Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

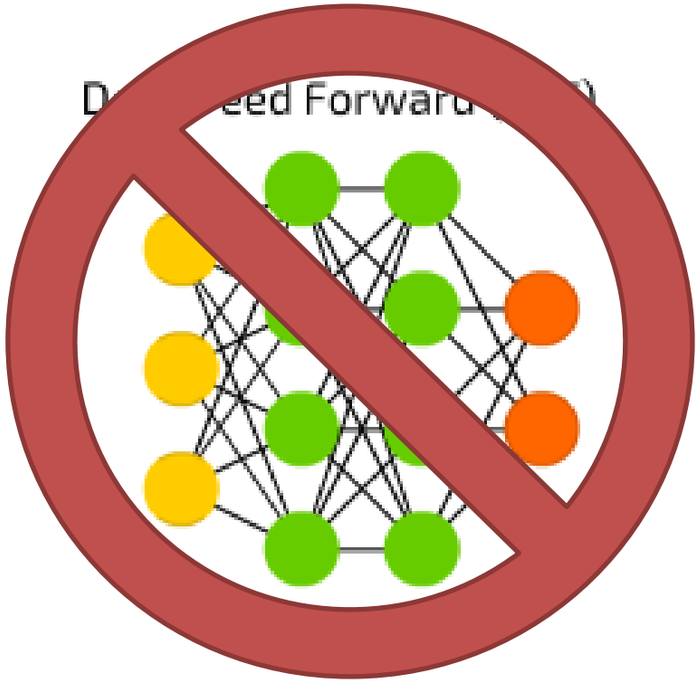
Sistemas Urbanos Inteligentes

Redes Neuronales y Datos Tabulados

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

Ahora que conocemos los ensambles, ¿dónde están las debilidades de las redes en datos tabulados?

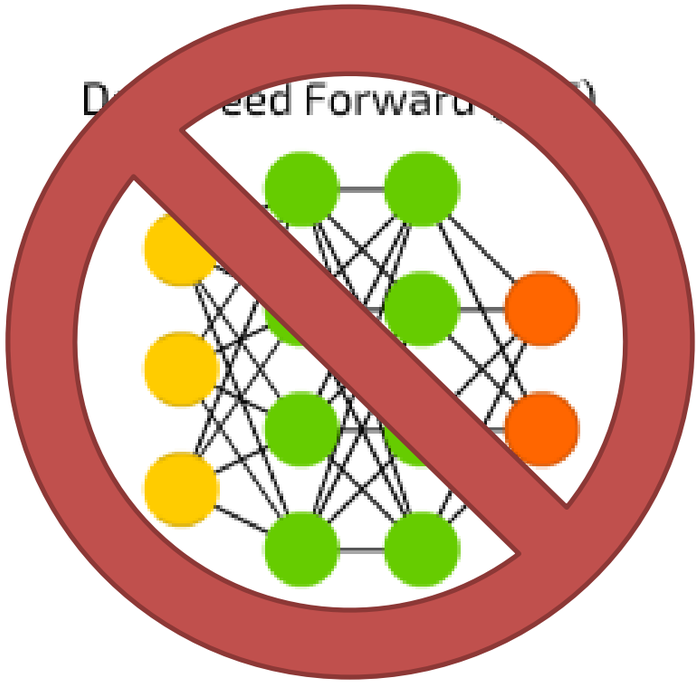
1. Discordancia en tipos de dato
2. Distribución y correlación en los datos



Ahora que conocemos los ensambles, ¿dónde están las debilidades de las redes en datos tabulados?

1. Discordancia en tipos de dato

   - Datos tabulados mezclan generalmente variables categóricas y numéricas.
   - Fuentes son heterogéneas, por lo que pueden tener distintas unidades y escalas de medición.
   - Variables son *sparse*, es decir, toman pocos posibles valores, incluso cuando son numéricas. Peor aún, puede haber datos faltantes.



