REGRESIÓN CUANTILICIA

Si está mirando el precio de una casa en un mercado en particular y su modelo predice que una casa se venderá por $262,458.45, ¿qué tan seguro puede estar de que la predicción de su modelo es exactamente correcta? Es de esperar que su instinto le diga que existe una posibilidad infinitesimalmente pequeña de que esto sea cierto, pero tal vez exista la posibilidad de que su modelo esté cerca de la respuesta real. Necesitamos una forma de predecir un rango de valores mientras tenemos cierta confianza en ese rango.

Se puede ver la regresión por cuantiles como una extensión de la regresión lineal. A diferencia de la regresión lineal regular que usa el método de mínimos cuadrados para calcular la media condicional del objetivo a través de diferentes valores de las características, la regresión por cuantiles minimiza la pérdida por cuantiles al predecir un determinado cuantil. El cuantil más popular es la mediana, o el percentil 50, y en este caso la pérdida del cuantil. Otros cuantiles podrían dar puntos finales de un intervalo de predicción; por ejemplo, un rango medio del 80 por ciento está definido por los percentiles 10 y 90.

Tradicionalmente, el modelo de regresión lineal para calcular la media toma la forma:

Imagen que contiene Gráfico

Descripción generada automáticamente

La mejor línea de regresión lineal se encuentra minimizando el error cuadrático medio, que se encuentra con la ecuación

Texto

Descripción generada automáticamente

Ahora, para la regresión de cuantiles, no está limitado a solo encontrar la mediana, sino que puede calcular cualquier cuantil (porcentaje) para un valor particular en las variables de características.

Por ejemplo: si tuviéramos que encontrar el cuantil 25 para el precio de una casa en particular, eso significaría que hay un 25 % de probabilidad de que el precio real de la casa esté por debajo de la predicción, mientras que hay un 75 % de probabilidad de que el precio está arriba.

Tomando una estructura similar al modelo de regresión lineal, la ecuación del modelo de regresión cuantil para el 𝜏-ésimo cuantil es

Gráfico

Descripción generada automáticamente

En lugar de ser constantes, los coeficientes beta ahora son funciones con dependencia en el cuantil. Encontrar los valores para estas betas en un valor de cuantil particular implica casi el mismo proceso que para la cuantificación lineal regular, excepto que ahora tenemos que reducir la desviación absoluta mediana.

Texto

Descripción generada automáticamente

Aquí, la función 𝜌 es la función de verificación que otorga pesos asimétricos al error según el cuantil y el signo general del error.

Texto

Descripción generada automáticamente

En este caso, u es el error de un solo punto de datos y la función max devuelve el valor más grande entre paréntesis. Esto significa que si el error es positivo, la función de verificación multiplica el error por 𝜏, y si el error es negativo, la función de verificación multiplica el error por (1- 𝜏).

Por ejemplo, si desea la mediana del percentil 10, eso significa que desea que el 90 % de los errores sean positivos y el 10 % negativos. Para encontrar la MAD más pequeña mientras esta declaración sea verdadera, se deben agregar pesos a los errores. En el caso del cuantil 10, se suma un peso de 0,9 a los pesos negativos mientras que a los positivos se le suma un peso de 0,1.

**Contraste**

Supongamos que un analista de bienes raíces quiere predecir los precios de las viviendas a partir de factores como la antigüedad de la vivienda y la distancia de los centros de trabajo. El objetivo típico será generar la mejor estimación del punto de precio de la vivienda dados esos factores, donde "mejor" a menudo se refiere a la estimación que minimiza las desviaciones al cuadrado de la realidad.

Por lo tanto, la regresión de cuantiles predice la proporción de pueblos (no hogares) con valores medios de vivienda por debajo de un valor.

**Explicación un poco más matemática**

Sea Y una variable de respuesta de valor real y X una covariable o variable predictora. El análisis de regresión estándar trata de llegar a una estimación ˆµ(x) de la media condicional E(Y |X = x) de la respuesta variable Y, dado X =x. La media condicional minimiza la pérdida de error cuadrático esperada y la aproximación de la media condicional se logra típicamente mediante la minimización de una función de pérdida de tipo error cuadrático sobre los datos disponibles.

La media condicional ilumina solo un aspecto de la distribución condicional de una variable de respuesta Y, pero ignora todas las demás características de posible interés. Esto condujo al desarrollo de la regresión por cuantiles.

La función de distribución condicional F(y|X = x) viene dada por la probabilidad de que, para X = x, Y sea menor que y ∈ R.



Para una función de distribución continua, el cuantil α Qα(x) se define entonces de tal manera que la probabilidad de que Y sea menor que Qα(x) es, para un X = x dado, exactamente igual a α. En general



Como ejemplo, considere las predicciones de los niveles de ozono del día siguiente, como en Breiman y Friedman (1985). La regresión de mínimos cuadrados intenta estimar la media condicional de los niveles de ozono. Por ejemplo, podría ser de interés encontrar un nivel de ozono que, con alta probabilidad, no se supere. Esto se puede lograr con la regresión por cuantiles, ya que brinda información sobre la dispersión de la variable de respuesta.

**Intervalos de predicción**

¿Qué tan confiable es una predicción para una nueva instancia? Esta es una pregunta relacionada de interés. Considere nuevamente la predicción de los niveles de ozono del día siguiente. Algunos días, podría ser posible identificar los niveles de ozono del día siguiente con mayor precisión que en otros días.

Con la predicción estándar, se devuelve una estimación de un solo punto para cada nueva instancia. Esta estimación puntual no contiene información sobre la dispersión de las observaciones en torno al valor predicho.

La regresión de cuantiles se puede utilizar para construir intervalos de predicción. Un intervalo de predicción del 95% para el valor de Y está dado por

Imagen que contiene Texto

Descripción generada automáticamente

Es decir, una nueva observación de Y, para X = x, está con alta probabilidad en el intervalo I(x).

<https://www.r-bloggers.com/2021/04/quantile-regression-forests-for-prediction-intervals/>

<https://towardsdatascience.com/quantile-regression-ff2343c4a03>

<https://towardsdatascience.com/quantile-regression-from-linear-models-to-trees-to-deep-learning-af3738b527c3>