Revisión del libro de Molnar

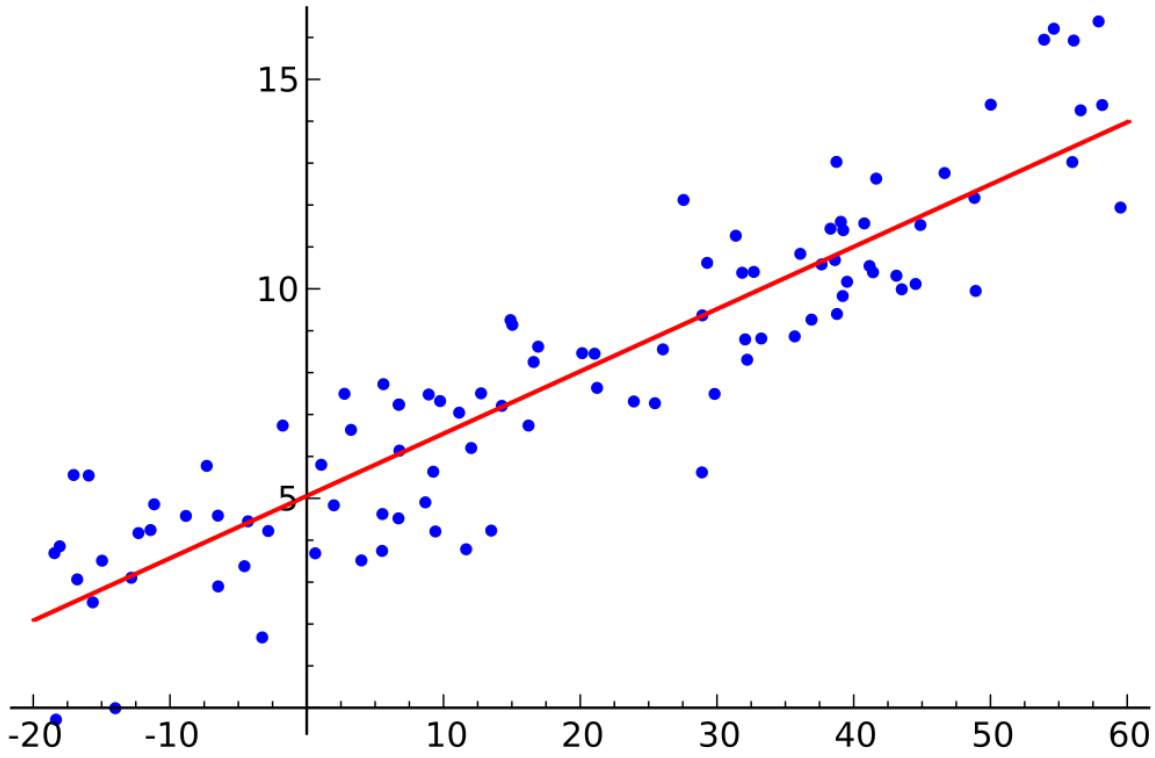
Modelos intrínsecamente interpretables

Javier París

Table of contents

1 Regresión lineal

1.1



1.2 Interpretación de los coeficientes en atributos numéricos

Dada la ecuación lineal a ajustar por el modelo:

$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n$

Se puede interpretar cada uno de los coeficientes ω_i como: **manteniendo todos los demás factores constantes, un incremento de una unidad en x_i se asocia con un incremento de ω_i en y.**

1.3 Interpretación según atributo

- **Atributos numéricos**: incremento en ω_i por incremento en una unidad en x.
- **Atributos categóricos**: incremento en ω_i por comparación con la categoría base.

- ω_0: valor esperado de y con todos los demás atributos como valor base ($x = 0$). Con todos los atributos normalizados, este valor pasa a ser el valor esperado para el dato promedio.

1.4 Feature importance

El valor de cada coeficiente está intrínsecamente relacionada con la importancia de cada atributo, pero también es altamente dependiente de la variación de dicho atributo. Por eso es importante **normalizar** los atributos para poder comparar los **coeficientes** entre sí.

Además, la correlación entre atributos puede hacer que los coeficientes no tengan sentido. Para eso se pueden usar ténicas de regularización como **LASSO**.

1.5 Pros / Cons

1.5.1 Ventajas

- Interpretación directa de los coeficientes. No hay caja negra.
- Ampliamente aceptado y utilizado.
- Garantía de encontrar el modelo óptimo (algebraicamente).