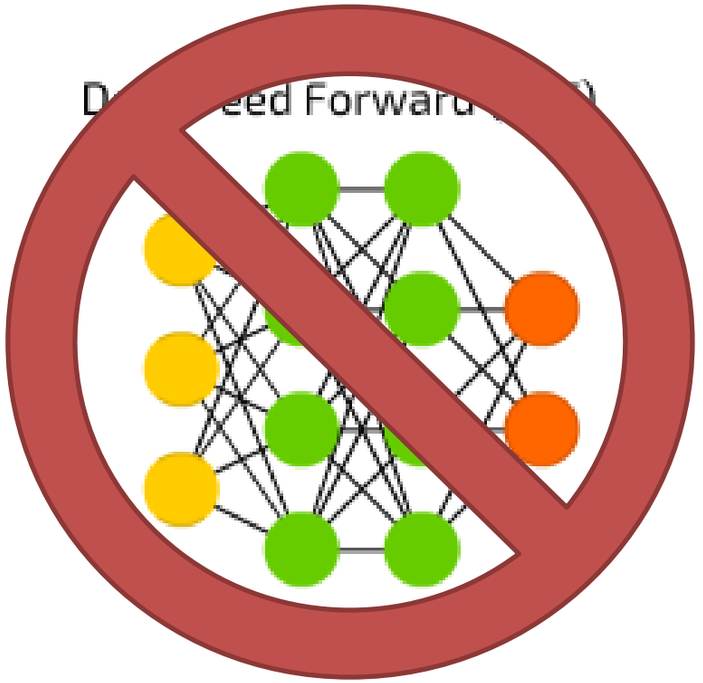
Solución tradicional sería a través de modelación

Ahora que conocemos los ensambles, ¿dónde están las debilidades de las redes en datos tabulados?

2. Distribución y correlación en los datos

- Columnas en datos tabulados presentan generalmente una alta correlación, por lo que basta un subconjunto pequeño de variables para poder hacer la predicción.

- Existe habitualmente un alto desbalance en las clases a predecir.



Solución tradicional sería a través la ingeniería de datos/características y del muestreo

Centrémonos inicialmente en el punto 1. Discordancia en tipos de dato

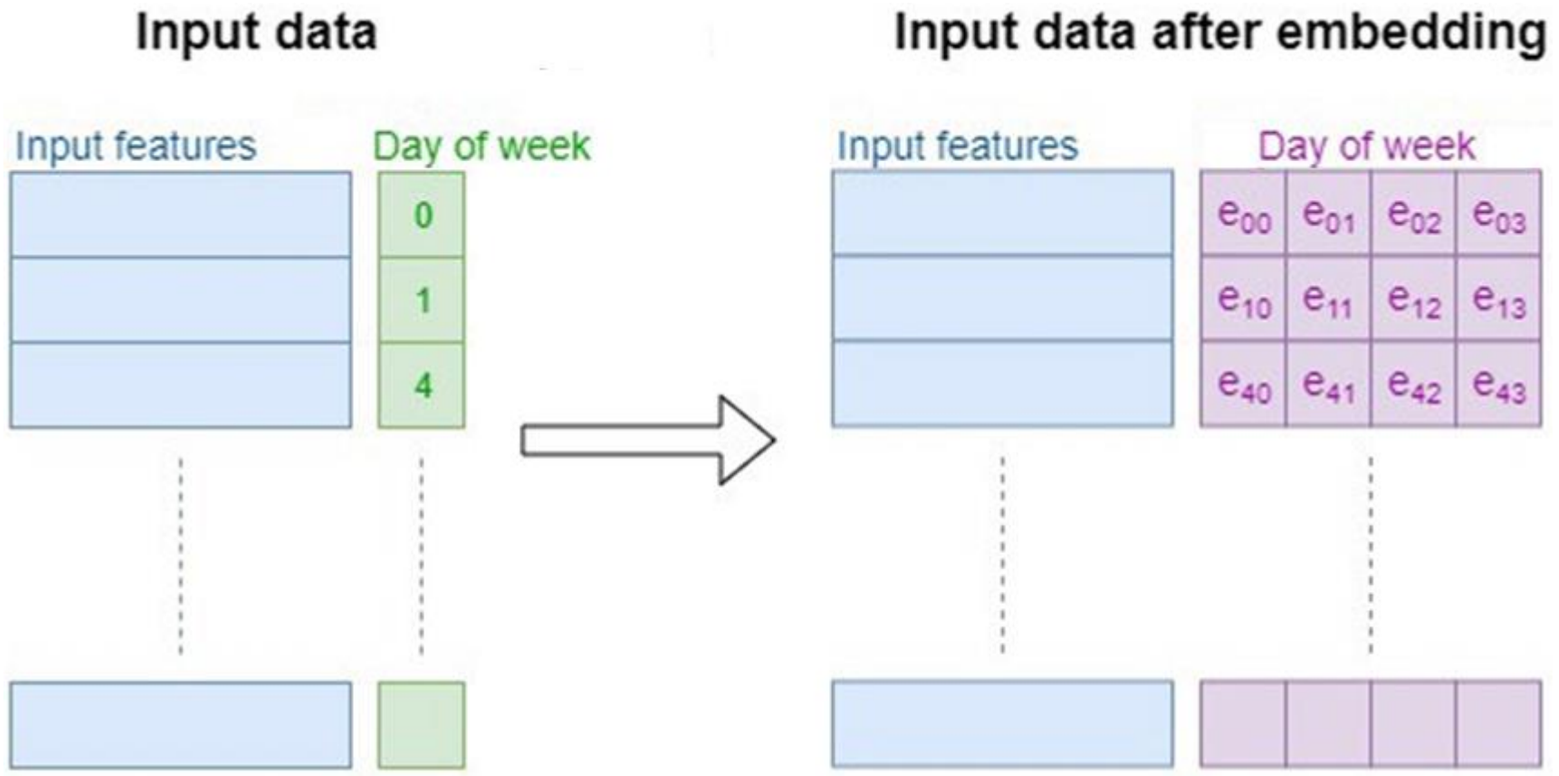
- Si bien la modelación es una herramienta fundamental, en este curso siempre intentaremos tomar inicialmente el enfoque del aprendizaje.
- Al analizar los problemas, notamos que “todo se reduce” a la incapacidad de las redes de trabajar con datos que no sean numéricos (continuos), como los categóricos.
- En vez de adaptar las redes para que funcionen con datos categóricos, buscaremos aprender transformaciones que muevan los datos categóricos a un espacio numérico.
- Esto tiene la ventaja que es posible capturar información semántica en este nuevo espacio (si es aprendida de forma *end-to-end*).
- Para hacer esto, utilizaremos el concepto de *embedding*.

