Interpretabilidad

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Junio de 2022

Contenido

Introducción

- Los modelos de aprendizaje de máquinas son diseñados fundamentalemmenta para detectar patrones y/o generalizar bien (i.e., predecir).
- Con este objetivo en ocasiones es óptimo usar modelos con muchas características difíciles de intepretar (i.e., vectorizacion de imágenes o textos).
- Para predecir es común usar modelos no lineales.
- Estas características hacen difícil hacer inferencia e intepretar los modelos (e.g, análisis de causalidad).

Introducción

- Vamos a estudiar algunas técnicas de interpetabiliad dejando el caso particular de análisis causal para otra sesión.
- Los temas principales son:
  1. Importancia de variables en árboles y bosques aleatorios.
  2. Esperanza condicional y dependencia parcial.
  3. Local interpretable model explanation (LIME).
  4. Juegos cooperativos y Shapley Additive Explanations (SHAP).
  5. Aplicación: Interpretando modelos GANs de crimen en Bogotá.

Contenido

Árboles

- Dado un nodo t del árbol, sea $v(t)$ la variable que se usa en ese nodo para dividir en el árbol. Sea $L(v(t))$ la pérdida en ese nodo si no se hiciera la división y $L_1(t)$ y $L_2(t)$ la pérdida en cada una de las divisiones que se realizan.
- Definamos la importancia de usar la variable l en el árbol T como:

$$I_l^2(T) = \sum_t i_t^2 I(v(t) = l)$$

donde $i_t^2 = (L(v(t)) - L(1) - L(2))^2$

- Para un bosque aleatorio se define como:

$$I_l^2 = \frac{1}{B} \sum_b I_l^2(T_b)$$

- Se normaliza la importancia más alta en 100.

Ejemplo: Importancia relativa

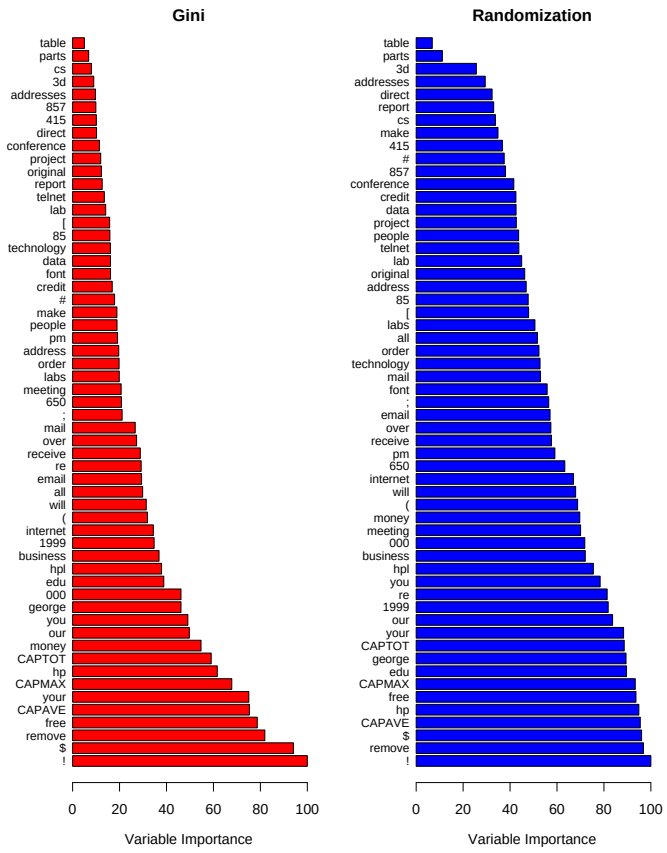


FIGURE 15.5. *Variable importance plots for a classification random forest grown on the* `spam` *data. The left plot bases the importance on the Gini splitting index, as in gradient boosting. The rankings compare well with the rankings produced by gradient boosting (Figure 10.6 on page 316). The right plot uses* OOB *randomization to compute variable importances, and tends to spread the importances more uniformly.*

Contenido

Gráficos de Dependencia Parcial

- Una forma de estimar la importancia de variables en modelos arbirarios (cajas negras) son los gráficos de dependencia parcial de Friedman.
- Más adelante vamos a ver bajo que circunstancias pueden utilizarse para estimar efectos causales.

Gráficos de Dependencia Parcial

- Sea X_S y $Z = X_C$ un conjunto de variables y su complemento respectivamente.
- Sea h una función de aprendizaje: $Y = h(X_S, Z)$. Definimos la dependencia parcial $h_{X_S}(x_S)$, de Y en X_S en el punto x_S como:

$$E_Z[h(x_S, Z)] = \int h(x_S, Z) dF(z)$$

- Obsérvese que aquí se está integrando sobre la distribución marginal de Z.
- Esto es diferente a la esperanza condicional: $E[h(X_S, Z) \mid X_S = x]$, en donde se integra con respecto a la distribución condicional de Z a $X = x$.