1.5.2 Desventajas

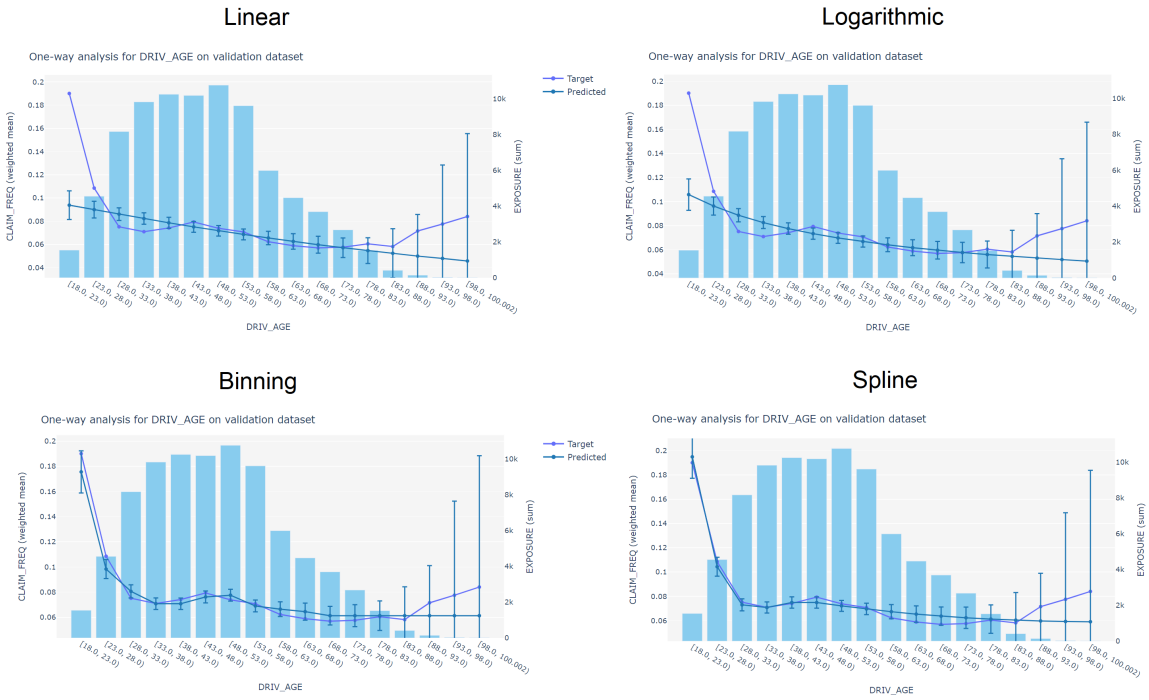
- No puede ajustar funciones no lineales.
- No modela relación entre atributos.
  - Atributos muy correlacionados pueden dar a coeficientes sin sentido (infinitas soluciones del sistema de ecuaciones).

1.6 Conclusiones

- Es un modelo fácilmente interpretable localmente. Es estrictamente monótono.
- Para poder ver la importancia de los atributos, es necesario normalizarlos y usar regularización.

2 Modelos basados en Regresión lineal

2.1



2.2 GLM (Generalized Linear Models)

Este tipo de modelos se basa en una regresión lineal ajustando la regresión a una distribución de probabilidad no gaussiana, y aplicándole una función de enlace.

$g(E(y|x)) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + ... + \omega_n x_n$

El caso de la regresión logística es un caso particular de GLM donde $g = ln$ y la distribución es binomial.

2.3 Interpretación GLM

La interpretación de los coeficientes depende de la función de enlace:

- **Identidad**: Los coeficientes siguen interpretándose como la suma
- **Logarítmica**: Los coeficientes pasan a ser multiplicativos (como en la regresión logística)

- **Otras**: depende de la función de enlace, y en muchos casos no tienen interpretación directa.

2.4 GAM (Generalized Additive Models)

Los GAM son una extensión de los GLM que permiten ajustar funciones no lineales a los atributos.

$g(E(y|x)) = \omega_0 + f_1(x_1) + f_2(x_2) + ... + f_n(x_n)$

Normalmente se usan *splines* para ajustar funciones no lineales a los atributos.