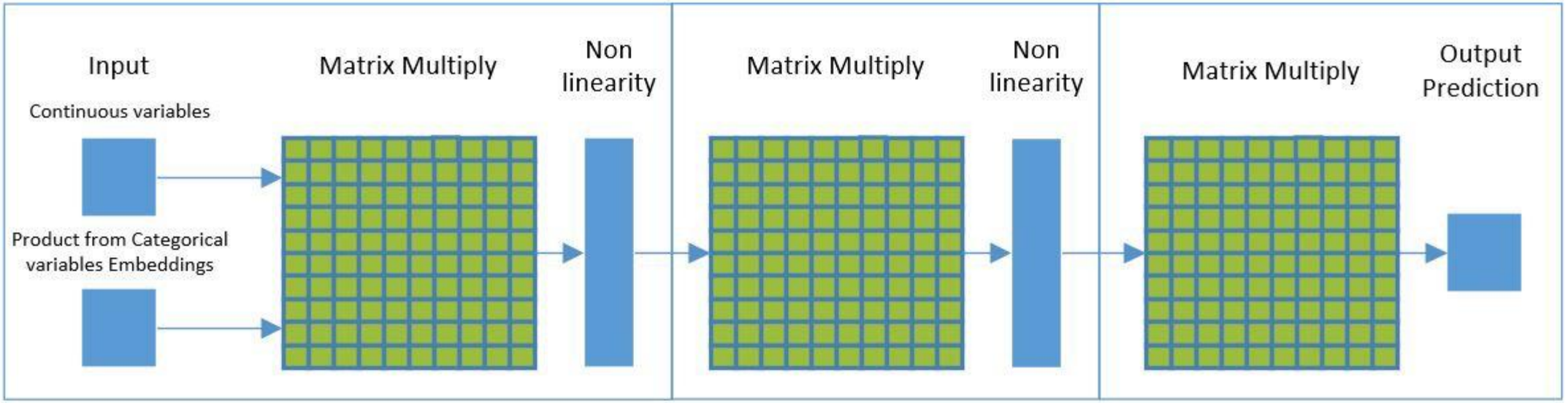
¿Qué es un *embedding*?

- En esencia, un *embedding* es un esquema para transformar datos que viven en espacios no numéricos, a un espacio vectorial continuo, generalmente de menor dimensionalidad (de ahí el nombre).

- Un *embedding* puedes puede ser diseñado o aprendido para una tarea.

- En ML, no solo se usan para transformar datos categóricos, sino también para grafos, texto y otros.


Input data
Input features
Day of week
0
1
4
Input data after embedding
Input features
Day of week
e00 e01 e02 e03
e10 e11 e12 e13
e40 e41 e42 e43


Input
Continuous variables
Product from Categorical variables Embeddings
Matrix Multiply
Non linearity
Matrix Multiply
Non linearity
Matrix Multiply
Output Prediction

¿Como integramos este concepto de *embedding* en una red neuronal?

- En realidad, las redes neuronales lo que hacen es aprender *embeddings*: transformaciones de datos de un espacio a otro.
- En algunos casos, como el *autoencoder*, esa transformación busca reducir la dimensionalidad, pero no es algo obligatorio.
- Hasta ahora, solo hemos hecho transformaciones entre espacios vectoriales continuos.
- ¿Qué cambios debemos hacer entonces en las redes para aprender transformaciones de espacios categóricos a continuos?

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Redes Neuronales y Datos Tabulados

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación