Aprendizaje Estadístico Supervisado

Natalia da Silva

2024

Aprendizaje estadístico supervisado

Problema supervisado:

$$\mathrm{Y} = \mathrm{f}(\mathrm{X}) + \epsilon$$

donde Y: la variable de interés. X: variables para explicar el problema. ϵ: aspectos no explicados.

Métodos para aprender $\mathrm{f}()$

- Obtener valores futuros de Y
- Entender los efectos de X sobre Y

Esquema

- Interpretabilidad en aprendizaje estadístico
- Gráfico de dependencia parcial (PDP)
- Extensiones del PDP (ICE y ALE)

Clase basada en: Interpretable Machine Learning

Interpretabilidad

- No hay una definición matemática de interpretabilidad
- En términos generales la interpretabilidad es el grado en que un humano puede entender de la causa de la decisión o predicción de un método estadístico.

Importancia de la Interpretabilidad

- Muchas veces en los modelos predictivos no nos interesa del porqué de la decisión
- Otras veces saber el porqué de la decisión nos ayuda a entender más del problema, los datos y las razones de porqué el modelo puede fallar.

Aún si el objetivo es puramente predictivo:

- Interpretabilidad en los ML es útil para detectar sesgos
- Un modelo interpretable puede ser auditado por eso también la importancia
- Posible mejorar el ajuste

Muchos métodos para interpretar resultados y muchas formas de clasificarlos …

Taxonomía de los métodos de Interpretabilidad

Los métodos interpretables pueden ser resumidos de acuerdo a su resultado:

- Estadísticas de resumen para las variables explicativas (ej medida de importancia)
- Visualización de resumen para las variables explicativas, algunos de los anteriores tb se visualizan pero otros resumenes solo tienen sentido visualizado (ej PDP)
- Carácteristicas intrínsecas al modelo como ser la estructura de un árbol o los pesos en un modelo lineal.
- Datos, en esta categoría entran los modelos que generan datos para hacer el modelo interpretable (explicación contrafactual)
- Modelos instrinsecamente interpretables, una solución para interpretar modelos de caja negra es aproximarlos con un modelo interpretable.

A su vez hay herramientas que son específicas para un modelo y otras que son para todos (model- specific, model-agnostic)

Taxonomía de los métodos de Interpretabilidad

Intrínsecos o post hoc

Distingue si el método logra la interpretabilidad restringiendo la complejidad del modelo o aplicando métodos que analizan el modelo luego de entrenarlo

1. **Intrínseco, Model-specific:** Modelos que son interpretables debido a características propias del modelo. Por ejemplos, los coeficientes de un modelo de regresión, también dada la estructura sencilla los árboles entran en este grupo. Este tipo de interpretabilidad puede dificultar la comparación entre modelos.
2. **Post hoc, Model-agnostic :** Aplicación de métodos interpretables luego que el modelo fue entrenado. Ejemplos: importancia permutada de las variables, gráfico dependencia parcial.

De acuerdo al tipo de efectos

Interpretabilidad: describir la relación entre la respuesta y los predictores (en conjunto y por separado).

- Impacto de $\mathbf{X_s}$ en el modelo estimado $\hat{\mathbf{f}}()$
- Importancia de $\mathbf{x_s}$ en el poder predictivo del modelo
- Efecto causal de $\mathbf{x_s}$ sobre $\mathbf{Y}$

Alcance de la interpretabilidad

- Transparencia del método, ¿cómo el algoritmo crea el modelo?
- Interpretación global del modelo, ¿cómo el modelo entrenado hace predicciones?
- Interpretación global del modelo en un nivel modular, ¿cómo las partes del modelo afectan la predicción?
- Interpretación local para una sola predicción, ¿porqué el modelo hace ciertas predicciones para una observación particular?
- Interpretación local para un grupo de predicciones, ¿porqué el modelo hace predicciones específicas para un grupo de observaciones?

Evaluación de la interpretabilidad

No hay un consenso sobre la interpretabilidad en ML y tampoco es claro como medirla pero hay alguna investigación sobre como hacerlo.

Algunas ideas:

- Evaluación a nivel de la aplicación (tarea real): poner la explicación en el producto y testeado por el usuario final. Comparar el modelo con la decisión que tomaría un humano (ej: rayos X)
- Evaluación a nivel de usuario (tarea simple): es una simplificación del anterior, experimentos no llevados adelante por expertos
- Evaluación a nivel de la función….

Propiedades de la explicación

Queremos explicar las predicciones de ML, se recae en métodos explicativos que son algoritmos que generan explicaciones.

Se debe evaluar cuan buena es una explicación o el método para explicar

Propiedades de los métodos de explicación:

- Potencia expresiva: lenguaje o estructura de la explicación que el método es capaz de generar.
- Transparencia: describe cuanto el método de explicación recae en mirar el ML como sus parámetros.
- Portabilidad: describe el rango de ML para que el método explicable puede ser usado.
- Complejidad de algoritmo: describe la complejidad computacional de método que genera la explicación.

Propiedades de la explicación

Propiedades de explicaciones individuales:

- **Precisión:** Qué tan bien una explicación predice datos no observados.
- **Fidelidad:** Qué tan bien la explicación aproxima la predicción del modelo caja negra. Alta fidelidad es una de las propiedades más importantes de una explicación.
- **Consistencia:** Qué tanto difieren las explicaciones entre modelos que han sido entrenados con el mismo objetivo y tienen similares predicciones.
- **Estabilidad:** Cuan similares son las explicaciones para observaciones similares.
- **Comprensibilidad:** Qué tan bien se entiende la explicación por humanos
- **Certeza:** La explicación refleja la certeza del ML?

Propiedades de la explicación

- **Grado de importancia:** Como la explicación refleja la importante de las variables o parte de ella.
- **Novedad:** La explicación si una observación a ser explicada viene de una región nueva no contenida en los datos de entrenamiento.
- **Representatividad:** Cuantas observaciones cubre una explicación

4) Modelos interpretables

- Regresión lineal
- Regresión logística
- GLM, GAM y más
- Árboles de decisión
- ....
- Otros

5) Model agnostic

Principal ventaja de los métodos “Model agnostic” sobre los “Model specific” es su gran flexibilidad para ser usados en todos los ML.

Deseable para los métodos “Model agonostic”:

- Flexibilidad del modelo, el método funciona con cualquier ML
- Flexibilidad de la explicación, no limitado a cierta forma de la explicación
- Flexibilidad en la representación, el sistema de explicación debería ser capáz de usar distintas representaciones de las variables explicativas

Gráfico de dependencia parcial (PDP)

- Muestra el efecto marginal de una o dos variables explicativas en el valor predicho del ML.
- Muestra si la relación entre la respuesta y la variable explicativa es lineal, mónotona o más compleja. Cuando se aplica a un modelo de regresión lineal, el PDP siempre muestra una relación lineal.

