Índice general

0.1. Regresión no paramétrica

Uno de los principales objetivos del aprendizaje estadístico es poder construir una buena herramienta para predecir. La teoría en la que nos basamos para poder construir estos predictores es la teoría de Esperanza Condicional. Repasemos un poco esta teoría ya que necesitamos de ella para poder entender mejor como manejarnos luego con los datos.

0.1.1. Un poco de teoría

Predicción Sea Y una variable aleatoria, $\underline{X} = (X_1, X_2, \ldots, X_n)$ un vector aleatorio, existirá alguna función $g(\underline{X})$ que nos sirva para predecir a Y. Para encontrar dicha función se calcula el error cuadrático medio:

$$ECM = \mathbf{E}[(Y - g(\underline{X}))^2]$$

Y se busca la función g que minimiza el error.

El mejor predictor constante será $\mathbf{E}(Y)$.

El mejor predictor lineal se llama **Recta de regresión**. Si pensamos el caso para el vector bidimensional (X, Y), entonces la recta de regresión es la ecuación que resulta de minimizar

$$ECM = \mathbf{E}[(Y - (aX + b))^2].$$

Aplicando propiedades de la esperanza, buscamos las derivadas parciales en función de a y de b y despejando resulta

$$g(X) = \hat{Y} = \frac{\mathbf{cov}(X, Y)}{\mathbf{var}(X)}(X - \mathbf{E}[X]) + E[Y]$$

El mejor predictor será la **Esperanza Condicional**. La teoría ahora es un poco más complicada, necesitamos entender como se comportan las variables aleatorias condicionadas.

0.1.2. Variables aleatorias condicionadas

Vectores discretos

Definición: Sean X e Y variables aleatorias discretas con $p_X(x) > 0$, la función de probabilidad condicional de Y dado que $X = x$ es

$$p_{Y|X=x}(y) = \mathbf{P}(Y = y|X = x) = \frac{\mathbf{P}(Y = y, X = x)}{\mathbf{P}(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

Se define como 0 a $p_{Y|X=x}(y)$ cuando $p_X(x) = 0$.

(Para cada valor x que tome la variable aleatoria X, tendremos una variable aleatoria $Y|X = x$ diferente, con su correspondiente función de probabilidad)

Propiedad: Sean X e Y vectores aleatorios discretos tal que
$p_{Y|X=x}(y) = p_Y(y)$ para todo $x \in \mathbb{R}$, entonces X e Y son independientes.

De la definición de esperanza para variables aleatorias discretas resulta

$$\mathbf{E}[Y|X=x] = \sum_{y \in R_Y} y p_{Y|X=x}(y), \forall x \in R_X$$

Vectores continuos

Definición: Sea (X,Y) un vector aleatorio con densidad conjunta $f_{X,Y}(x,y)$ y densidad marginal $f_X(x)$, entonces para cualquier valor de X con el cual $f_X(x) > 0$, la función de densidad condicional de Y dado $X = x$ es

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

De la definición de esperanza para variables aleatorias continuas resulta

$$\mathbf{E}[Y|X=x] = \int_{y \in R_Y} y f_{Y|X=x}(y) dy, \forall x \in R_X$$

Esperanza condiiconal

A las funciones $\mathbf{E}[Y|X=x]$, que son funciones de x, se las llama funciones de regresión, y se las denota $\varphi(x)$

Llamemos $\varphi(x) = \mathbf{E}[Y|X=x]$, luego $\varphi : sop_X \to \mathbb{R}$. La variable aleatoria llamada Esperanza condicional de Y dado X, denotado por $\mathbf{E}[Y|X]$, se define por $\varphi(X) = \mathbf{E}[Y|X]$.

Definición: La variable aleatoria esperanza condicional de Y dada X se define como $\varphi(X) = \mathbf{E}[Y|X]$ con φ una función medible tal que $\mathbf{E}((Y - \varphi(X)) * t(X)) = 0$ para toda función t medible $t : R_X \to \mathbb{R}$ tal que $Y * t(X)$ tiene esperanza finita(t es cualquier función en el plano, $\varphi(X)$ será el mejor predictor lineal de Y basado en X).

Obs: La esperanza condicional siempre existe y a demás es única con probabilidad 1

Propiedades

1. $\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y|X]]$
2. X e Y vectores aleatorios, s y r funciones medibles tales que las variables aleatorias $r(X)*s(Y)$, $r(X)$ y $s(Y)$ tienen esperanza finita, entonces $\mathbf{E}(r(X) * s(Y)|X) = r(X) * \mathbf{E}(s(Y)|X)$
3. Y1,Y2, V.A. con esperanza finita, X vector aleatorio, E(aY1+bY2—X)=aE(Y1—X)+ bE(Y2—X)
4. $\mathbf{E}(Y|X) = \mathbf{E}(Y)$ si X e Y son independientes
5. $\mathbf{E}(r(X)|X) = r(X)$

Predicción: "La esperanza condicional de Y dado X es la función de la variable aleatoria X que mejor predice o se aproxima a Y"

Recordamos la definición:
Sea Y una V.A., $\underline{X} = (X_1, X_2, \ldots, X_n)$ un vector aleatorio, existirá alguna función $g(\underline{X})$ que nos sirva para predecir a Y. Para encontrar dicha función se calcula el error cuadrático medio:

$$ECM = \mathbf{E}[(Y - g(\underline{X}))^2]$$

Y se busca la función g que minimiza el error.

La esperanza condicional de Y dada $\underline{X}$ cumple que:

$$\mathbf{E}[(Y - g(\underline{X}))^2] \geq \mathbf{E}[(Y - \mathbf{E}(Y|\underline{X}))^2]$$

Para cualquier función $g(\underline{X})$

0.1.3. Estimación de la función de regresión

Supongamos que realizamos un experimento estadístico muchas veces, y observamos cada vez el valor de la variable aleatoria X. El promedio de dichas observaciones (bajo la mayoría de las circunstancias) va a converger a un valor fijo a medida que la cantidad de observaciones aumenta. Este valor es la Esperanza de X, $E[X]$. Esta afirmación se basa en la **Ley de los grandes números** que dice que

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E[X]$$

si las variables $X_1, \ldots, X_n$ son *i.i.d.*

En R, el promedio se calcula con la función 'mean()' (Que en inglés es media). Al igual que como lo hicimos con la estimación de una probabilidad, puedo realizar simulaciónes para un valor de n cada vez mayor y observar como las estimaciones se acercan cada vez más al valor verdadero. Esta herramienta nos será de mucha utilidad, ya que hay experimentos que son muy complicados, con funciones difíciles de integrar o sumar, y si logramos simular el experimento, el cálculo estimado de la esperanza es muy sencillo. En materias de estadística previas ya se estudió cuales son los mejores estimadores para la media, varianza, covarianza, y otros parámetros, y las propiedades de insesgadez y consistencia de los mismos.

Buscamos ahora aprender a estimar la función de regresión. Supongamos que $\varphi(x)$ es la verdadera función de regresión. Hay diferentes formas de estimar a $\varphi(x)$ que solo asumen que es suave. Estos métodos se llaman métodos de estimación no paramétrica. Sabemos, de la teoría de probabilidades, que la función de regresión está dada por

$$\varphi(x) = E[Y|X = x]$$

Lo que buscamos en estudiar alguna relación entre Y y X mediante la forma $Y \sim \varphi(X)$ de manera de minimizar el error cuadrático medio.

Una forma razonable para estimar a $\varphi(x)$ sería calcular el promedio de todos los valores de y para cada valor de x, siguiendo la idea de que el promedio es una muy buena estimación para la esperanza de una variable aleatoria.

Entonces, si X es discreta y tenemos un conjunto de observaciones del vector (X, Y), (x, y), podríamos estimar a $\varphi(x)$ por

$$\hat{\varphi}(x) = \frac{\sum_{i=1}^{n} y_i * \mathbf{I}\{x_i = x\}}{\sum_{i=1}^{n} \mathbf{I}\{x_i = x\}}$$

Observamos que esta fórmula calcula exactamente lo explicado, sumamos los valores de y que corresponden a la observación $X = x$ y lo dividimos por el total de esas observaciones, entonces para cada valor de x estamos calculando el promedio de las y. Eso dará una buena estimación de la función de regresión cuando las variables son discretas. Ejemplo urnas y bolitas

Ahora, ¿qué pasa si las variables son continuas? Es muy probable que estos casos ocurra que para cada valor de x obtenga un solo valor de y, por lo que la curva resultante pasará por todas las observaciones y no será suave, tampoco dará una buena estimación de la función de regresión. Entonces en este caso, una forma razonable para estimar a $\varphi(x)$ puede ser calcular el promedio de las respuestas cercanas a x, por ejemplo dentro de una ventana definida en un entorno de x, $(x-h, x+h)$. El procedimiento de encontrar promedios locales es la idea central de lo que se conoce como suavizado.

0.1.4. Estimador de Nadaraya Watson

Veremos uno de los mejores estimadores no paramétricos para la función de regresión: El estimador de **Nadaraya-Watson**.

Volviendo a la definición de función de regresión (para el caso continuo)

$$\varphi(x) = E[Y|X = x] = \frac{\int_{-\infty}^{\infty} y * f(x, y)dy}{f(x)}$$

Entonces podemos usar estimaciones para las densidades usando núcleos. Recordemos que

$$\hat{f}_h(x) = \frac{1}{nh_X} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_X}\right).$$

Con la misma idea, tendremos una estimación para la densidad conjunta

$$\hat{f}_h(x, y) = \frac{1}{nh_X h_Y} \sum_{i=1}^{n} K_X\left(\frac{x - x_i}{h_X}\right) K_Y\left(\frac{y - y_i}{h_Y}\right).$$

Si definimos

$$\hat{\varphi}(x) = \frac{\int_{-\infty}^{\infty} y * \hat{f}(x, y)dy}{\hat{f}(x)}$$

Reemplazando y resolviendo resulta el estimador de Nadaraya-Watson

$$\hat{\varphi}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) * y_i}{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} = \sum_{i=1}^{n} w_i(x) y_i$$

Teorema: Sea $X \in \mathbb{R}^p$, si $n \to \infty$, $h \to 0$ y $nh \to$ ínf, entonces $\hat{\varphi}(x) \to \varphi(x)$.

Este estimador es un promedio ponderado de las variables respuesta y. La forma de los pesos está dado por la función K y el tamaño de los pesos está parametrizado por h. El valor óptimo de

h se consigue tratando de minimizar el error cuadrático medio estimado mediante una técnica que lleva el nombre de validación cruzada.

Si definimos el error de valización cruzada como

$$CV(h) = \hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n}\hat{L}_i(h)$$

con

$$\hat{L}_i(h) = (y_i - \hat{\varphi}^{(-i)}(x_i))^2$$

entonces

$$h_{cv} = arg\,\min_h \hat{L}(h)$$

.

Se puede observar que si h tiende a cero entonces el estimador diende a y_i, pequeñas ventanas reproducen los datos y disminuyen el sesgo, mientras que si h tiende a infinito entonces el estimador tenderá al promedio de las y dando curvas super suaves con menor variabilidad.

En R, para usar el estimador de N-W se puede utilizar la función `ksmooth`, donde aclaramos cual es el núcleo que queremos usar y el valor de h (bandwith). Ejemplo: Lidar.

0.1.5. Polinomios locales (loess)

Busca el ajuste a polinomios locales usando mínimos cuadrados y núcleos. Se parte del desarrollo de Taylor de orden p de la función de regresión en un punto que esté en el vecindario de x y luego se incluyen pesos usando núcleos, resultando el siguiente problema de minimización:

$$\min_\beta \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p)^2 K\left(\frac{x - x_i}{h}\right)$$

El estimador de N-W puede verse como un caso particular de este estimador. La solución a este problema se encuentra con el método llamado mínimos cuadrados pesados. Podremos desarrollarlo mejor luego del capítulo de regresión lineal.

0.1.6. Vecinos más cercanos

Lo veremos en detalle en el capítulo de clasificación. Lo que busca es estimar a $\varphi(x)$ promediando los valores de las y_i que se encuentran más cerca de x, resultando

$$\hat{\varphi}(x) = \frac{1}{j}\sum_{i\in J_k} y_i$$

con $J_k = \{i : x_i$es una de las k observaciones más cercanas a $x\}$

0.1.7. Estimadores basados en Splines

Considera la suma de residuos al cuadrado como criterio de ajuste

$$RSS = \sum_{i=1}^{n}(y_i - \varphi(x))^2$$

Para poder resolver el problema de minimización, se agrega una restricción porque si no resulta en una curva que pasa por todos los datos. Se agrega un factor que penaliza la no suavidad de $\varphi(x)$

$$\hat{\varphi}_\lambda(x) = arg\min_{\varphi}(\sum_{i=1}^{n}(y_i = \varphi(x))^2 + \lambda \int \varphi''(x)^2 dx)$$

Este problema de minimización tiene una solución única definida como Spline cúbica con las siguientes propiedades:

- $\hat{\varphi}(x)$ es un polinomio cúbico entre dos valores sucesivos de x
- entre los puntos x_i la curva tiene sus primeras dos derivadas continuas
- en $x_{(1)}$ y $x_{(n)}$ la segunda derivada de la curva es cero.