UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS

FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS

“ El propósito de la demostración no es verificar, es comprender. ”
-ARNOLD ROSS

MM-422

MÉTODOS LINEALES

Informe:
Estadística no-paramétrica

25 de abril de 2019
Primer periodo

Elaboradas por:

Luis Felipe Flores Machado

Número de cuenta:

20141030258

Índice general

Capítulo 1

SECCIÓN PREAMBULAR

"Si he logrado ver más lejos, ha sido porque he subido a hombros de gigantes"

-Isaac Newton

1.1 Introducción

La **estadística no-paramétrica** es la rama de la estaditca que no está basada solamente en el análisis de familias parametrizadas de distribuciones de probabilidad. Por ejemplo, en modelos en que se busca determinar (o estimar) los valores reales de la varianza y la media de conjunto de datos a la cual se le asume una distribución conocida teórica a partir de un muestreo.

La estadística no-paramétrica está basada en no tener especificado los parámetros del modelo o bien no tener ni especificada la estructura de la distribución subyacente de los datos. En este último caso se le conoce como *modelo libre de distribución*.

Antes de proseguir, citamos el siguiente *Kendall*[1]:

> “ Las hipótesis estadística se realizan con aspiraciones a entender el comportamiento de los observables que las variables aleatorias modelan... por ejemplo tenemos la hipótesis de que (a) *una distribución normal tiene una media y varianza específica*; la hipótesis de que (b) *tiene una media específica pero una varianza no especificada*, así como la hipótesis de que (c) *no tiene una media especificada ni tampoco la varianza especificada* y así como la hipótesis de que (d) *dos hipótesis no especificadas son idénticas*.
>
> Es de notar que en los ejemplos (a) y (b), las observaciones están asumidas de provenir de una distribución con estructura específica (en este caso, la distribución normal) y que entonces la hipótesis está totalmente interesada sólo en los valores de uno o ambos parámetros de dicha distribución. Por razones obvias, este tipo de hipótesis es llamada *“hipótesis paramétrica”*.
>
> La hipótesis (c) es de distinta naturaleza completamente ya que ninguno de sus parámetros está especificado y por ende la hipótesis trata totalmente sobres si los datos realmente pertenecen o no a dicha distribución sin importar qué tan bien se “ajustan” a ciertos parámetros. Este tipo de hipótesis se puede categorizar razonablemente como *“hipótesis no-paramétrica*.
>
> La hipótesis (d) de igual manera es no paramétrica, sin embargo, adicionalmente tenemos que la distribución subyacente tampoco se especifica, y la característica que entonces nos interesa es totalmente enfocada en similitud de tendencias estocásticas de ambos conjuntos de datos. Por ende, este tipo de hipótesis puede ser llamada razonablemente como *libres de distribución*.
>
> No respetando dicha distinción, la literatura usual de estadística comúnmente aplica la etiqueta de *no paramétrico* a los procedimientos de prueba que recién hemos llamado como *libres de distribución*, perdiendo así una clasificación útil. ”

Ejemplos de estadísticos no paramétricos son la estadística descriptiva y la inferencia estadística. Ambas serán posteriormente discutidas.

En la siguiente sección expandimos discusión sobre estos tipos anteriormente mencionados. Se dará además una revisión detallada de varios ejemplos de modelos y métodos no-paramétricos y otros libres de distribución para ilustrar su real diferencia y aplicaciones.

[1] Stuart A., Ord J.K, Arnold S. (1999), Kendall’s Advanced Theory of Statistics: Volume 2A–Clssical Inference and the linear model, Sexta edición, secciones 20.2 y 20.3

1.2 Definiciones básicas

Una primera definición heurística y no tan precisa yace entre las dos siguientes definiciones:

1. El primer significado de *no-paramétrico* cubre técnicas que dependen de datos pertenecientes a ninguna familia paramétrica de distribuciones de probabilidad. Esto, como mencionado anteriormente son:

   - Métodos libres de distribución, que no dependen de asunciones respecto a la pertenencia de nuestros datos a una distribución de probabilidad determinada. Por ende, es literalmente *lo opuesto* a la estadística paramétrica.
   - Estadísticos no paramétricos, que es un **estadístico** definido como una función de la muestra o los datos de la muestra, en contraste a dependiende de un parámetro externo a determinar.