Gráficos de Dependencia Parcial

- La dependencia parcial $h_{X_S}(x_S)$, de Y en X_S en el punto x_S como:

$$E_Z[h(x_S, Z)] = \int h(x_S, Z) dF(z)$$

- Un estimador de la dependencia parcial es:

$$E_Z[h(x_S, Z)] = \frac{1}{n} \sum_{i=1}^{n} h(x_S, z_i) \tag{1}$$

Gráficos de Dependencia Parcial

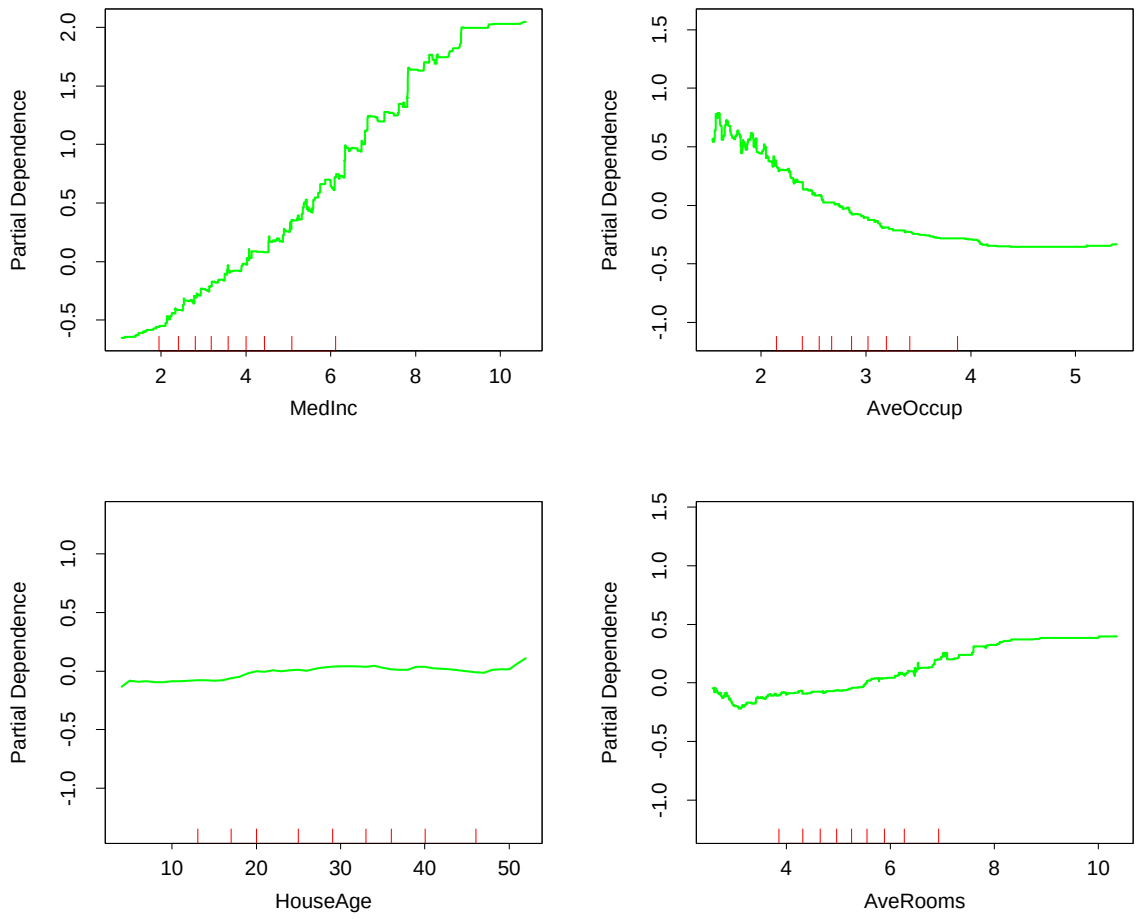


FIGURE 10.15. *Partial dependence of housing value on the nonlocation variables for the California housing data. The red ticks at the base of the plot are deciles of the input variables.*

Gráficos de Dependencia Parcial

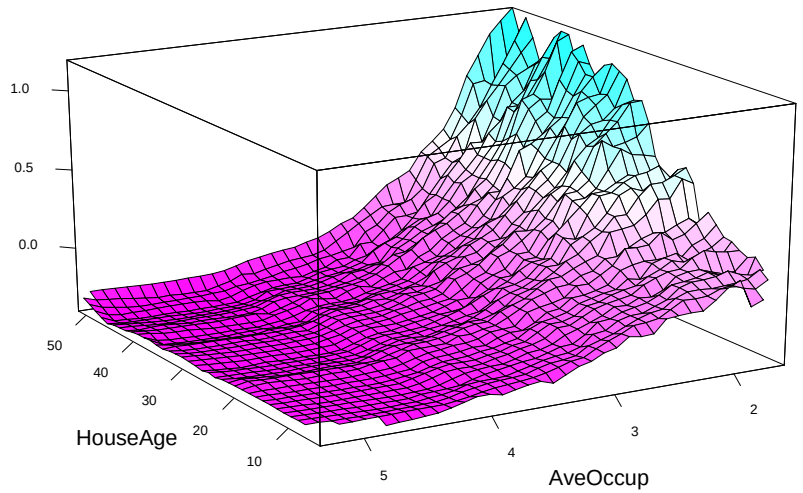


FIGURE 10.16. *Partial dependence of house value on median age and average occupancy. There appears to be a strong interaction effect between these two variables.*