Gráfico de dependencia parcial (PDP)

Separamos los predictores en dos grupos:

- X_S: la o las variables explicativas cuyo efecto sobre la respuesta queremos describir
- x_C son las otras variables explicativas utilizadas en el modelo

La función de dependencia parcial para regresión es:

$$f_S(x_S) = E_{X_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) dP(x_C)$$

- Marginalizando sobre x_C se obtienen una función que depende solamente de las variables en S e interacciones con otras variables incluídas.

Gráfico de dependencia parcial (PDP)

- La función de dependencia parcial $\hat{f}_{x_S}$ es estimada calculando el promedio en los datos de entrenamiento (Monte Carlo)

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

- $x_C^{(i)}$ son los valores de las variables que no estamos interesados en el conjunto de datos.
- La función nos dice que para un valor determinado en las variables en S cuál es el efecto marginal promedio en las predicciones.

Gráfico de dependencia parcial (PDP)

- Un supuesto en PDP es que las variables explicativas en C no están correlacionadas con las variable en S.
- Si este supuesto es violado el promedio calculado para el PDP incluirá puntos que son muy improbables o inluso imposibles.
- Para clasificación el PDP presenta la probabilidad para cierta clase dada diferentes valores de las variables en S. Para muchas clases se puede dibujar una linea o gráficos por clase.
- Para predictoras categóricas, para cada categoría se obtiene el PDP estimado forzando todos los datos a la misma categoría.

PDP pasos

1. Selecciono una o dos variables de interés x_S
2. Definimos una grilla para x_S
3. Para cada valor de la grilla: remplazo la variable de interés con el valor de la grilla y promedio las predicciones.
4. Dibujo la curva

Ejemplo PDP

Datos contienen conteos diarios de alquileres de bicicletas en Washington DC y contienen variables climáticas y de estación

El objetivo es predecir cuantas bicicletas se alquilaran dependiendo del clima y el día.

Ejemplo PDP

Variables:

- Conteo de bicicleta alquiladas.
- Estaciones del año
- Indicadoa si el día fue feriado o no
- Año, 2011 o 2012
- N´Umero de días desde 01.01.2011 arrancan los registro
- Temperatura en Celsius
- Humedad
- Velocidd del viento
- Entre otras

Se ajusta un RF para predecir el número de bicicletas alquiladas en un día y se estima el PDP.

Ejemplo PDP

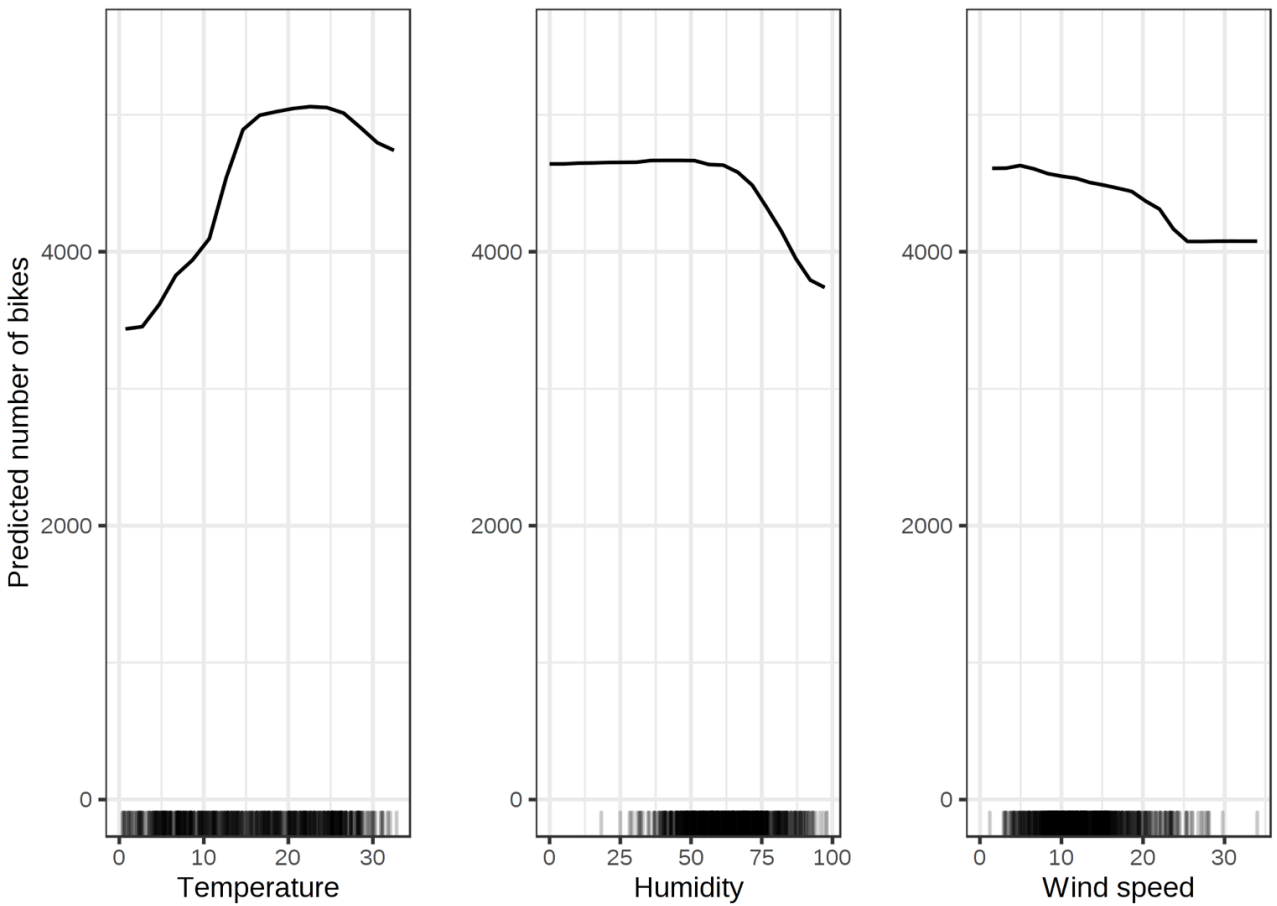


FIGURE 5.2: PDPs for the bicycle count prediction model and temperature, humidity and wind speed. The largest differences can be seen in the temperature. The hotter, the more bikes are rented. This trend goes up to 20 degrees Celsius, then flattens and drops slightly at 30. Marks on the x-axis indicate the data distribution.

For warm but not too hot weather the model predicts on average a high number of rented bicycles

Ejemplo PDP

- Para días calurosos pero no muy calurosos, el modelo predice en promedio un número elevado de bicicletas alquiladas.
- Los ciclistas potenciales son cada vez más inhibido en el alquiler de una bicicleta cuando la humedad supera el 60%.
- A más viento menor cantidad de gente alquila bicicletas. El número predicho de bicicletas alquiladas no cae cuando la velocidad del viento va de 25 to 35 km/h. No hay muchos datos de entrenemiento ahí por lo que puede ser que el modelo no de predicciones con sentido en ese rango.