2.5 Pros / Cons

2.5.1 Ventajas

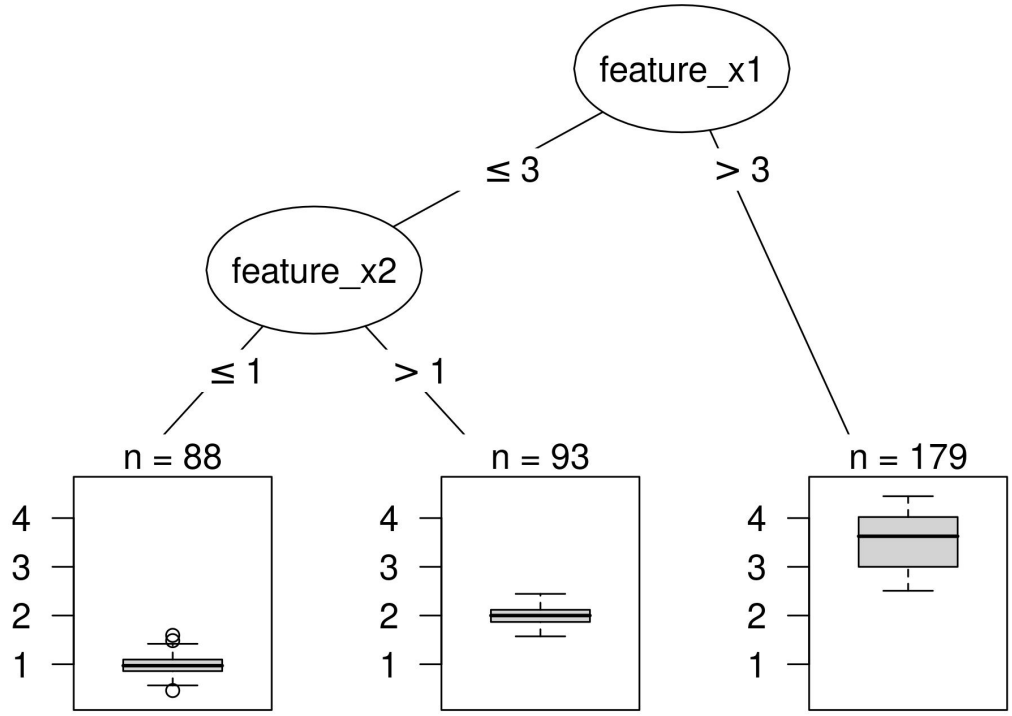
- Se pueden ajustar gran cantidad de funciones no lineales.
- Áun mantienen parte de la interpretabilidad de los modelos lineales.

2.5.2 Desventajas

- Estos modelos son difíciles de pre-ajustar y son altamente dependientes de los datos.
- Son menos interpretables.
- Asumen ciertas características de los datos.

3 Árboles de decisión

3.1



3.2 Interpretación de los nodos

Un árbol de decisión se puede interpretar como una división del espacio de atributos en subconjuntos, donde cada subconjunto se asocia con un nodo hoja.

$y = \sum_{a=1}^{A} c_a I(x \in R_i)$

Los árboles de decisión se interpretan por la decisión en cada nodo, y la importancia de cada atributo se mide por la cantidad de veces que se usa en el árbol.

3.3 Pros / Cons

3.3.1 Ventajas

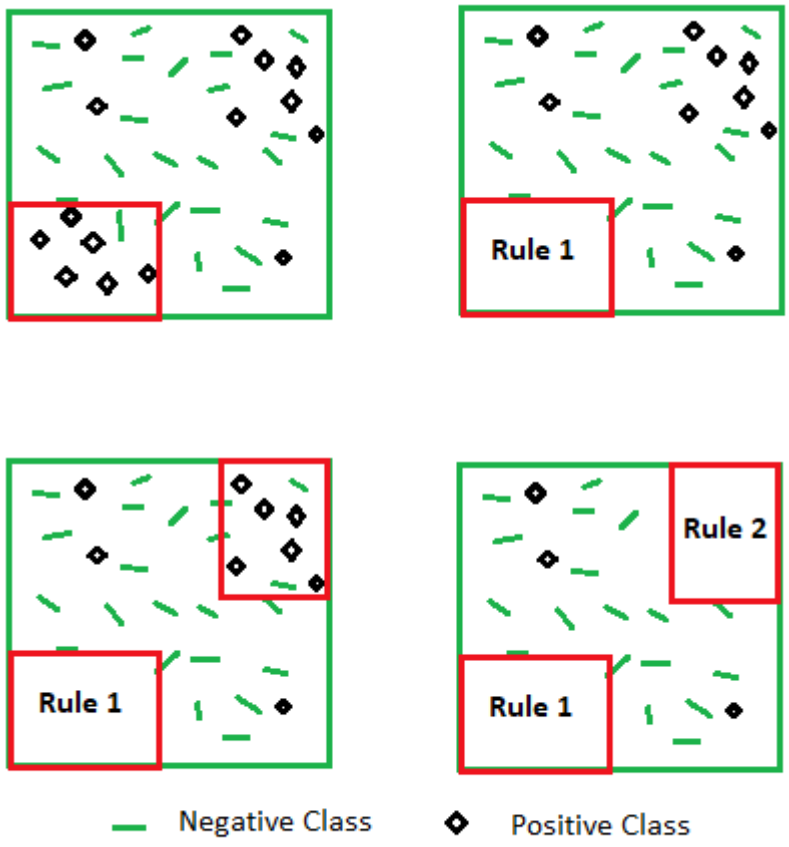
- Es capaz de modelar funciones no lineales y relaciones entre atributos.
- Tiene una visualización directa y sencilla.
- Es fácilmente interpretable debido a su naturaleza de explicar las decisiones como "what if".