Gráficos de Dependencia Parcial

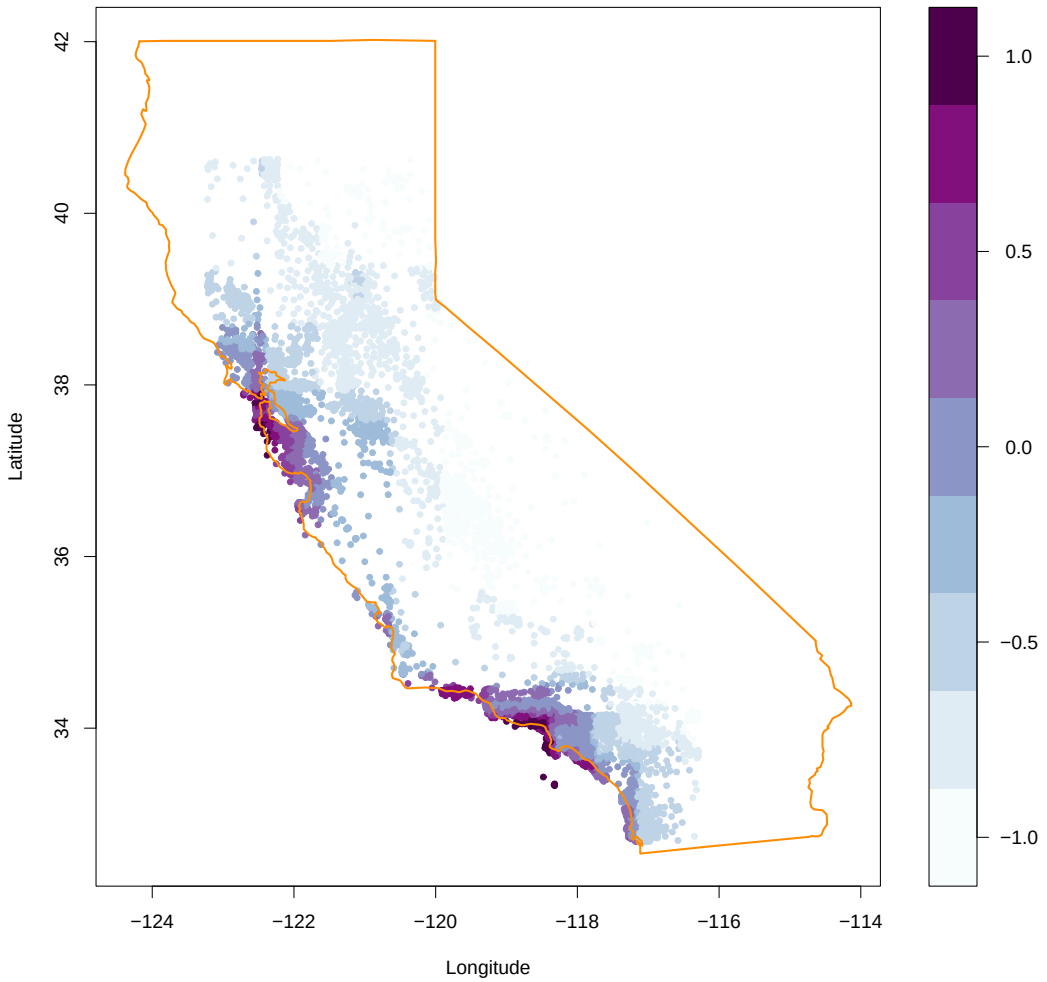
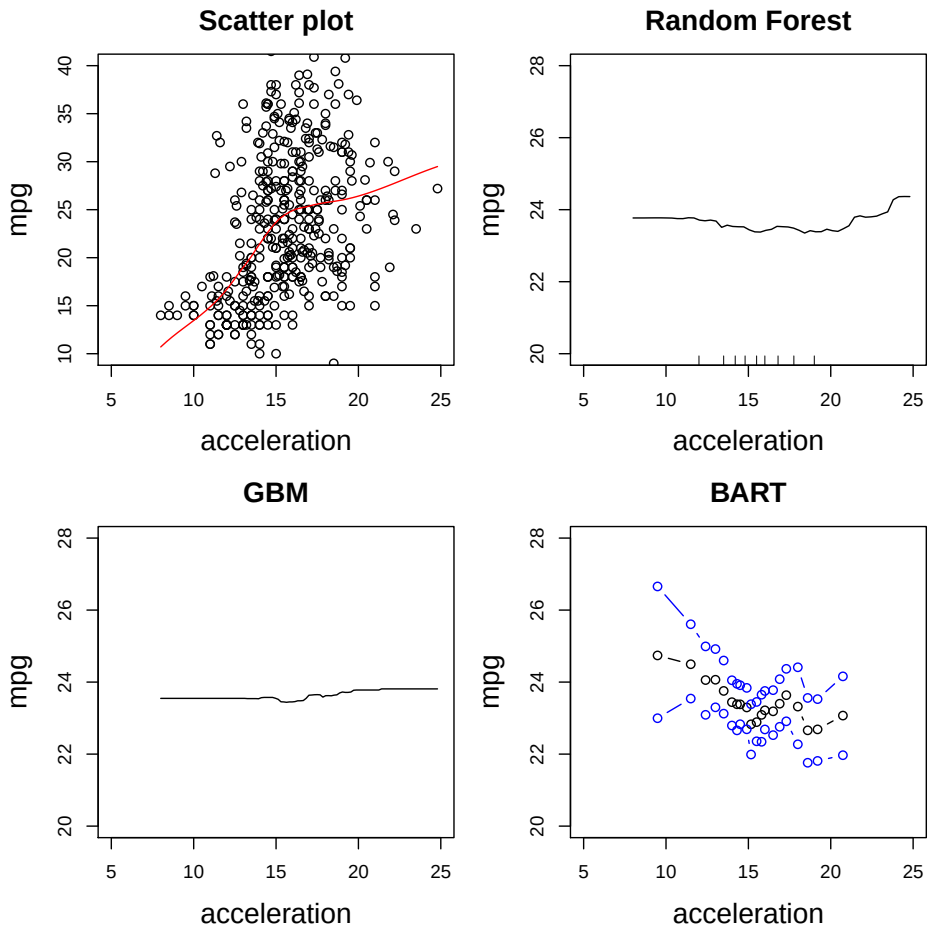