Ejemplo PDP

PDP para predictora categórica, miramos el efecto de la estación en la predicción del alquiler de bicicletas

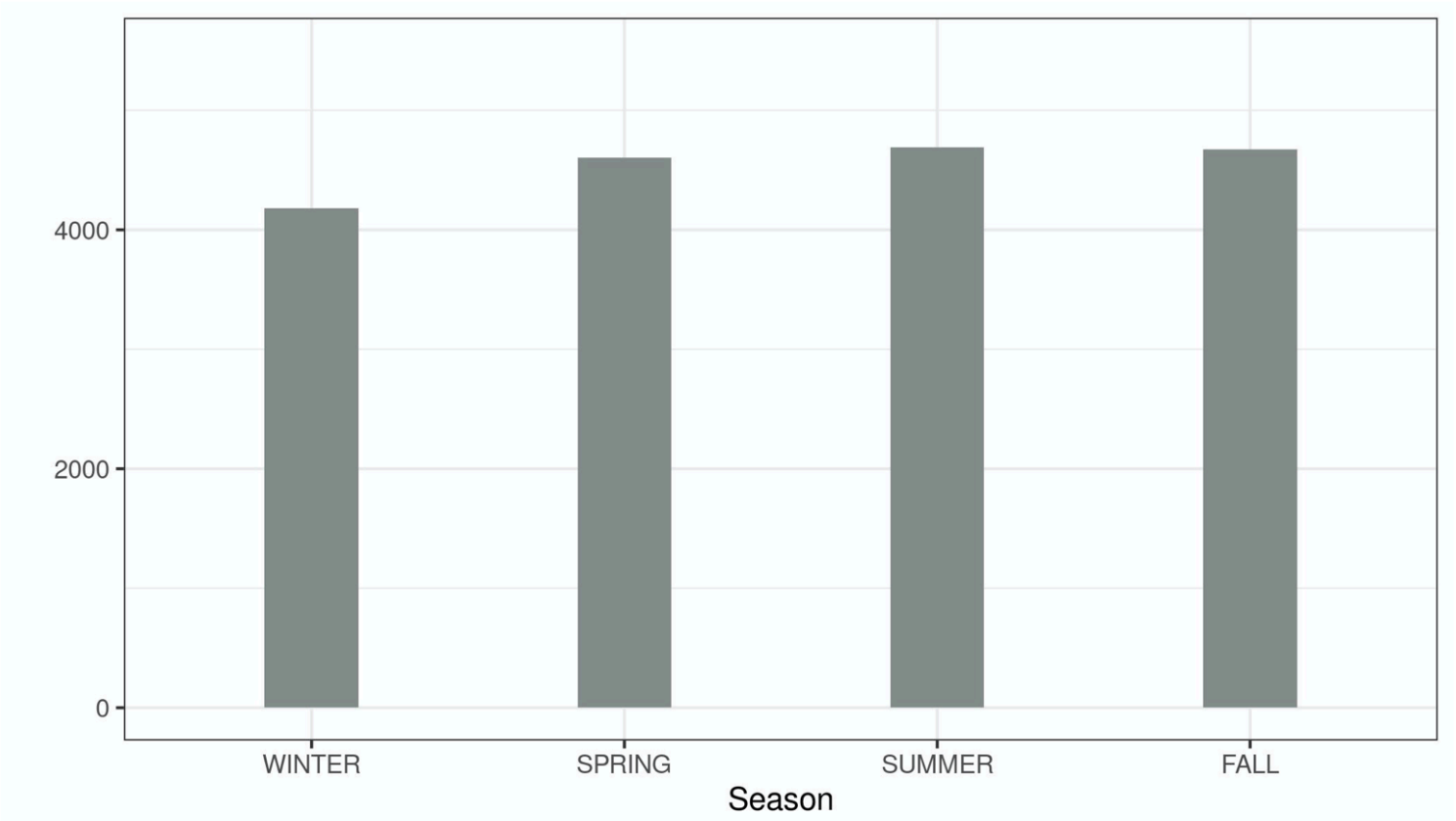


FIGURE 8.2: PDPs for the bike count prediction model and the season. Unexpectedly all seasons show similar effect on the model predictions, only for winter the model predicts fewer bicycle rentals.

Ventajas PDP

- El cálculo es intuitivo
- Si no hay correlación entre x_C y x_S el PDP perfectamente captura como x_S influencia la predicción en promedio.
- Sencillo de implementar
- Interpretación causal en el modelo

Desventaja

- El máximo número de variables en un PDP con sentido es 2.
- Algunos PDP no muestran la distribución de $\mathbf{x}_C$ en los datos, problema porque puedo sobre interpretar los resultados en lugares donde no observó datos o muy pocos.
- El supuesto de independencia es el principal problema en PDP, $\mathbf{x}_S$ no está correlacionada con otras $\mathbf{x}_C$
- Efectos de heterogenidad pueden estar ocultos porque los PDP solo muestran el efecto marginal promedio.

Esperanza condicional individual (ICE)

- ICE muestra una linea por observación, muestra cómo cambia la predicción cuando cambia una observación

Para cada observación en $\{(x_S^{(i)}, x_C^{(i)})\}_{i=1}^{N}$ la curva $f_S^{(i)}$ es dibujada contra $x_S^{(i)}$ mientras $x_C^{(i)}$ permanece constante.

- ICE permite visualizar la dependencia en la predicción de una variable para cada observación separadamente
- PDP es el promedio de las lineas del ICE.
- En el caso que hay interacción entre x_C y x_S es mejor que el PDP.

Ejemplo ICE

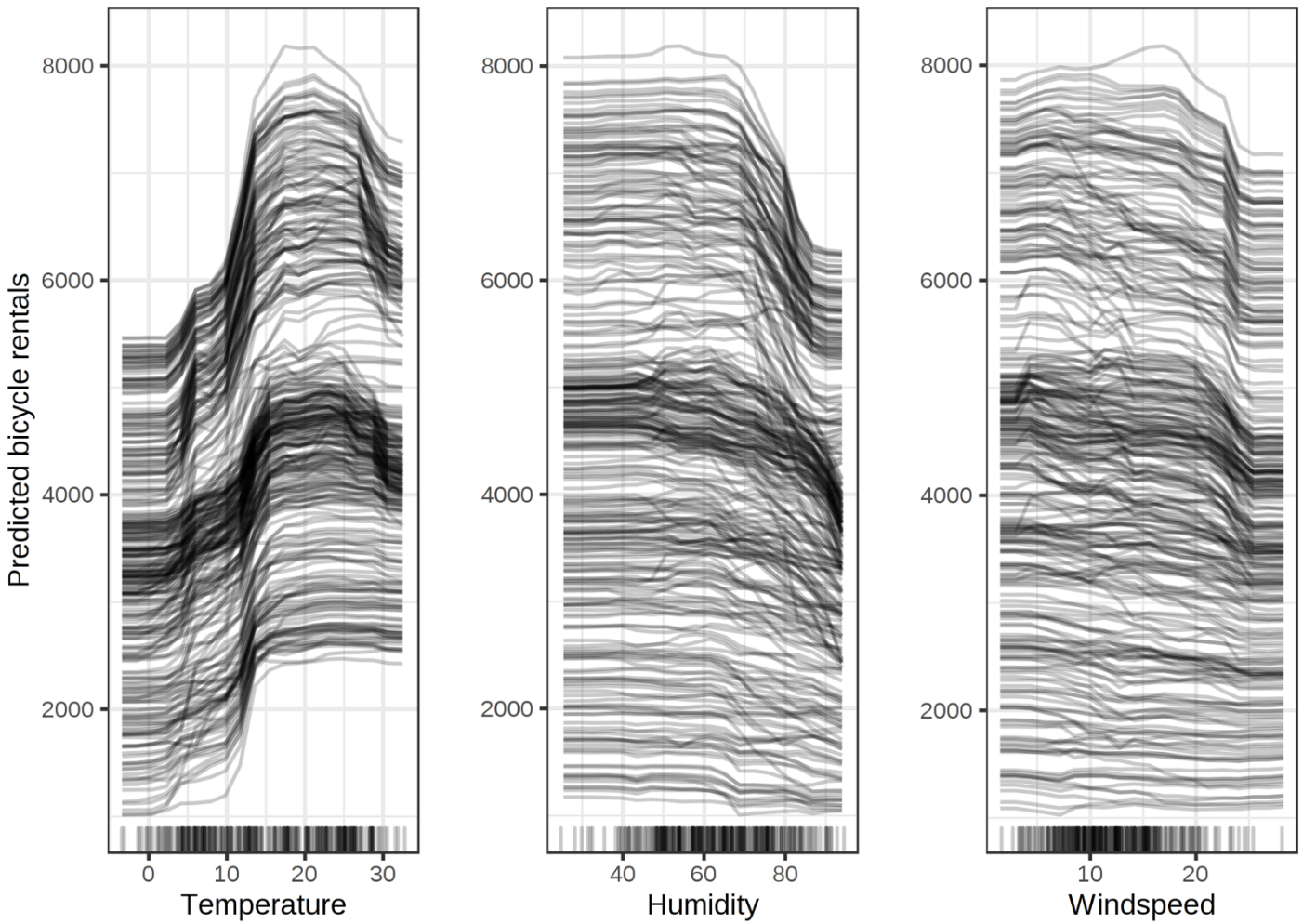


FIGURE 5.7: ICE plots of predicted bicycle rentals by weather conditions. The same effects can be observed as in the partial dependence plots.

ICE centrado

- Hay un problema con las gráficas ICE ya que muchas veces es difícil decir si dos curvas difieren porque comienzan en distintas predicciones
- Una solución es centrar las curvas a cierto punto en la variable explicativa y mostrar la diferencia en la predicción a este punto

c-ICE se define como:

$$f_{cent}^{(i)} = f^{(i)} - \mathbf{1}\hat{f}(x^a, x_C^{(i)})$$

Ejemplo ICE,centrado

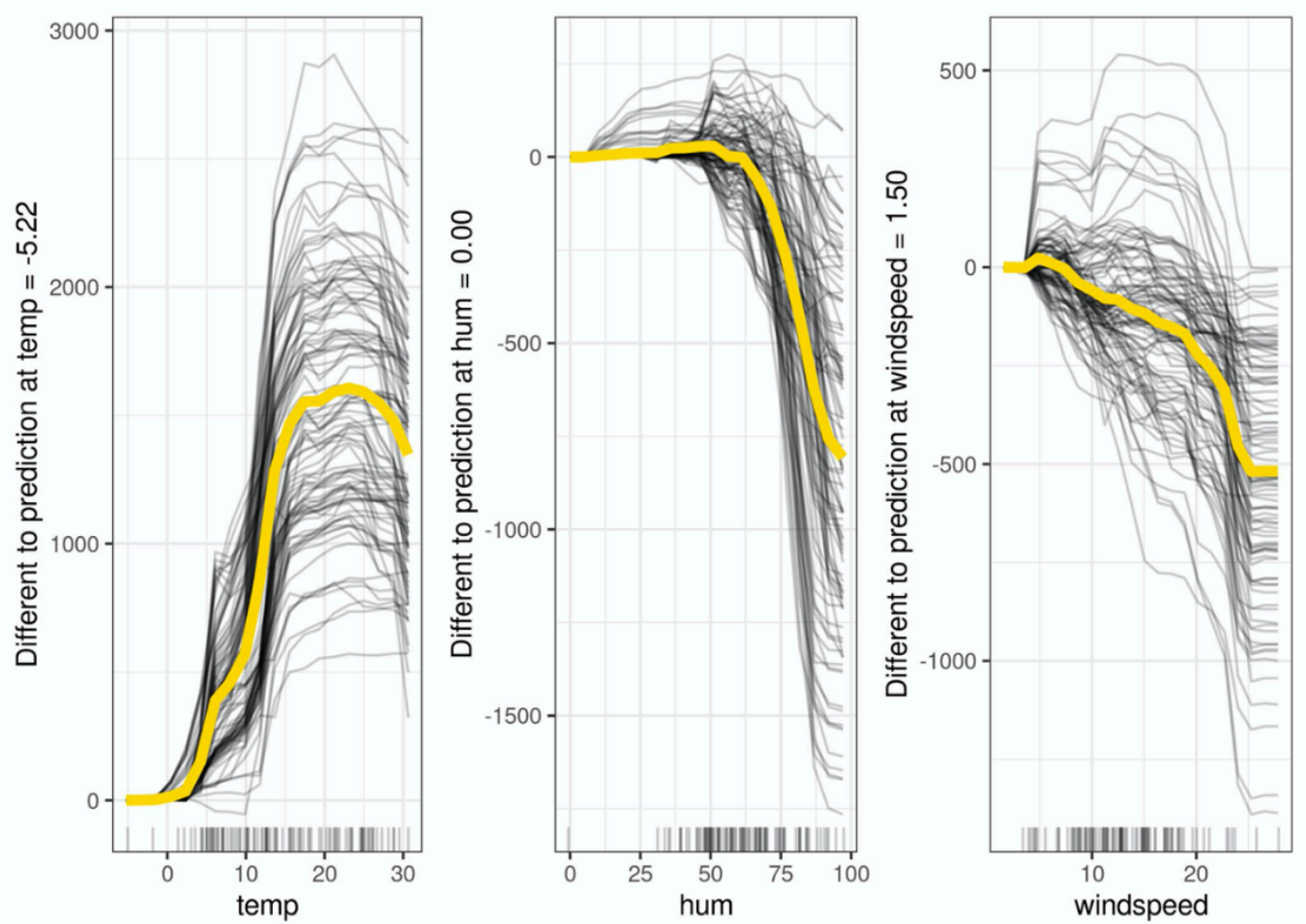


FIGURE 9.4: Centered ICE plots of predicted number of bikes by weather condition. The lines show the difference in prediction compared to the prediction with the respective feature value at its observed minimum.

