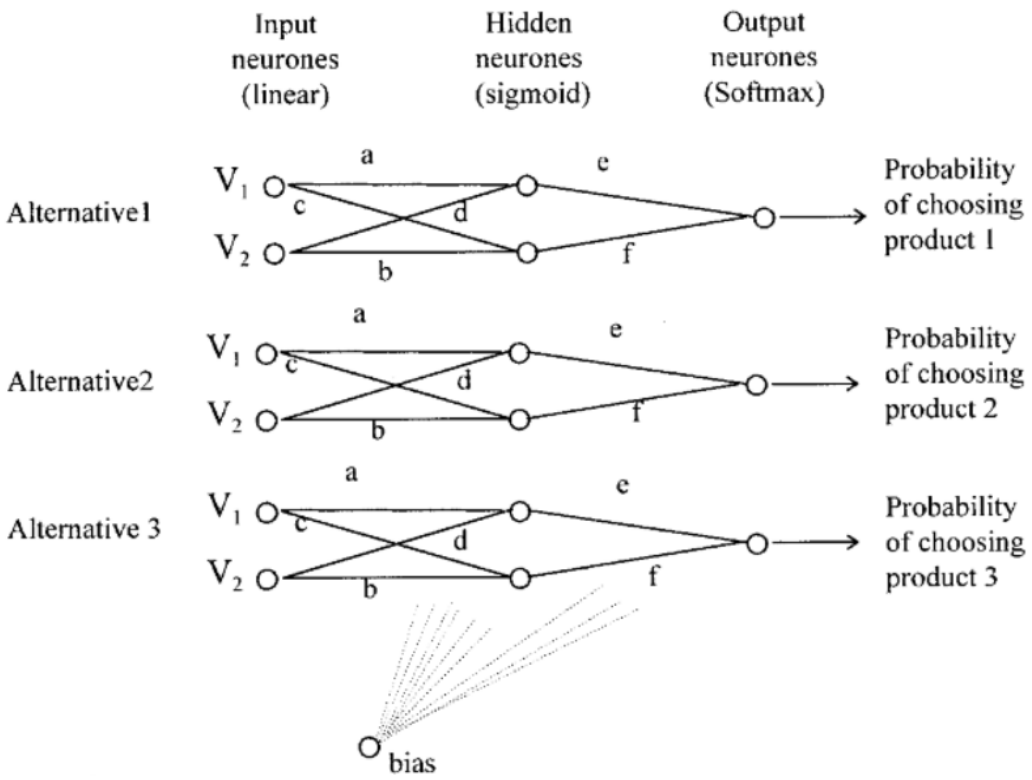
$$\text{Utility} = \begin{cases} -0.1 \cdot V_2 + 5 & \text{if } V_1 = 0 \text{ (Chocolate for children)} \\ +0.1 \cdot V_2 + 5 & \text{if } V_1 = 1 \text{ (Chocolate for consumer)} \end{cases}$$

Con la función de utilidad se calcula la probabilidad y en base a ella se genera la elección aleatoria

$$P_i(Y = j) = \frac{e^{v_{ij}}}{\sum_{j=1}^{J} e^{v_{ij}}}$$

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 1: Elección de marca de chocolate con datos sintéticos



Rendimiento del modelo

Comparación de la probabilidad de elección con lo predicho

$$R^2 = 1 - \frac{\text{MSE}}{\text{Var}}$$

→ Error cuadrático medio del modelo

→ Varianza de la probabilidad a modelar

Que tan mejor es mi modelo que el asar

$$U^2 = 1 - \frac{L(X)}{L_0}$$

→ Logverosimiltud del modelo estimado

→ Logverosimiltud de modelo aleatorio

Que tan significativos son los coeficientes

$$t = \frac{\hat{\theta}_k - \bar{\theta}_k}{\hat{\sigma}_\theta} \sim N(0,1)$$

$t_{crítico,\alpha} = 1{,}96$ para $\alpha = 95\%$

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 1: Elección de marca de chocolate con datos sintéticos