FIGURE 10.17. *Partial dependence of median house value on location in California. One unit is* $100,000, *at* 1990 *prices, and the values plotted are relative to the overall median of* $180,000.

Gráficos de Dependencia Parcial

CAUSAL INTERPRETATIONS OF BLACK-BOX MODELS



(A) Scatter plot and partial dependence plots using different black-box algorithms.

Contenido

Local Intepretable Model-Agnostic Explanations

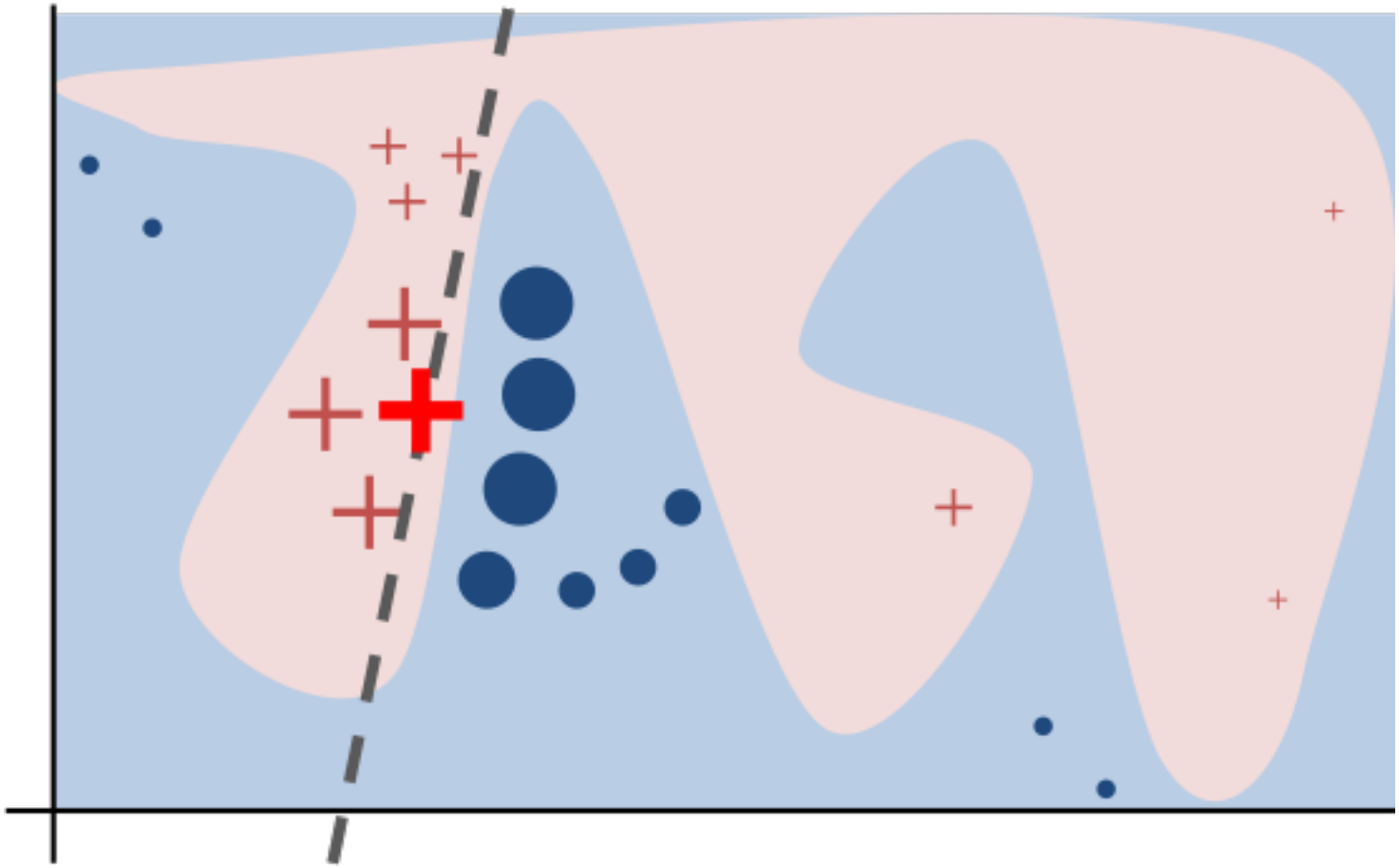
El problema que resuleve LIME es:

$$\max_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2}$$

donde:

- L es una función de pérdida.
- $f : \mathscr{X} \to \mathscr{Y}$ es la función de aprendizaje que queremos intepretar (localmente).
- $g : \mathscr{X}' \to \mathscr{Y}$ es una función que interpreta.
- $\mathscr{X}'$ es un espacio de variables más fácil de interpretar que $\mathscr{X}$.
- $\pi(x)$ es una función que le da peso a muestras cerca a x'.
- $\Omega(g)$ es una medida de la complejidad de g.
- G es un espacio de funciones intepretables.

Local Intepretable Model-Agnostic Explanations

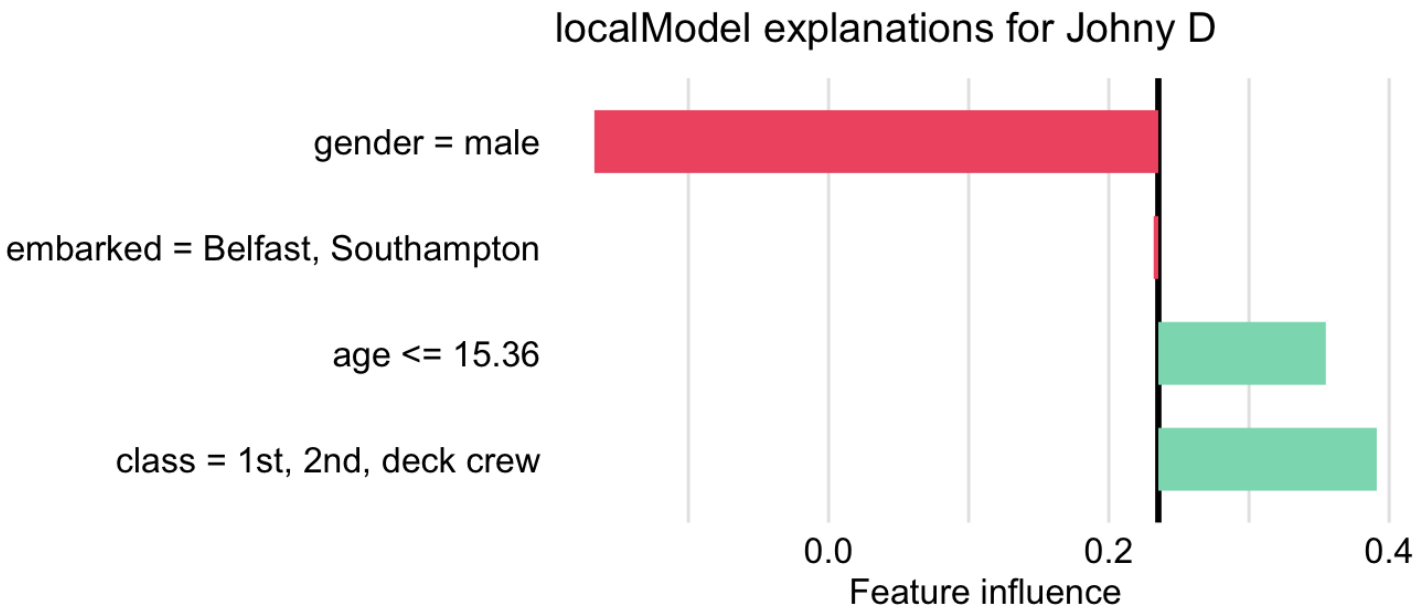


Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

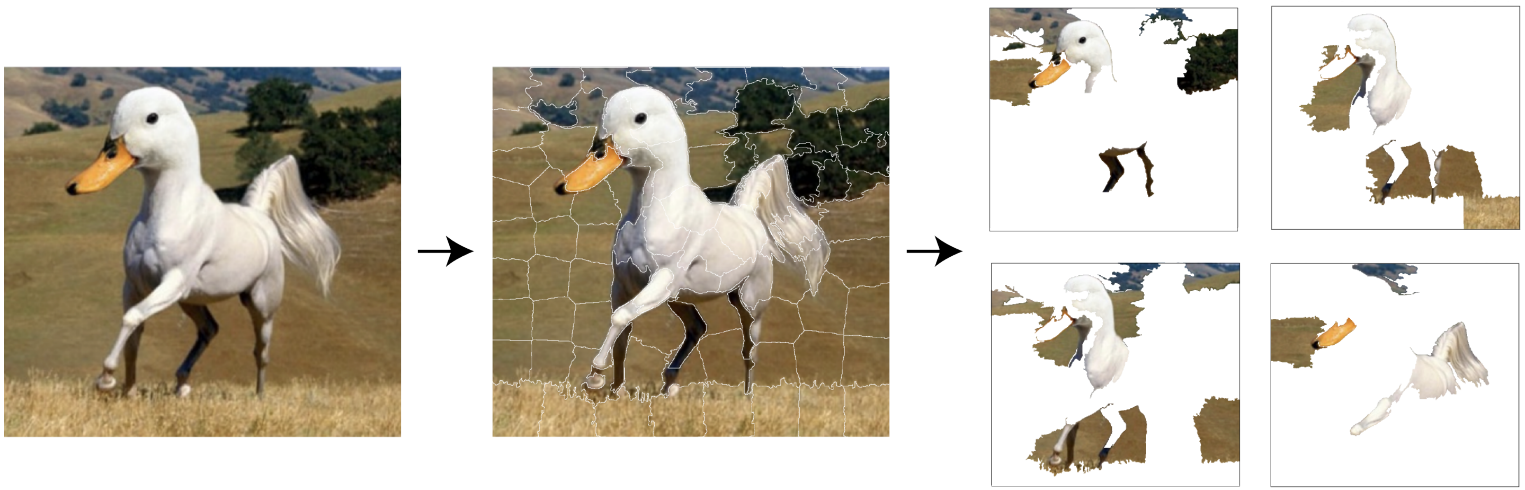
$\mathscr{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
$z_i' \leftarrow sample\_around(x')$
$\mathscr{Z} \leftarrow \mathscr{Z} \cup \langle z_i', f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathscr{Z}, K)$ $\triangleright$ with z_i' as features, $f(z)$ as target Return w