ICE Ventajas-Desventajas

Ventaja:

1. Más intuitivos que PDP.
2. Puede descubrir relaciones heterogeneas

Desventajas:

1. Puede solamente mostrar una sola variable con sentido.
2. ICE tiene el mismo problema que PDP si la variable de interés está correlacionada con las otras algunos puntos en las lineas pueden ser puntos sin sentido.
3. Si hay muchas curvas puede ser muy confuso, se puede usar transparencias o dibujar una muestra de lineas

Efecto local acumulado (ALE)

- ALE describe como las variables explicativas influyen la predicción del ML en promedio.
- Los gráficos ALE son rápidos y una alternativa insesgada a PDP.
- Tiene el mismo objetivo que el PDP pero trata de resolver una de las debilidades del PDP que es cuando x_C y x_S están correlacionadas.

ALE, motivación

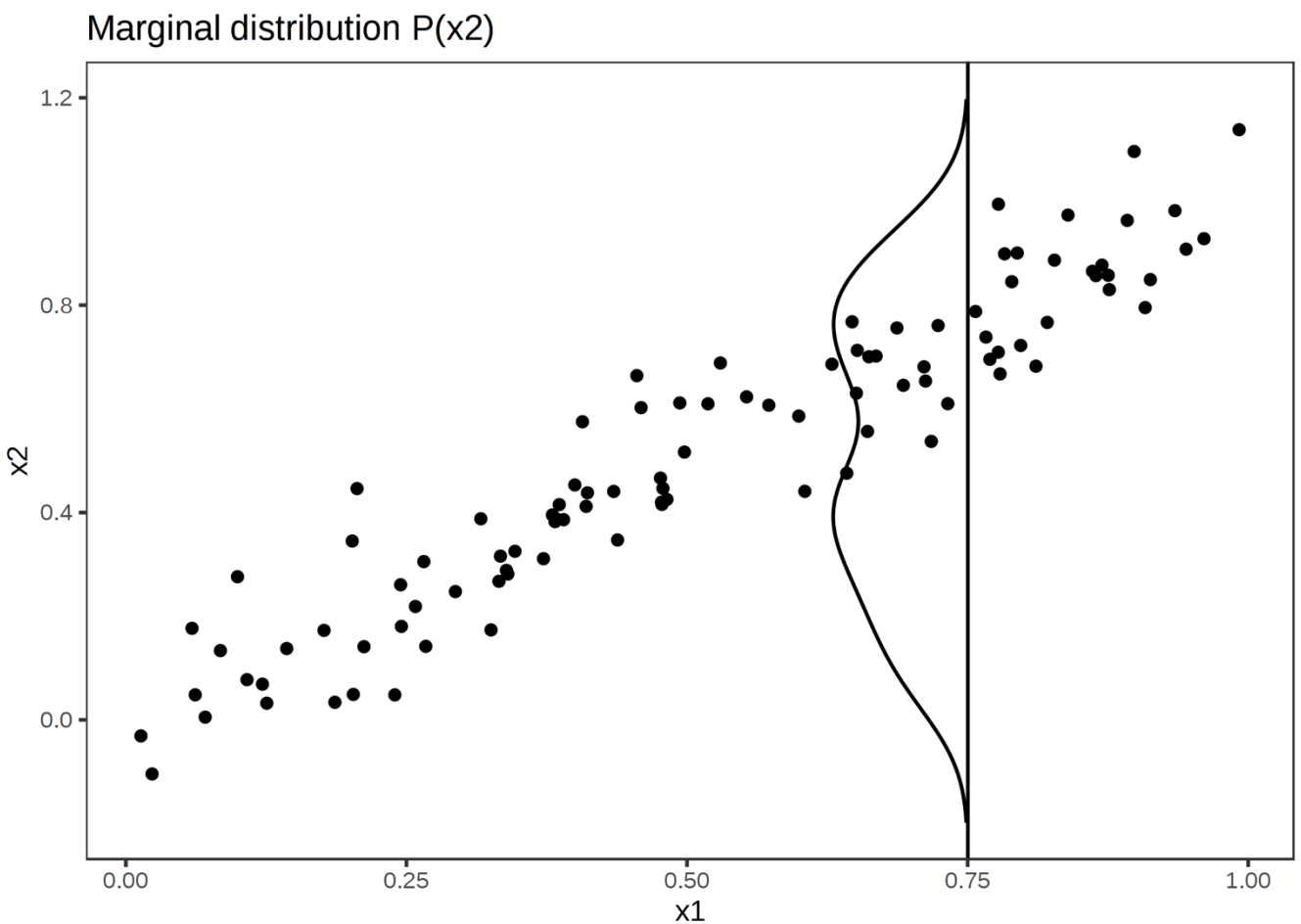
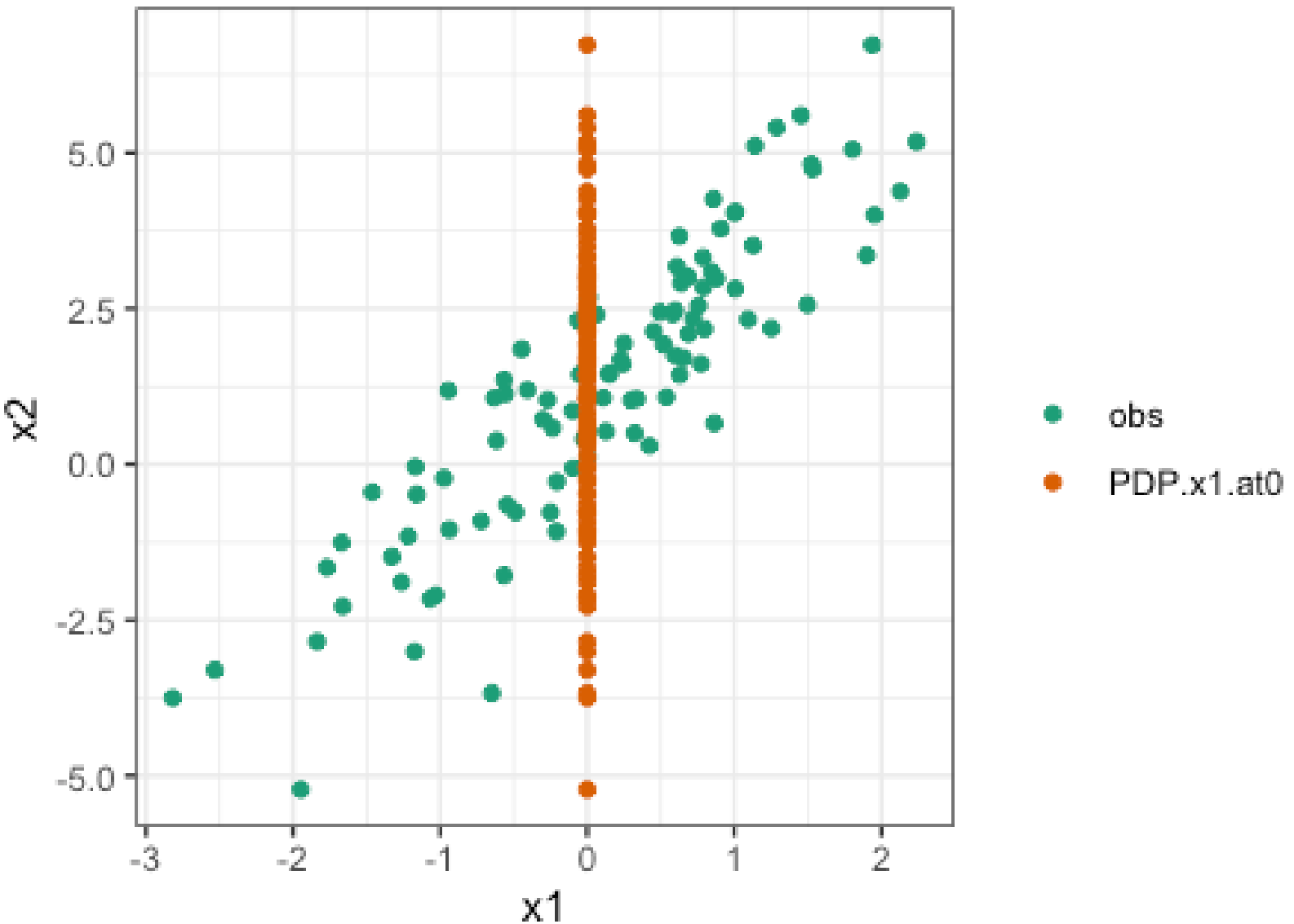


FIGURE 5.10: Strongly correlated features x1 and x2. To calculate the feature effect of x1 at 0.75, the PDP replaces x1 of all instances with 0.75, falsely assuming that the distribution of x2 at x1 = 0.75 is the same as the marginal distribution of x2 (vertical line). This results in unlikely combinations of x1 and x2 (e.g. x2=0.2 at x1=0.75), which the PDP uses for the calculation of the average effect.

ALE, motivación



ALE, motivación

¿Qué se puede hacer para obtener efecto estimado que respete la correlación entre la variables?

- Se podría promediar sobre la distribución condicional de la variable de interés (x_1).
- Esto es para un valor de la grilla de x_1 se promedian solo las predicciones de observaciones con valores similares de x_1 . Esta solución se llama M-plot (marginal plot).
- Los M-plot no soluciona el problema del PDP
- M-plots elimina del promedio de predicciones los casos poco probables, pero mezclan el efecto de las variables con el efecto de las variables correlacionadas.

Partial dependence plot

Muestra el efecto marginal de una o más variables sobre la variable de respuesta en un modelo de ML.

$$\hat{f}_{x_s}(x) = E[\hat{f}(x, X_c)] = \int \hat{f}(x, X_c) dP(X_c)$$

- x_s predictoras de interés
- X_c resto de las predictoras
- $\hat{f}()$: Modelo ajustado

Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232.

Partial dependence plot

Estimación: $\hat{f}_{x_s}(x) = \frac{1}{n} \sum_i \hat{f}(x, x_c^i)$

Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine.” Annals of statistics (2001): 1189-1232.

M-plot, ejemplo correlacionadas

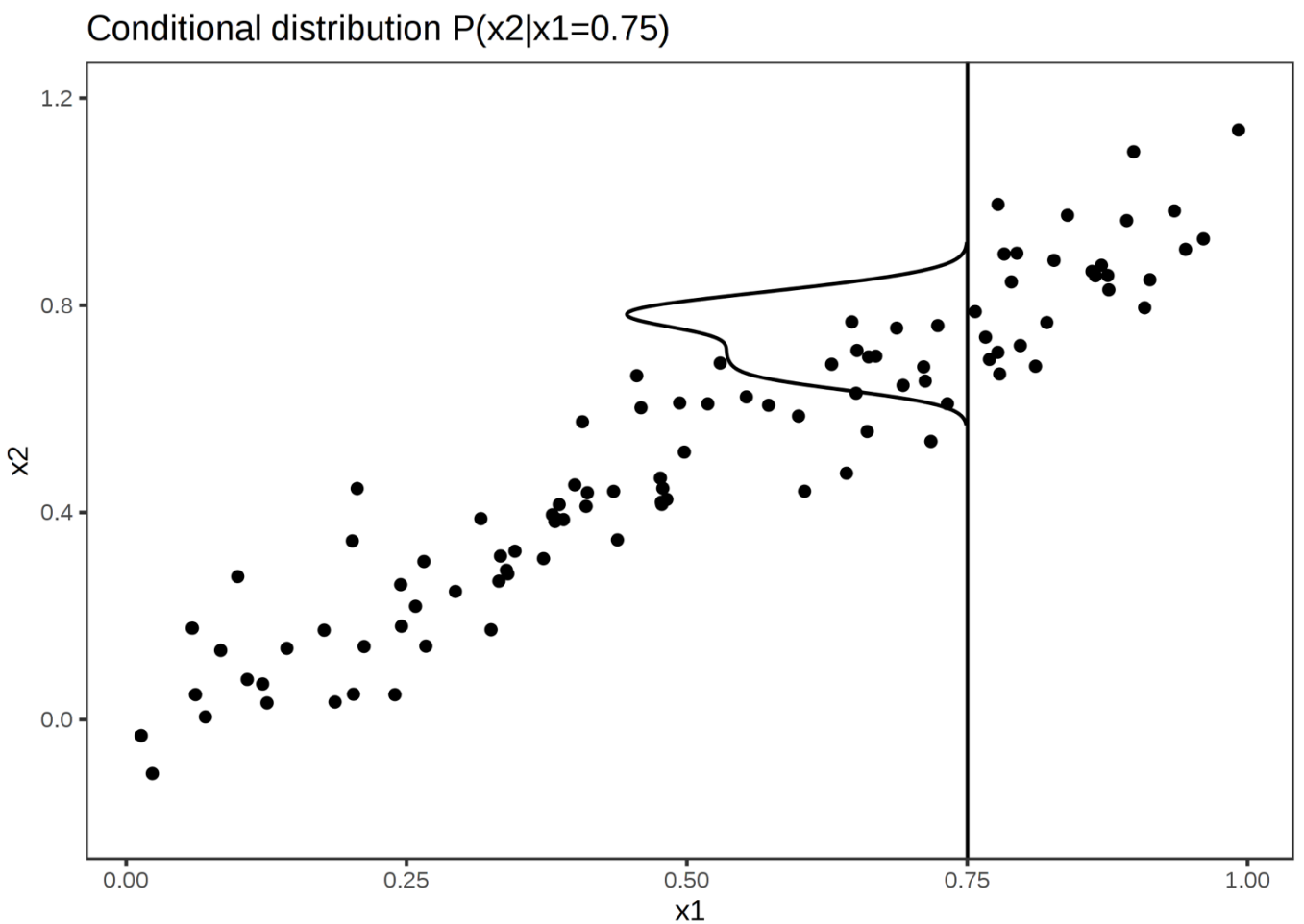


FIGURE 5.11: Strongly correlated features x1 and x2. M-Plots average over the conditional distribution. Here the conditional distribution of x2 at x1 = 0.75. Averaging the local predictions leads to mixing the effects of both features.

ALE plot

- ALE plots resuelven el problema, también basado en la distribución condicional de las variables, pero en vez de calcular el promedio de las predicciones usan la diferencia en las predicciones
- Entonces para el efecto de x_1 en un valor específico a el método ALE usa todas las observaciones con x_1 similares a a obtiene las predicciones del modelo asumiendo que $x_1 = a + 1$ restado a las predicciones del modelo asumiendo que $x_1 = a - 1$
- Esta forma de cálculo da el efecto puro de x_1 sin mezclar el efecto de las variables correlacionadas

ALE plot, intuición

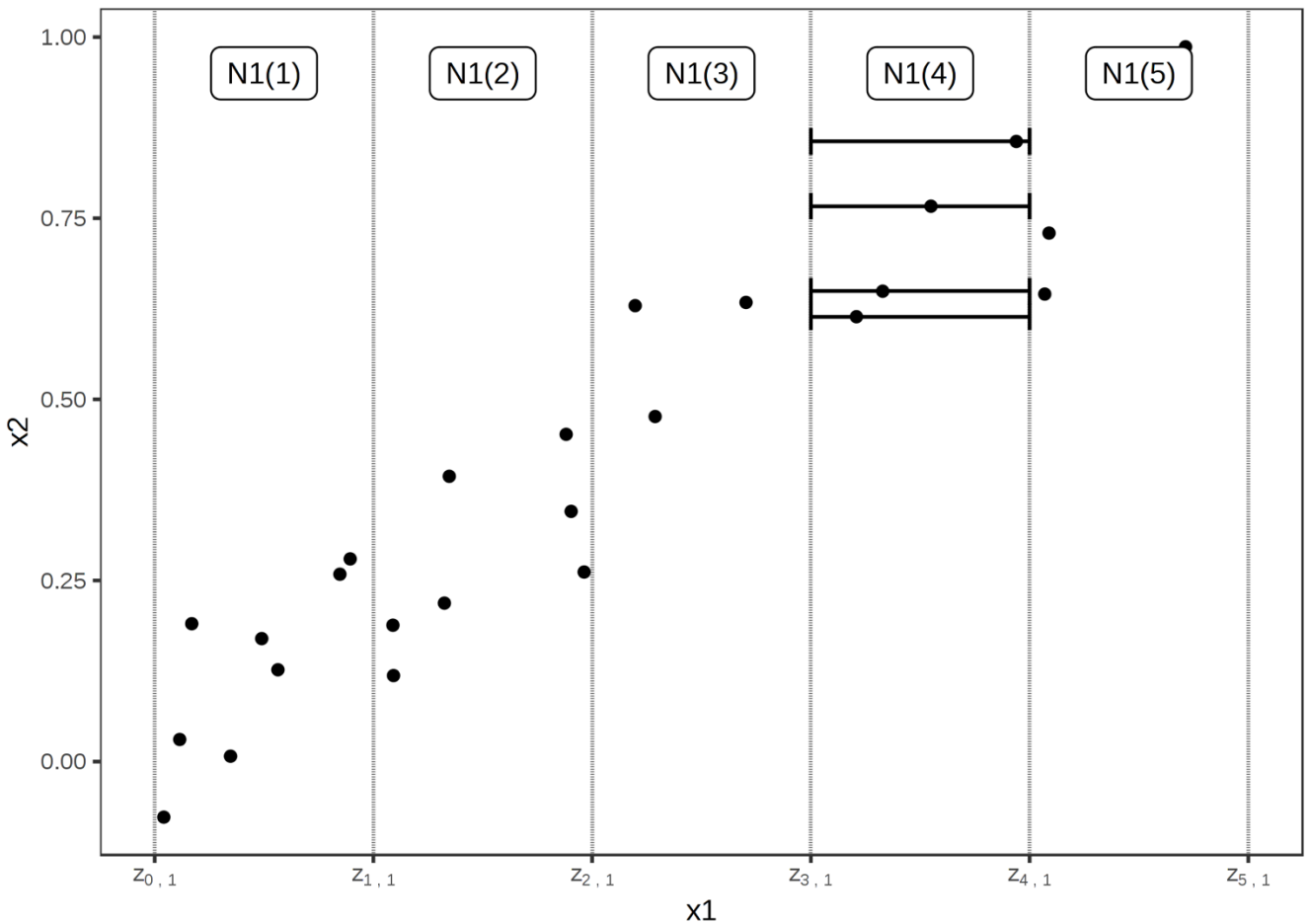


FIGURE 5.12: Calculation of ALE for feature x1, which is correlated with x2. First, we divide the feature into intervals (vertical lines). For the data instances (points) in an interval, we calculate the difference in the prediction when we replace the feature with the upper and lower limit of the interval (horizontal lines). These differences are later accumulated and centered, resulting in the ALE curve.

PDP,M-plot, ALE plot

Resumen como cada gráfico resume el efecto de una variable en un valor espec ífico de la grilla:

- **PDP:** muestra como el modelo predice en promedio cuando cada observación tiene el valor v en la variable de interés ignorando si este valor de grilla tiene o no sentido para todas las observaciones.
- **M-plots:** muestra como el modelo predice en promedio para cada observación que tiene valores cercanos a v para esa variable. El efecto podría ser por ese predictor pero también por otro correlacionado.
- **ALE-plots** muestra como el modelo predice el cambio en pequeñas ventanas de características entorno a un valor v para observaciones es esa ventana.

PDP,M-plot, ALE plot

- Los tres métodos reducen la función de predicción a una función que depende de una o dos variables de interés.
- Todos reducen la función promediando el efecto de otras variables explicativas pero difieren en si el promedio es de todas las predicciones o de la diferencia de las predicciones o si el promedio es sobre la distribución marginal o condicional.

PDP,ICE-plot, M-plot, ALE plot

- PDP, promedia las predicciones sobre la distribución marginal

$$\hat{f}_{x_S,PDP}(x_S) = E_{X_C}[\hat{f}(x_S, X_C)] = \int_{x_C} \hat{f}(x_S, x_C) P(x_C) dx_C$$

- ICE-plot, para cada observación en $\{(x_S^{(i)}, x_c^{(i)})\}_{i=1}^{N}$ la curva $f_S^{(i)}$ se dibujada para cada $x_S^{(i)}$ mientras $x_C^{(i)}$ permanece fijo

PDP,ICE-plot, M-plot, ALE plot

- M-plot promedia las predicciones sobre la distribución condicional

$$\hat{f}_{x_S,M}(x_S) = E_{X_C/X_S}[\hat{f}(X_S, X_C)/X_S = x_s] = \int_{x_C} \hat{f}(x_S, x_C) P(x_C/x_S) dx_C$$

- ALE-plot promedia el cambio en las predicciones y las acumula sobre la grilla.

$$\hat{f}_{x_S,ALE}(x_S) = \int_{z_{0,1}}^{x_S} E_{X_C/X_S}[f^S(X_S, X_C)/X_S = z_S] dz_S - cte$$

$$= \int_{z_{0,1}}^{x_S} \int_{x_C} f^S(z_s, x_C) P(x_C/z_S) dx_C dz_S - cte$$

ALE plot

Tres diferencias con M-plot:

1. Promediamos el cambio en la predicción no la predicción, el cambio es definido como el gradiente $\hat{f}^{S}(x_C, x_S) = \frac{\delta \hat{f}(x_S, x_C)}{\delta x_S}$
2. Hay una integral adicional sobre z, se acumulan los gradientes sobre el rango de variables sobre el conjunto S
3. Restamos una constante del resultado este paso centra ALE-plot tal que el efecto promedio sobre los datos es cero

Estimación, ALE plot

- Para estimar el efecto dividimos la variable en muchos intervalos y calculamos la diferencia en las predicciones.
- Primero se estima el efecto sin centrar:

$$\tilde{f}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j \in N_j(k)} [f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)})]$$

Estimación, ALE plot

- ALE calcula la diferencia de las predicciones donde se remplaza el valor de la variable de interés por el valor de la grilla.
- La diferencia en la predicción es el efecto de la variable para una observación particular en cierto intervalo.
- La suma a la derecha agrega el efecto de todas las observaciones al interior de un intervalo que aparece como el vecindario $N_j(k)$ y se divide por el número de observaciones en ese intervalo. Este es un promedio local
- La suma a la izquierda significa que acumulamos el promedio del efecto en todos los intervalos
- El ALE de una variable con un valor que cae en en el tercer intervalo es la suma del efecto de los primeros tres (acumulado)

Estimación, ALE plot

El efecto se centra para que el efecto promedio sea cero

$$\hat{f}_{j,ALE}(x) = \tilde{f}_{j,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{j,ALE}(x_j^{(i)})$$

- El valor de ALE puede interpretarse como el efecto medio de la variable en cierto valor comparado con el promedio de la predicción en los datos

Si ALE es -2 en $x_j = 3$ significa que cuando la variable j-ésima tiene el valor 3 entonces las predicción es dos puntos más chica que la predicción promedio.

Ejemplo, ALE plot

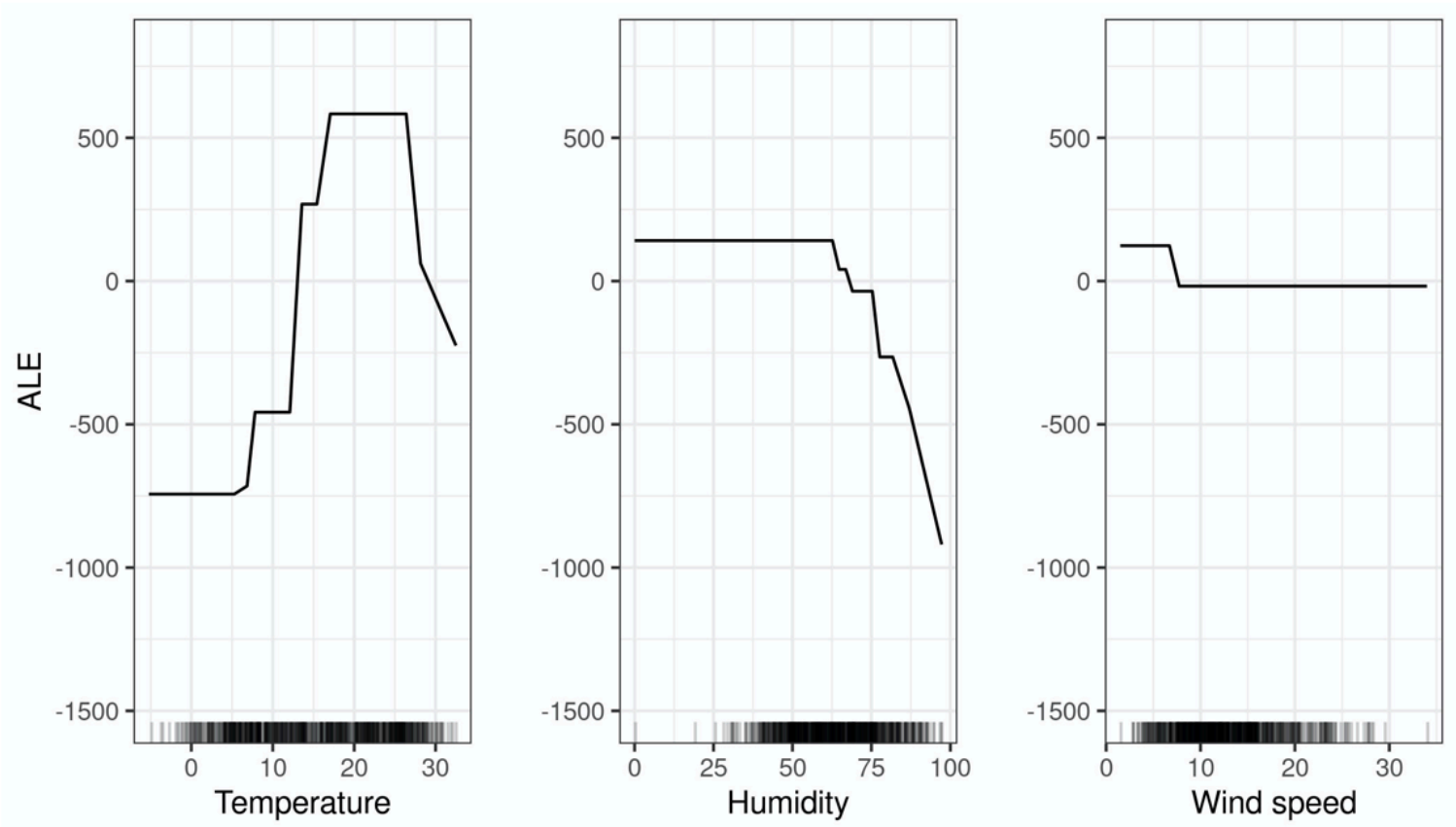


FIGURE 8.11: ALE plots for the bike prediction model by temperature, humidity and wind speed. The temperature has a strong effect on the prediction. The average prediction rises with increasing temperature, but falls again above 25 degrees Celsius. Humidity has a negative effect: When above 60%, the higher the relative humidity, the lower the prediction. The wind speed does not affect the predictions much.

Ventajas, ALE plot

- ALE-plot son insesgados
- Más rápidos de calcular que PDP
- Interpretación clara

Desventajas, ALE plot

- Pueden ser rugososo cuando se incrementa el número de intervalos
- ALE-plot nos son acompañadas por curvas ICE como PDP
- Implementación es más compleja que PDP

Implementar en R

Explorar el paquete `DALEX` y como usarlo con `tidymodels`

también ver `iml` y `pdp`

Referencias