3.3.2 Desventajas

- No son capaces de modelar relaciones lineales.
  - Un mismo cambio en un dato puede hacer que la predicción no cambie o que cambie drásticamente.
- Son altamente inestables. Dependen altamente del dataset y de la decisión de qué atributos elegir en qué orden.
- El número de nodos final aumenta exponencialmente con la profundidad del árbol.

4 Modelos basados en reglas

4.1



4.2 Reglas de decisión

$IF(cond[\ \&\ ]) \rightarrow THEN(class)$

Las reglas de estos modelos son altamente interpretables ya que se asemejan al lenguaje natural.

Cada reglar se puede medir principalmente con 2 valores, que suelen ser inversamente proporcionales:

- **Soporte/Cobertura**: cuántas veces se cumple la regla.
- **Precisión**: cuántas veces la regla acierta.

4.3 Pros / Cons

4.3.1 Ventajas

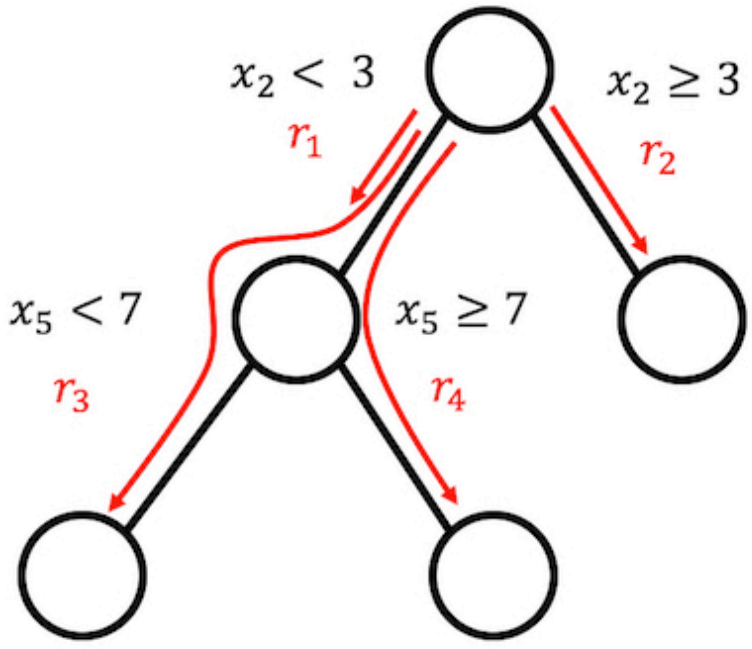
- Son altamente interpretables.
- Son muy similares a los árboles, aunque en general más compactos.
- Son robustos (no inestables) frente a cambios en los datos o outliers.
- Solo utilizan los atributos relevantes.

4.3.2 Desventajas

- No sirven para problemas de regresión
- En muchos casos por el método o por motivos de complejidad, los atributos deben ser categóricos
- No son capaces de modelar relaciones lineales.
- En muchos casos se da overfitting.

5 Rule fit

5.1



5.2 Implementación

Modelo que ajusta nodos hoja de un árbol de decisión como atributos de un modelo lineal.

- Se generan reglas a partir de un árbol de decisión.
- Cada nodo hoja se interpreta como un atributo binario.
- Se añaden los atributos numéricos originales.
- Se entrena un modelo lineal con estos atributos utilizando regularización LASSO.

5.3 Interpretación

La importancia de cada atributo se mide como en un modelo lineal:

- **Atributos numéricos**: incremento en ω_i por incremento en una unidad en x.
- **Reglas**: incremento en ω_i si se cumple la regla.

5.4 Pros / Cons

5.4.1 Ventajas

- Todas las ventajas del modelo lineal
- Añade las interacciones entre atributos al modelo
- Es fácil localmente interpretable, ya que las reglas suelen aplicar a regiones pequeñas

5.4.2 Desventajas

- Puede generar muchas reglas que hagan el modelo demasiado complejo
- La interpretación sigue fallando en el mismo caso que la lineal: solo es interpretable si el resto de atributos son constantes

6 Otros

6.1 Naive Bayes

$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$

Es fácilmente interpretable la importancia de cada atributo $P(x|C_k)$.

6.2 KNN

Al ser un modelo basado en instancias (datos) no puede tener ciertas interpretaciones, como global o modular.

El modelo es interpretable en tanto en cuanto sus atributos (una instancia concreta) son interpretables. Es decir, su interpretabilidad se reduce con el número de atributos.