Ejemplos: Probabilidad de sobrevivir Titanic



Ejemplos: Clasificación Imágenes VGG16

Ejemplos: Clasificación Imágenes VGG16



Ejemplos: Clasificación Imágenes VGG16

Label: standard poodle
Probability: 0.18
Explanation Fit: 0.37

Label: goose
Probability: 0.15
Explanation Fit: 0.55

Contenido

Juegos coalicionales

- Partimos de un conjunto de jugadores y estudiamos los incentivos a formar subgrupos.
- Los acuerdos entre las posibles coaliciones son de obligatorio cumplimiento (de la misma forma que en la solución de Nash al problema de negociación, en este sentido, toda asignación de utilidades es un equilibrio).
- Luego la pregunta fundamental es, qué coaliciones se formarían que respeten los acuerdos y cómo se reparten el valor obtenido por la coalición.

Juegos coalicionales: Utilidad transferible

- Los juegos de utilidad transferible corresponden a situaciones donde los agentes pueden transferirse la utilidad de lo que puedan recibir de participar en una coalición (e.g., valores monetarios, un bien divisible, etc.).
- Utilidades no tranferibles son por ejemplo la reputación, influencia política, etc.
- Un juego coalicional es (N, v) donde N es un conjunto finito de jugadores y $v : P(N) \to R$ tal que $v(\emptyset) = 0$. La función v se conoce como la función coalicional o **función carcaterística** del juego y sus valores representan lo que una coalición podria obtener (o esperar obtener) en caso de formarse.

Juegos coalicionales: Utilidad transferible

- Los juegos de utilidad transferible corresponden a situaciones donde los agentes pueden transferirse la utilidad de lo que puedan recibir de participar en una coalición (e.g., valores monetarios, un bien divisible, etc.).
- Utilidades no tranferibles son por ejemplo la reputación, influencia política, etc.
- Un juego coalicional es (N, v) donde N es un conjunto finito de jugadores y $v : P(N) \to R$ tal que $v(\emptyset) = 0$. La función v se conoce como la función coalicional o **función carcaterística** del juego y sus valores representan lo que una coalición podria obtener (o esperar obtener) en caso de formarse.

Juegos coalicionales: Utilidad transferible

- Los juegos de utilidad transferible corresponden a situaciones donde los agentes pueden transferirse la utilidad de lo que puedan recibir de participar en una coalición (e.g., valores monetarios, un bien divisible, etc.).
- Utilidades no tranferibles son por ejemplo la reputación, influencia política, etc.
- Un juego coalicional es (N, v) donde N es un conjunto finito de jugadores y $v : P(N) \to R$ tal que $v(\emptyset) = 0$. La función v se conoce como la función coalicional o **función carcaterística** del juego y sus valores representan lo que una coalición podria obtener (o esperar obtener) en caso de formarse.

Juegos coalicionales

- Una hipótesis fundamental es que la utilidad que podría obtener una coalición no depende del resto de los jugadores que no son miembros (i.e., el modelo más general se conoce como juegos en forma particional).

Juegos coalicionales: Ejemplos

- El consejo de seguridad de las naciones unidas tiene 15 miembros y $v(S) = 1$ si y solo si $|S| \geq 9$ y además, todos los miembros permanentes están en S (los miembros permanentes son: Estados Unidos, Francia, Inglaterra, China y Rusia).

Conceptos de solución: Preguntas fundamentales

- Que coaliciones se forman?
- Cómo se dividen la utilidad alcanzada cuando forman la coalición.
- Qué coalición y cómo dividir la utilidad alcanzada recomendaría un planificador.
- Nos vamos a concentrar en situaciones en las que se forma la coalición completa y trataremos de responder a la segunda pregunta.

Conceptos de solución: Preguntas fundamentales

- Que coaliciones se forman?
- Cómo se dividen la utilidad alcanzada cuando forman la coalición.
- Qué coalición y cómo dividir la utilidad alcanzada recomendaría un planificador.
- Nos vamos a concentrar en situaciones en las que se forma la coalición completa y trataremos de responder a la segunda pregunta.

Conceptos de solución: Preguntas fundamentales

- Que coaliciones se forman?
- Cómo se dividen la utilidad alcanzada cuando forman la coalición.
- Qué coalición y cómo dividir la utilidad alcanzada recomendaría un planificador.
- Nos vamos a concentrar en situaciones en las que se forma la coalición completa y trataremos de responder a la segunda pregunta.

Conceptos de solución: Preguntas fundamentales

- Que coaliciones se forman?
- Cómo se dividen la utilidad alcanzada cuando forman la coalición.
- Qué coalición y cómo dividir la utilidad alcanzada recomendaría un planificador.
- Nos vamos a concentrar en situaciones en las que se forma la coalición completa y trataremos de responder a la segunda pregunta.

Conceptos de solución

- Sea $T(N)$ un conjunto de juegos coalicionales sobre un conjunto de jugadores N (denotamos por N el conjunto y n el número).
- Un concepto de solución (puntual) de un juego es una función $\phi : T(N) \to R^n$.
- Si para todo $v, \sum_{i=1}^{n} \phi_i(v) = v(N)$, decimos que ϕ es una imputación (eficiente).

Valor de Shapley

- El valor de Shapley, $\phi^{SV} : T(N) \to R^n$ es un concepto de solución (puntual) basado en la contribución marginal de cada jugador al valor de cada coalición.

Valor de Shapley

- Sea S y $N \setminus S \setminus i$ dos coaliciones ordenadas que no contiene a i (S no contiene a i): $(v(S \cup i) - v(S))$ es la contribución de i a la coalición.
- Existen $(\mid S \mid !)(N - \mid S \mid -1)!$ formas de elegir las dos coaliciones inciales y al normalizar por $N!$ se normaliza por el número total de formas de ordenar i unido con S y $N \setminus S \setminus i$.

$$\phi_i^{SV}(v) = \sum_{S \subset N \setminus i} (v(S \cup i) - v(S)) \frac{(\mid S \mid !)(N - \mid S \mid -1)!}{N!} \quad (3)$$

- Obérvese que $\sum_{S \subset N \setminus i} \frac{(|S|!)(N-|S|-1)!}{N!} = \frac{(N-1)!}{N!}$, y sumando sobre i se obtiene 1.

Ejemplo: Juego de voto mayoritario con un veto

Example

Supongamos que $N = 3$ y el jugador 1 tiene poder de veto. Es decir: $v(S) = 1$ si $|\, S \,| = 2$ y $1 \in S$. Consideremos el jugador 1. Existen 3 posibles coaliciones que no incluyen al jugador 1: $\{2\}, \{3\}, \{2,3\}$. Su contribución marginal es 1 en todas las coaliciones. Entonces:

$$\phi_1^{SV}(v) = \frac{1+1+2}{6} = \frac{2}{3} \tag{4}$$

Un análisis similar muestra que el valor de Shapley de todos en el juego es: $(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$

Contenido

Modelos lineales de atribución

Definition

El modelo explicativo es una función lineal de de variables binarias:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{5}$$

donde $z' \in \{0,1\}^M$, M es el número de características explicativas y $\phi_i \in \mathbb{R}$.

Theorem

La única g que satisface aproximación local, missingness y consitencia es:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{6}$$

donde $|z'|$ es el número de entradas z',y $z' \subseteq x'$ re;resenta todos los z' vectores subconjuntos de los vectores x' con entradas diferentes de cero.

SHAP: SHapley Additive exPlanation

- Dado que $f_x(z')$, donde z' es un subconjunto de variables de x' no siempre se puede evaluar, corresponde a evaluar la función con datos faltantes, SHAP propone:

$$f_x(z') = E[f(z) \mid z_S] \tag{7}$$

  donde z_S son los indices que no son cero de z.
- En la demostración de aproximación local se muestra que $\phi_0 = f_x(0) = E[f(z)]$.

SHAP: SHapley Additive exPlanation

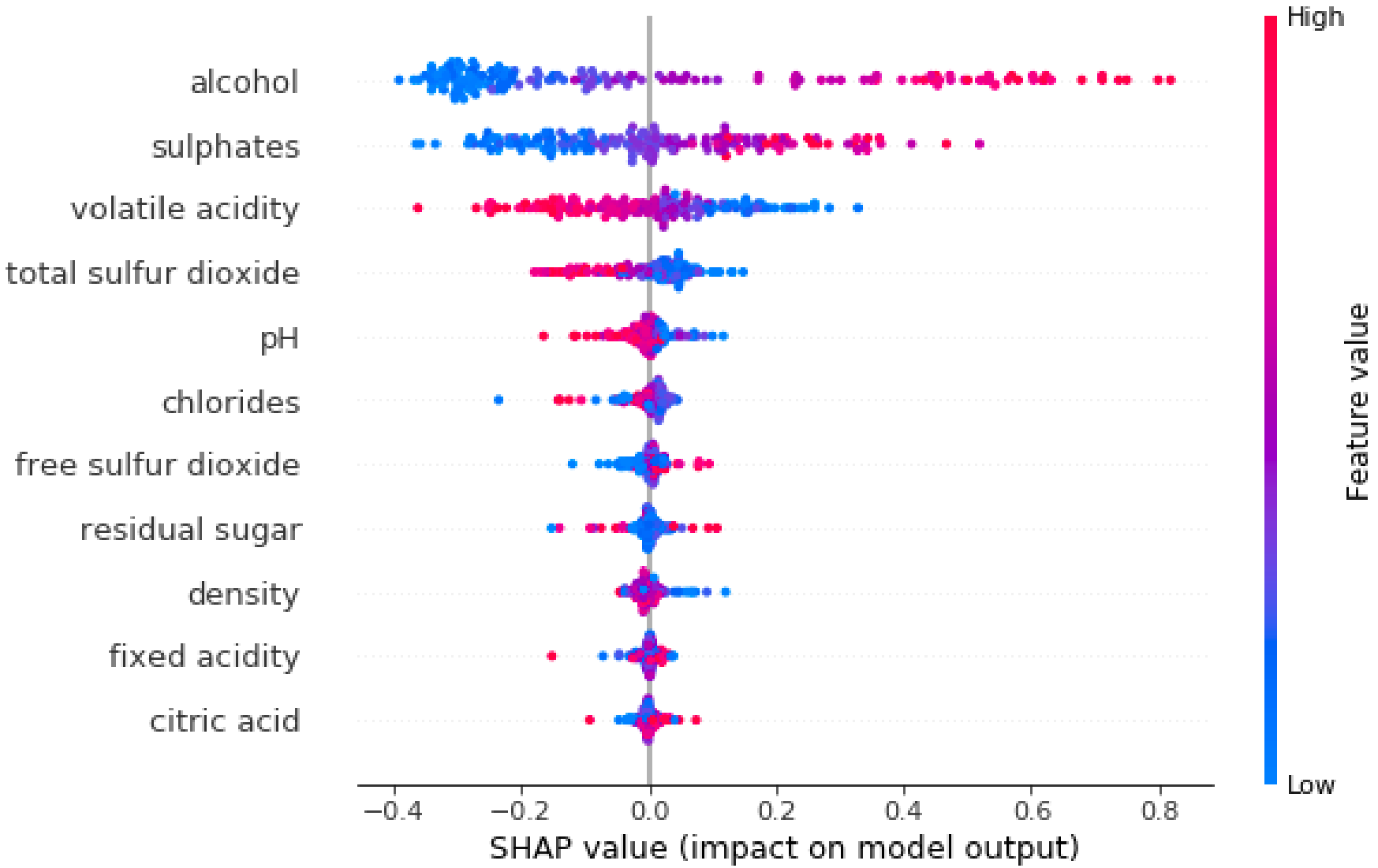


Figure: Resultado: Rating de calidad de variedaded de vino

Valores de SHAP: CGANS predicción de crimen

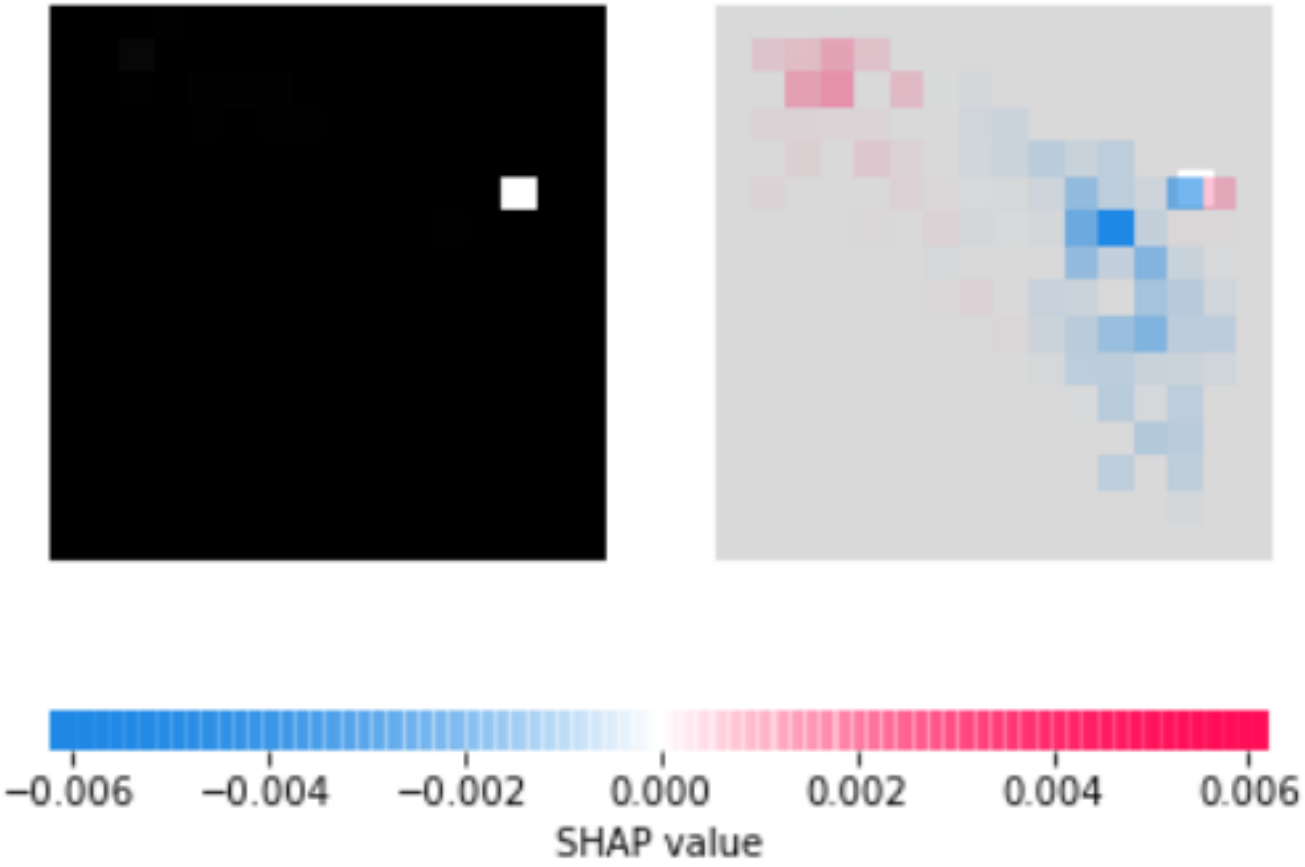


Figure: SHAP values (right) for a crime image (left)