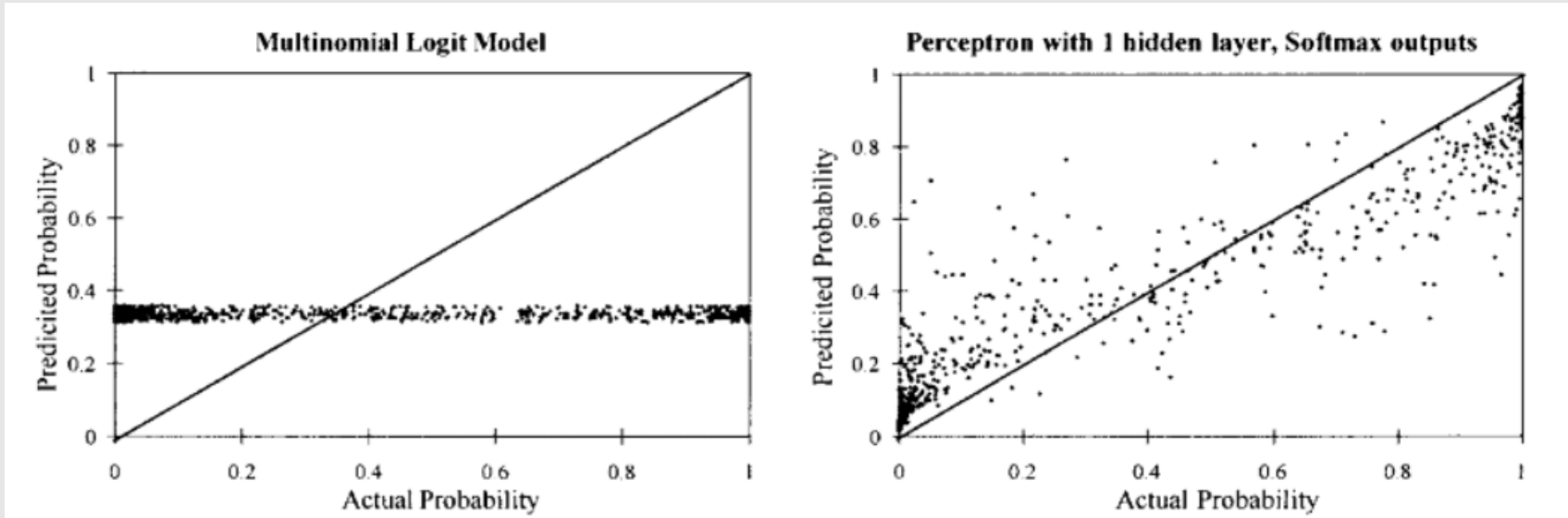


Table I. Comparison of performance measures for the MNL model and ANN model

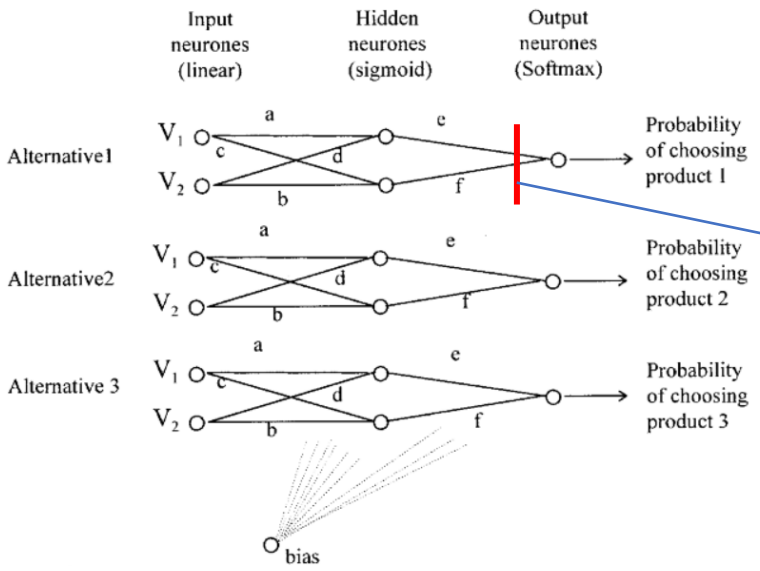
Performance measures	MNL	ANN
U^2	0.049	0.52
R^2	0.002	0.845
t-stat. for variable 1 (P-value)	−0.67 (0.50)	
t-stat. for variable 2 (P-value)	0.072 (0.94)	

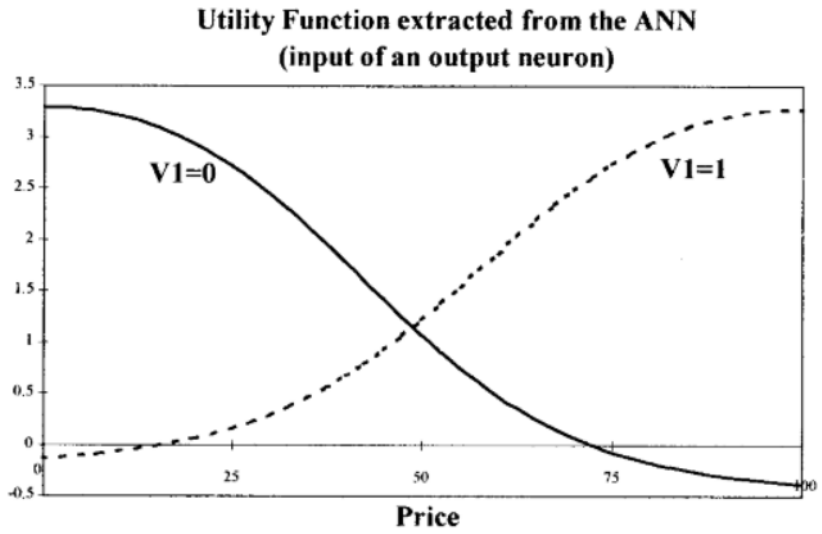
La función de utilidad lineal no capturó lo que la red neuronal sí

Entonces se puede explorar no linealidades

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 1: Elección de marca de chocolate con datos sintéticos





Entonces hay una interacción entre variables

Agregando un termino de interacción:

Table I. Comparison of performance measures for the MNL model and ANN model

Performance measures	MNL	ANN
U^2	0.049	0.52
R^2	0.002	0.845
t-stat. for variable 1 (P-value)	−0.67 (0.50)	
t-stat. for variable 2 (P-value)	0.072 (0.94)	

Table II. Performances of the respecified MNL model

Performance measures	MNL
U^2	0.059
R^2	0.99
t-stat. for variable 1 (P-value)	$-10.7\ (10^{-26})$
t-stat. for variable 2 (P-value)	$-9.9\ (10^{-22})$
t-stat. for variable 3 (P-value)	$11.23\ (10^{-28})$

Mejor rendimiento del modelo y coeficientes representativos

En un mercado compuesto por segmentos de con comportamientos diferentes la especificación de interacción resulta relevante y de difícil especificación

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 2: Pronosticar compra individuales de café instantaneo

Efectos de variables de marketing sobre la compra de café instantáneo

- ✓ 89 tiendas (6 tiendas)
- ✓ 5 ciudades australianas
- ✓ 25 000 compras (4952 compras)
- ✓ 1500 hogares (937 hogares)
- ✓ Entre los años 1993 y 1994

46 marcas de café (3 principales 63% de la participación de mercado)

Datos: 4952 compras

Entrenamiento: 2 899 compras
Validación cruzada: 322 compras
Test: 1 731 compras

Alternativas

A. Nescafé Blend, grande
B. National Roast, grande
C. Nescafé Blend, pequeño
D. National Roast, pequeño
E. Hbrand, grande

Variables

A. Precio por cantidad
B. % de reducción del precio normal
C. Dummy especificas de cada producto
D. Propias de cada marca
E. Lealdad a la marca
F. Lealdad al tamaño

Complejidad optima dado por:

$$E = -\sum_{i=1}^{M}\sum_{j=1}^{J} Y_{ij}\,\ln(S_{ij})$$

La red tiene 4 neuronas ocultas por alternativa

Se eliminaron hogares con pequeños volúmenes de compra (al menos 12 compras en 2 años y un compra en cada semestre)
Se controló la variedad de productos disponibles en las tiendas

CASO 2: Pronosticar compra individuales de café instantaneo

Rendimiento del modelo: $R^2, U^2, 2 * [logL(modelo\ 2) - logL(modelo\ 1)]$, t

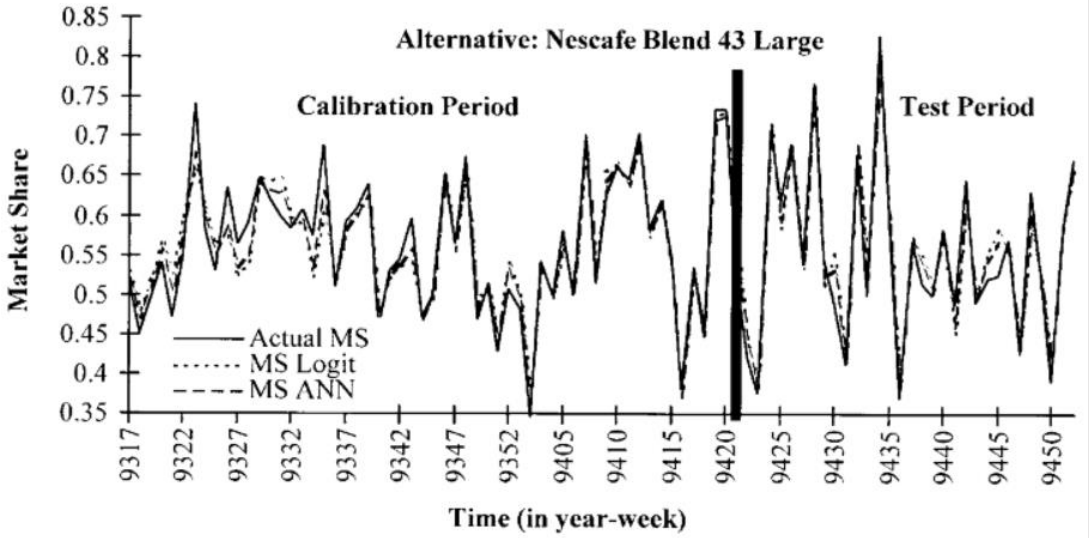
Table III. Comparison of model performances on calibration and test periods

Models	U^2 calibration	Average R^2 test	R^2 over the test period for each alternative				
			A	B	C	D	E
MNL	0.870	0.903	0.939	0.927	0.713	0.822	0.715
ANN	0.883	0.916	0.955	0.938	0.740	0.811	0.705

La red neuronal es mejor que MNL aun que no por mucho.

Puede existir un aporte débil de no linealidades

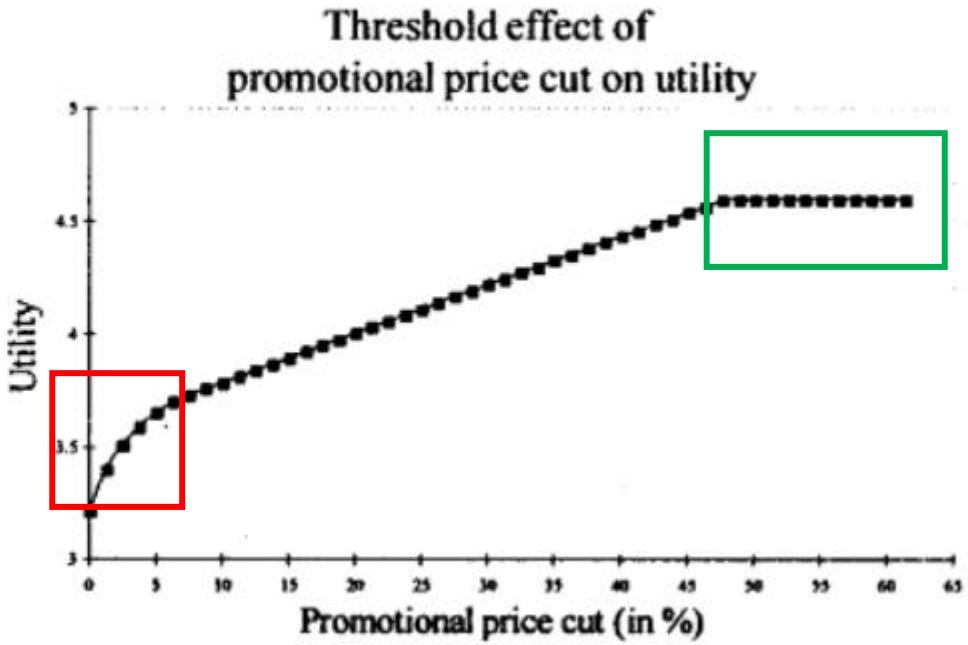
Hay muchos ejemplos de algunos productos que aportan en la reducción de los errores



La relación entre las variables de decisión y la probabilidad de elección parecen ser estacionarias

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 2: Pronosticar compra individuales de café instantaneo



Leal

No Leal

Fuente: Yves Bentz, Dwight Merunka (2000)

CASO 2: Pronosticar compra individuales de café instantaneo

Variables

A. Precio por cantidad
B. % de reducción del precio normal
C. Dummy especificas de cada producto
D. Propias de cada marca
E. Lealdad a la marca
F. Lealdad al tamaño
$x_9. E * B$
$x_{10}. E^2$
x_{11}. Dummy existe o no promoción

Modelos

O. Variables A, B, C, D, E, F
a. O + x_9
b. O + x_{10}
c. O + x_{11}
d. O + $x_9 + x_{10}$
e. O + $x_9 + x_{10} + x_{11}$

Table IV. Performances for respecified Logit models (non-linear utility function)

Model specifications		O	a	b	c	d	e
Performance measures	U^2 (calibration)	0.8701	0.8708	0.8709	0.8719	0.8717	0.8737
	χ^2 (significance level)	0	7.66 ($5.6\ 10^{-3}$)	8.34 ($3.9\ 10^{-3}$)	19.16 ($1.2\ 10^{-5}$)	16.30 ($2.9\ 10^{-4}$)	36.58 ($5.6\ 10^{-8}$)
	R^2 over test set	0.9027	0.9005	0.9003	0.9065	0.9025	0.9101
t-statistics for model variables	Depromoted price	−6.4	−6.4	−6.4	−6.2	−6.4	−6.3
	Price cut	6.2	6.6	6.1	0.1	6.6	1.8
	Brand loyalty	30.3	27.4	6.5	30.0	6.6	6.6
	Size loyalty	12.7	12.7	12.7	12.7	12.7	12.7
	x_9		−2.9			−2.9	−3.3
	x_{10}			−2.7		−2.8	−2.7
	x_{11}				4.3		4.5
t-statistics for model constants	A	6.8	6.4	6.4	6.4	6.1	5.7
	B	5.0	4.6	4.8	4.5	4.5	3.9
	C	−7.0	−7.1	−7.1	−7.2	−7.2	−7.4
	D	−8.2	−8.3	−8.3	−8.4	−8.4	−8.5

Critical values for t-statistics are: 1.645 (sig. level = 10%), 1.960 (s.l. = 5%), 2.576 (s.l. = 1%), 3.290 (s.l. = 0.1%).

Fuente: Yves Bentz, Dwight Merunka (2000)

7. CONCLUSIONES

Se puede representar con redes neuronales un modelo logit con una función de utilidad lineal o no lineal

El enfoque hibrido debería utilizarse cuando los datos sean muy complejos

No se pueden comparar ambos enfoques ya que la diferencia depende de la complejidad de los datos

La red neuronal aprende relaciones complejas sin el criterio experto (evitando el sesgo de especificación), permitiendo utilizar esta herramienta como diagnostico del MNL

Se puede extraer y graficar una función de utilidad de la red neuronal

No siempre es necesario utilizar este enfoque (relaciones no lineales debiles)

8. CRITICA

No se muestra como evoluciona la función de perdida al entrenar la red neuronal

No se menciona como se utilizaron los datos categóricos en la red neuronal

La descripción de la arquitectura de la red fue complicada de entender sobre todo como utilizar la función (Softmax)

Se destaca la importancia de una capa oculta (teorema de aproximación universal)

La cantidad de datos es baja para aplicar una red neuronal

Cuando la función de utilidad es compleja o existen muchas variables igual se necesita mucho tiempo para parametrizar e interpretar las funciones de utilidad que se pueden extraer