   La estadística de orden, la cual está basada en un *ranking* de las observaciones (u orden inducido), es un ejemplo de estadístico no paramétrico, pero no necesariamente uno libre de distribución en el caso que nos estemos auxiliando de asunciones extras.

2. El segundo significado de la palabra *no paramétrico* cubre técnicas que no asuen la estructura del modelo mismo, o bien no la fijan. Típicamente, el modelo en estos casos crece en tamaño para acomodarse dinámicamente a la complejidad de los datos.

   En estas técnicas, variables individuales son típicamente asumidas de pertenecer a una distribución paramétrica y también se asumen los tipos de conexiones entre dichas variables. Estas técnicas son, entre otras:

   - Regresión no paramétrica, la cual consiste en modelar la estructura de la relación entre variables de forma no paramétrica, aunque pueden haber aún asunciones paramétricas de la distribución de los residuales del modelo.
   - Modelos bayesianos jerárquicos no-paramétricos, como modelos basados en procesos de Dirichlet, los cuales permiten que el número de variables latentes aumente dinámicamente lo que sea necesario para ajustarse a los datos; sin embargo, variables individuales aún pueden seguir distribuciones paramétricas e incluso el proceso mismo que controla el ritmo de crecimiento de las variables latentes puede seguir una distribución paramétrica.

Una lista breve de modelos a consideras es la siguiente:

1. Histograma
2. Estimación de denidad de kernel/nucleo
3. Regresiones no-paramétricas y semi-paramétricas
4. Análisis envolvente de datos
5. KNN (K-nearest neighbors); “Algoritmos de k-cercanía”
6. Máquina de vectores de soporte
7. Método de momentos

Estos serán considerados en la siguiente sección

1.3 Aplicaciones y propósito de uso

Métodos no paramétricos son utilizados ampliamente para estudiar poblaciones que de manera intrínseca un orden o *ranking* (tal como los ranking de películas en base a su puntuación por críticos). El uso de métodos no paramétricos puede que sea necesario cuando los datos tienen un una estructura de orden más no una representación numérica obvia, al como preferencias de visita en una lista de lugares turísicos (uno no puede fácilmente cuantificar el "deseo de visitar cierto sitio", pero sí compararlos)

Ya que los métodos no paramétricos, por lo general, realizan menos asunciones, su aplicabilidad es mucho más amplia que la de los métodos paramétricos correspondientes. En particular, pueden ser aplicados a situaciones en que menos información es conocida. Además, son por definición más robustos.

Otra justificación para el uso de métodos no paramétricos es la simplicidad de su uso. En algunos casos, incluso cuando el uso de modelos paramétricos es justificado los métodos no paramétricos pueden ser más fáciles y económicos de utilizar. Debido además a, ambas, simplificad y robustez, los métodos no paramétricos son usualmente vistos por algunos estadísticos como ideales para permitir menor margen de uso impropio y malos entendidos.

Sin embargo, la amplitud de aplicabilidad y robustez de las pruebas no paramétrices viene con un costo: En casos donde una prueba paramétrica es apropiada, las pruebas no paramétricas tienen menor poder de predicción, en otras palabras, es necesario un tamaño de muestra mayor para realizar una predicción con el mismo nivel de confinanza al del método paramétrico.

Capítulo 2

MODELOS Y SU USO

"El nitrógeno de nuestro ADN, el calcio de nuestros dientes, el hierro de nuestra sangre, el carbono de nuestras tartas de manzana se hicieron en los interiores de las estrellas en proceso de colapso.
¡Estamos hechos de sustancia estelar!."

-Carl Sagan

2.1 Modelos no-paramétricos

Los modelos no-paramétricos difieren de los modelos paramétricos en que la estructura del modelo no está especificada a priori (y por ende, ingresada a la teoría mediante una selección empírica) si no que está en su lugar determinada por los datos.

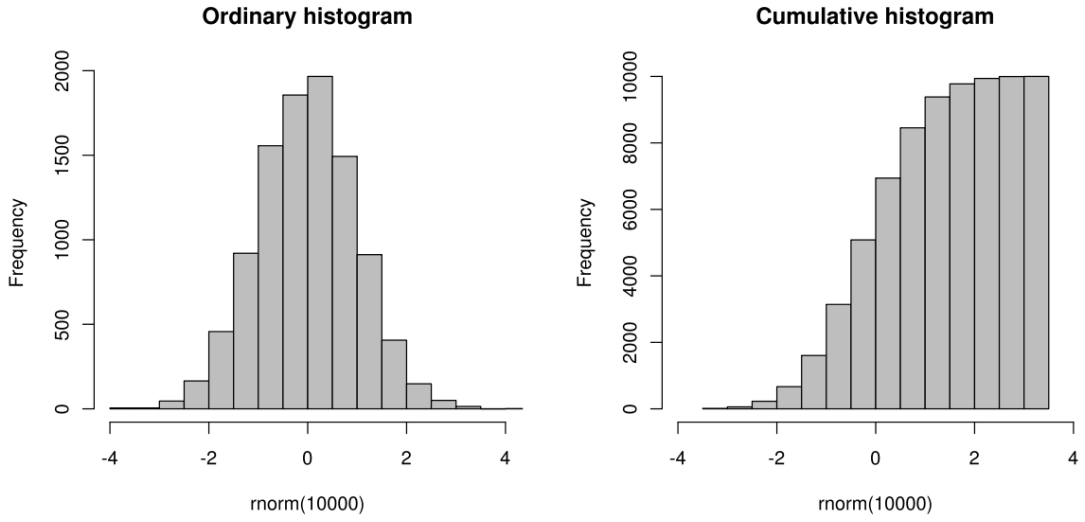
El término "no-paramétrico" no pretende significar que hay absoluta carencia de parámetros, si no que el número y naturaleza de los parámetros son flexibles y no fijos de antemano.

Algunos ejemplos de tales modelos son los siguientes:

2.1.1 Histograma

- Un **histograma** como modelo no-paramétrico para estimar la distribución de probabilidad de donde los datos fueron elegidos.

  Un ejemplo de histograma se ilustra a continuación:



  En la figura anterior observamos un histograma ordinario y un histograma acumulativo de los mismos datos. En este caso tenemos que los datos fueron extraidos de una distribución normal con media 0 y desviación estandar 1 (es decir, la distribución normal estándar).

  La diferencia entre ambos histogramas es que el primero se utiliza para aproximar la **función de densidad de probabilidad** mientras que la otra aproxima la **función de probabilidad acumulada**.

  Un histograma es una representación precisa de la distribución de datos numéricos. Esto estima la distribución de probabilidad de variables aleatorias y fue introducida por primera ver por Karl Pearson.

  La diferencia crucial de un histograma con un sencillo gráfico de barras es que el gráfico de barras estrictamente relaciona dos variables, pero los histogramas solo hablan de una sola variable y su distribución de posiles valores medidos.

  En cuanto a cómo decidir el número y grosor de las barras, no hay realmente una pauta clara. No existe, en general una "mejor" elección en cuanto a ello, ya que diferentes tamaños de barras pueden revelar diferentes características de los datos y ser diferentemente valiosas dependientes del contexto.

Por ejemplo, utilizar barras menos delgadas donde la densidad de probabilidad subyacente a los datos es pequeña, reduce el ruido debido a la aleatoriedad del muestro; utilizar barras más angostas donde la densidad es alta provee una mayor precisión en la estimación de dicha densidad. No obstante, lo más común es utilizar barras de igual grosor para todo el gráfico.

Algunos teóricos han intentado determinar un número óptimo de barras pero usualmente estos vienen con el costo de realizar asunciones fuertes de la distribución. Dependiendo de la distribución de los datos reales y la meta en mente, diferentes acercamientos son apropiados. A continuación presentaremos algunas de ellas:

1. **Asignación directa en base a grosor**: El número k de barras puede ser asignado directamente mediante un grosor h sugerido por, quizá los márgenes del gráfico impreso, u otras caracerísticas visuales. La fórmula es la siguiente:
$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil .$$
donde hemos utilizado la función techo.
2. **Asignación de raiz cuadrada**: Esta es utilizada por muchos programas de hojas de cálculo y graficadoras como Microsoft Excel. La fórmula es:
$$k = \lceil \sqrt{n} \rceil$$
donde n es el número de datos de la muestra.
3. **Fórmula de Sturges**: Se deduce a partir de la distribución binomial y se asume que los datos tienen una buena aproximación a una distribución normal. Esto puede aplicarse con confianza cuando tenemos muchos datos, y bajo condiciones amparadas por el teorema del límite central.

   La fórmula es la siguiente:
$$k = \lceil \log_2 n \rceil + 1,$$
   Es bueno notar que, por el lento crecimiento de la función logaritmo, incluso si tenemos datos altamente normales, siempre asumimos que tenemos un n "grande" (empíricamente, mayor que 30) o si no, el número de barras predichas será muy pequeño para realmente mostrar cualidades importantes de la distribución.
4. **Regla de Rice**: Es usualmente utilizada como una alternativa a la fórmula de Sturges cuando nuestros datos no son tan grandes, sin embargo aún se asume normalidad. La fórmula es:
$$k = \lceil 2n^{1/3} \rceil, k = \lceil 2n^{1/3} \rceil$$
5. **Fórmula de Doane**: Las fórmula de Doane es una corrección de orden superior a la fórmula de Sturges que pretende mejorar el rendimiento tenemos datos con distribución más lejana a la de una normal. La fórmula es:
$$k = 1 + \log_2(n) + \log_2\left(1 + \frac{|g_1|}{\sigma_{g_1}}\right)$$
   donde g_1 es la *asimetría estadística* (skewness) estimada de nuestros datos o bien, de la distribución de la que sospechamos que pertenecen los datos, y
$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}.$$
   Notemos que el resultado de la fórmula anterior puede que requiera redondearse a conveniecia al siguiente o al anterior entero.

6. **Regla de referencia normal de Scott**: Aunque las anteriores fórmulas son más para correcciones estéticas cuando podemos calcular ciertas propiedades de la distribución, si estamos seguros que nuestros datos tienen uan distribución muy cercana a la normal, podemos de hecho minimizar el error cuadrático medio de dicha estimación mediante la siguiente elección de grosor:

$$h = \frac{3.5\hat{\sigma}}{n^{1/3}},$$

donde $\hat{\sigma}$ es la desviación estandar muestral de nuestros datos.

7. **Regla de Freedman–Diaconis**: Es una mejora a la regla de Scott anterior que implemente el rango intercuantil denotado por IQR, el cual es menos sensible que la desviación estandar cuando estamos en presencia de puntos atípicos en nuestros datos.

   La fórmula es la siguiente:

$$h = 2\frac{\text{IQR}(x)}{n^{1/3}},$$

   Donde $\text{IQR} = Q_3 - Q_1$ es la resta de los cuartiles 3 y 1. Además, la regla de Scott se ha encontrado ser inexacta para tamaños de muestra grandes (muy por encima de los $n = 200$), mientras que la de Freedman lo tolera.

   En caso de la fórmula de Freedman-Diaconis, lo que estamos optimizando no es el error cuadrático medio, si no el área necerrada entre nuestra distribución teórica y la distribución muestral.

8. **Asignación de Shimazaki and Shinomoto**: Esta asignación está basada en la función de costo-riesgo de tipo L^2:

$$\arg\min_h \frac{2\bar{m} - v}{h^2}$$

   donde $\bar{m} = \frac{1}{k}\sum_{i=1}^{k} m_i$ y $v = \frac{1}{k}\sum_{i=1}^{k}(m_i - \bar{m})^2$ son respectivamente la media y la varianza muestral del histograma. Con ello, se calcula el h que minimice la anterior expresión.

9. **Minimización del error cuadrático de validación cruzada**: Esta generaliza la regla de Scott hacia distribuciones que no necesariamente sean normales. Similar a la anterior, consiste en encontrar el h que minimice la siguiente expresión:

$$\arg\min_h \hat{J}(h) = \arg\min_h \left(\frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h} \sum_k N_k^2 \right)$$

   donde N_k es el número de datos en la $k-$ésima barra.

10. **Tamaños de barras variables:** A veces es mucho más conveniente no utilizar el mismo grosor para cada barra, debido a que quisieramos promover mejor ciertas características de la estimación de la distribución de datos en ciertas regiones mediante mayor o menor número de barras concentradas en dicha región.

Esto además evita que ciertas barras tengan pocas cuentas (frecuencia). Una manera usual de evitar dicho fenóemeno es motivarse por maximizar el poder de la prueba de chi cuadrado de Pearson, probando si las barras contienen todas un número igual de elementos e la muestra. Más específicamente, para un intervalo de confianza específico, α, es recomendable utilizar entre 1/2 y 1 veces la siguiente fórmula:

$$k = 4\left(\frac{2n^2}{\Phi^{-1}(\alpha)}\right)^{\frac{1}{5}}$$

donde $\Phi^{-1}(p) = \sqrt{2}\,\mathrm{erf}^{-1}(2p-1)$. Utilizando esta ecuación para el caso de $alpha = 0.5$, obtenemos valores entre: $1.88n^{2/5}$ y $3.77n^{2/5}$

2.1.2 Estimación de densidad de kernel

- La **estimación de densidad de kernel** que provee mejores estimaciones de la forma de la densidad de probabilidad que el histograma. Esta se se basa en encontrar una estimación $\hat{f}$ a la densidad f mediante la siguiente ecuación:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

donde K se lo que llamamos **kernel**, que es una función no negativa que *decae lo suficientemente rápido* a 0 cuando nos acercamos a infinito. h es llamado *ancho de banda*[1] que indica la precisión o fineza con la cual estaremos aproximando f y, finalmente, $K_h(x) = 1/hK(x/h)$ se conoce como el kernel escalado.

En teoría, nosotros quisieramos hacer a h tan pequeño como permitan los datos, sin embargo siempre habrá un intercambio entre el sesgo del estimador y la varianza del mismo. Por ello, existe un análisis profundo por realizar en cada caso para intentar minimizar el error cuadrático medio del estimador.

No obstante, una buena regla de oro para ello es utilizar la siguiente relación:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

donde $\hat{\sigma}$ es la desviación estandar de la muestra. Esta 'regla" realmente es sólo válida cuando utilizamos un kernel K gaussiano (discusión a continuación), sin embargo para tamaños de muestra bastante grandes puede ser igual de efectivo.

Sin embargo, es bueno tener cuidado con su uso cuando la distribución subyacente de nuestros datos (y por ende la densidad f) está bastante "lejana" de ser una gaussiana, pues puede fallar drásticamente y devolver valores demasiado grandes de error.

cabe además destacar que dicho parámetro h no necesita porqué ser constante entre todos los puntos de nuestros datos. En caso de no serlo, aunque vuelve más difícil el análisis de optimización y el tratamiento de datos, puede proveer un método mucho más robusto llamado **estimación adaptativa de densidad de kernel**.

[1] En inspiración a las aplicaciones de este modelo para procesos de señales.

Ahora, para entender la forma del estimador de kernel, consideremos el problema de estimar la función característica de una distribución, ya que por medio de ella estamos indirectamente estimando la densidad de probabilidad, al ser ambas transformada de Fourier de la otra.

Dada una muestra aleatoria $(x_1, x_2, \ldots, x_n)$, un estimador usual de la función característica $\phi(t) = \mathbb{E}\left[\exp(itX)\right]$ es:

$$\widehat{\varphi}(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itx_j}$$

Sin embargo, este estimador tiene el problema de que su transformada de fourier diverge en general. Por ello, necesitamos multiplicar por un factor de corrección que nos disminuya el valor para $|t|$ grande. Usualmente este factor es de la forma $\psi_h(t) = \psi(ht)$ donde ψ es una función cuadrado integrable tal que es 1 en el origen y luego desvanece a 0 en el infinito. El parámetro h (llamado aquí convenientemente también "ancho de banda") controla qué tan rápido se va a apasiguar $\widehat{\varphi}(t)$.

La más común elección de ϕ es la función gaussiana. En dicho caso la fórmula de inversión se puede aplicar y nuestor estimador de la densidad será:

$$\begin{aligned}\widehat{f}(x) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \widehat{\varphi}(t)\psi_h(t)e^{-itx}\,dt = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{1}{n}\sum_{j=1}^{n} e^{it(x_j-x)}\psi(ht)\,dt \\ &= \frac{1}{nh}\sum_{j=1}^{n}\frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-i(ht)\frac{x-x_j}{h}}\psi(ht)\,d(ht) = \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{x-x_j}{h}\Big),\end{aligned}$$

donde K es la transformada de Fourier de ϕ. Por ello, el estimador de densidad de kernel coincide con el estimador de la función característica.

2.1.3 Métodos de regresión no-paramétrica/semi-paramétrica

- **Métodos de regresión no-paramétrica y regresión semi-paramétricas** han sido desarrolados en términos de en base a los kernel, trazadores y ondículas[2].

  La regresión no-paramétrica es una clase de análisis de regresión en que el estadístico predictos no toma una forma predeterminada, si no que es construido acorde a la información deducida de los datos directamente.

  La regresión no paramétrica requiere, por ende, un tamaño de muestra mucho mayor que en la regresión basada en modelos paramétricos ya que los datos tienen tanto que proveer información para las estimaciones del modelo como ahora también la estructura misma del modelo.

  Hay varios ejemplos de regresión no paramétrica como los siguientes

  1. **Kringeaje** o **kringeado** (kringing): Originado en geoestadística, el kringeaje (también conocido como proceso de regresión Gaussiana) es un método de interpolación en el cua los valores interpolados son modelados mediante un proceso Gaussiano (proceso estocástico cuyos estados tienen una relación que se distribuye normal) que están gobernados por previas covarianzas.

     Bajo asunciones adecuadas en las covarianzas previas, el krinaje puede de hecho provee el **mejor predictor insesgado lineal** (conocido por sus siglas en inglés como BLUP). Por otro lado, métodos de interpolación basados en otros criterios como suavidad de los parámetros (por ejemplo, trazadores suaves) no necesriamente proveerán el BLUP ni tampoco los valores intermedios más probables.

     Este métodos es ampliamente utilizado en el régimen de análisis espaciales y espacio-temporales, además de experimentos computacionales (simulaciones). Esta técnica es además conocida como **predicción de Wiener–Kolmogorov** en honor a Norbert Wiener y Andrey Kolmogorov.

  2. **Regresión de kernel** (o regresión de nucleo): La regresión de kernel estima la variable dependiente confinua de un conjunto finito de datos mediante la convolución de posiciones de dichos datos mediante una de nucleo.

     Intuitivamente hablando, al función de nucleo tiene la tarea de determinar cómo difuminar la influencia de puntos de la data tal que sus valores puedan ser utilizados para predecir localidades cercanas.

     La idea central yace en tratar de encontrar m, la función de esperanza matemática condicional que asume que ya se dieron cierto conjunto de datos periféricos a uno particular, y entonces, en base a ellos, predecir el valor del dato particular en cuestión.

     Nadaraya y Watson, ambos en 1964, propusieron estimar a m mediante una media localmente pesada. El estimador de Nadaraya–Watson es:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i) y_i}{\sum_{j=1}^{n} K_h(x - x_j)}$$

[2]La transformada de ondícula es un tipo especial de transformada matemática que representa una señal en términos de versiones trasladadas y dilatadas de una onda finita (denominada óndula madre). Una ondícula puede pensarse como un "pulso" que se efectuó en determinado punto de la recta y luego se difunde, decayendo a cero a medida nos acercamos a infinito

Donde K_h es una función de nucleo o función de kernel basado en el ancho de banda h. El denominador debe de ser tal que dicha suma sea igual a 1.

Además del anterior estimador, tenemos otros. Por ejemplo:

Estimador de kernel de Priestley–Chao

$$\widehat{m}_{PC}(x) = h^{-1} \sum_{i=2}^{n} (x_i - x_{i-1}) K\left(\frac{x - x_i}{h}\right) y_i$$

Note que cada término es un aporte lineal en la respuesta con información "a dos pasos" (dependientes de x_i y x_{i-1}) junto con la información del kernel que decide cómo pesar dichos aportes.

Esto se mejora y generaliza con el siguiente estimador:

Estimador de Kernel de Gasser–Müller

$$\widehat{m}_{GM}(x) = h^{-1} \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K\left(\frac{x - u}{h}\right) du \right] y_i$$

donde, $s_i = \frac{x_{i-1}+x_i}{2}$. Note ahora que tenemos una dependencia mucho más complicada, pues la integral suma sobre todos los posibles aportes u desde uno particular s_i y su anterior.

3. **Regresión multiplicativa no-paramétrica:** La regresión multiplicativa no paramétrica es una forma de regresión no-paramétrica basada en la estimación mediante un kernel que es utilizado de forma multiplicativa (o bien, "pesando") a estimaciones a priori.

   La meta de estos métodos es estimar una respuesta (variable dependiente) basado en uno o más predictores (variables independientes) pero con especial precisión en caso que las siguientes condiciones se cumplan:

   - La forma de la dependencia o distribución de la variable respuesta sea desconocida.
   - Los predictores son se involucran en la expresión de la respuesta de forma acoplada. Es decir, la expresión de la dependencia que tiene la variable respuesta respecto a una predictora depende a su vez de las demás predictoras (acoplamiento)
   - La variable respuesta es cuantitativa o binaria (0 o 1).

   Además es notable que esta técnica puede tener validación cruzada. Esto, junto con su capacidad de modelar comportamientos complejos y de dependencias acopladas permite que sea un buen modelo para fenómenos de respuesta en organismos respecto a su ambiente.

   Una característica biológica del modelo de NPMR (regresión multiplicativa no paramétrica por sus siglas en inglés) es el hecho de que cualquier intolerancia del organismo hacia cualquiera una de las dimensiones del espacio de predictores resulta en una falla total del organismo en sí; Similarmente para el modelo de NPMR.

   Por ejemplo, asumamos que una planta necesita de un rango específico de humedad en un rango de temperaturas espeifico. Tanto si la humedad o la temperatura fallan de yacer en dicho rango, el organismo morirá. Esto se ve claro pues, si la temperatura está demasiado alta, no habrá cantidad de humedad que salve a la planta de no morir.

Matemáticamente esto funciona perfecto con el modelo NPMR ya que el producto de los pesos de un punto en el espacio de predicción será cero (o cercano a cero) si cualquiera de los predictores individuales es cero (o cercano a cero).

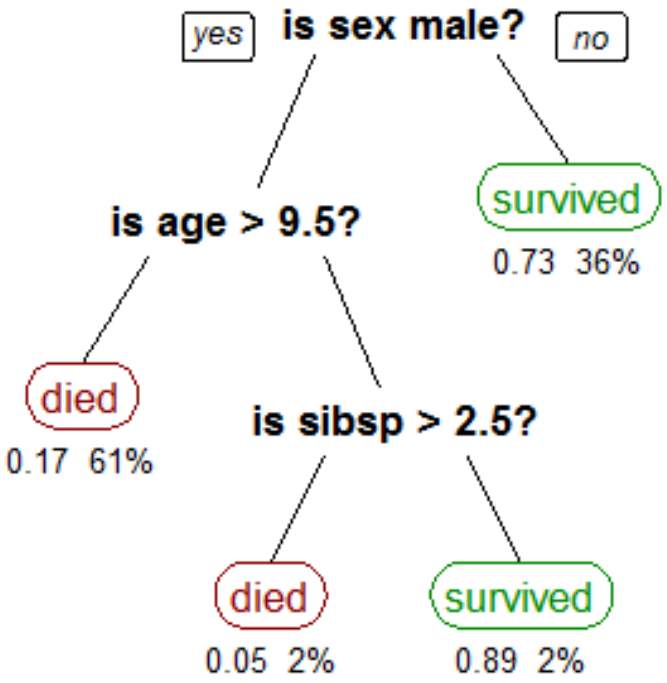
Note además que en este sencillo ejemplo la segunda condición expuesta arriba probablemente se cumpla también, pues existe una dependencia entre la humedad del ambiente con la temperatura y viceversa. Esto logra capturar correlaciones que provienen inicialmente de ciencias auxiliares como biología y física en un modelo puramente estadístico para aumentar la eficiencia de predicción sin considerar ecuaciones extras, si no que condiciones natas de dicho modelo.

Optimizar la elección de predictores y sus parámetros de suavidad (kernels) en un modelo multiplicativo es computacionalmente intensivo. Dado el tamaño usual del espacio de posibles predictores, la mayor desventaja de estos modelos es lo difícil de encontrar los mejores predictores entre todos ellos.

4. **Árboles de regresión/decisión:** Los algoritos de árboles de decisión se utilizan para explorar la estructura de cierto conjunto de datos para "aprender" a predecir de manera óptima una variable dependiente.

   A pesar de que originalmente el modelo (conocido entonces como árbol de clasificación y regresión, abreviado CART por siglas en inglés) fue formulado y aplicado sólo para predecir conjuntos de datos univariados, el marco teórico es fácilmente extendible para datos multivariados, incluyendo hasta series de tiempo.

   Observemos el siguiente diagrama descriptivo sobre los sobrevivientes del accidente del Titanic:



El interior de cada nodo corresponde a una de las variabes de entrada (variables independientes/de control) y las aristas muestran caminos hacia todos los posibles valores de dichas variables.

Las hojas representan posibles valores de la variable de respuesta dado los valores de control registrados en el camino de la raíz hasta la hoja.

El parámetro "sibsp" indita el número de familiares (ya sea hermanos, hijos o cónyugue) abordo, y el parámetro "age" nos ayuda a distinguir si la persona es un niño pequeño o ya un joven o adulto.

Estas condiciones nos ayudan a clasificar en diferentes categorías los datos que tenemos acerca de sobrevivientes de dicho accidente. Los valores abajo de cada hoja representa la probabilidad de haber sobrevivido dado el camino de condiciones seguido. Por ejemplo, esto muestra que la probabilidad de sobrevivir era buena si (i) se es mujer o (ii) se es niño varón menor de 9.5 años y menos de 2.5 familiares en promedio.

El anterior es un ejemplo de árbol de clasificación, sin embargo, existe los siguientes otros tipos de árboles de regresión:

- Árbol de clasificación
- Árbol de regresión numérica
- árbol de agregación de bootstrap (o empaquetadores)
- Árbol de potenciación de gradiente
- Bosque aleatorio (potenciado mediante análisis de componentes principales)
- Árbol de aprendizaje de decisión

Cada uno se utiliza dependiente de estructuras específicas, tamaños, ruidos y dimensiones de nuestros datos.

Definamos ahora la diferencia entre un modelo no-parampetrico y uno semi-paramétrico. a semiparametric model is a statistical model that has parametric and nonparametric components.

Un modelo estadístico es una familia parametrizada de distriuciones: $\{P_\theta : \theta \in \Theta\}$ indexada por un parámetro θ.

Un modelo paramétrico es un modelo en el que el parámetro de indexación θ es un vector en un espacio euclidio k-dimensional, para algún entero no-negativo k. Por ende, θ es finito-dimensional, y $\Theta \subseteq \mathbb{R}^k$.

En un modelo no-parametrico, el conjunto de posibles valores del parámetro θ es un conjunto de algún espaco V, que no es necesariamente finito-dimensional.

Por ejemplo, podriamos desear el considerar el conjunto de todas las distribuciones con media 0. Tales espacios son espacios vectoriales con una estructura topológica, pero puede que no sean finito dimensionales respecto a dicha estructura lineal. Así entonces, $\Theta \subseteq V$ para algún espacio infinito dimensional V.

En un modelo semi-parametrico, el parámetro θ tiene ambas, una componente en un espacio infinito-dimensional (usualmente una función de valor real definida sobre la recta real) y una componente finito-dimensional. Así, $\Theta \subseteq \mathbb{R}^k \times V$, donde V es un espacio infinito dimensional.

Puede parecer a simple vista que los modelos semi-paramétricos son modelos que incluyen los no-paramétricos como ub subconjunto o caso especial. Sin emargo, los modelos semi-paramétricos usualmente se consideran como una clase "más pequeña" de los modelos completamente no-paramétricos debido a que con frecuencia en los semi-paramétricos nos interesamos solamente en la parte finito-dimensional de θ. Esto es debido a que la parte infinito-dmensional se considera como un parámetro de ruido.

En contraste, cuando tenemos modelos no-paramétricos, el interés principal yace en estimar el parámetor infinito-dimensional. Eso hace que en los modelos semi-paramétricos la estimación se vuelve signficativamente más difícil que en los no-paramétricos.

Un ejemplo muy conocido de un modelo semi-paramétrico es el modelo de regresión de Cox (modelo de riesgos proporcionales de Cox). Si nos interesamos en estudiar el tiempo T que transcurre hasta que cierto evento como el de la muerte debido a cáncer o la falla de un bombillo de luz, el modelo de Cox especifica la siguiente función de distribución para T:

$$F(t) = 1 - \exp\left(-\int_0^t \lambda_0(u)e^{\beta' x}du\right),$$

donde x es un vector conocido de los datos, mientras que β y $\lambda_0(u)$ son parámetros desconocidos. Entonces tenemos que en este caso, como mencionado anteriormente para los modelos semi-paramétricos, que $\theta = (\beta, \lambda_0(u))$.

Aquí β es finito-dimensional es y de interés, pues contiene la información de cómo se combinan las influencias de nuestros datos para predecir el tiempo de interés T; mientras que $\lambda_0(u)$ es una función no-negativa desconocida del tiempo (conocida como la función de base de riesgo) y es usualmente considerada un parámetro de ruido que queremos evitar.

El conjunto de posibles candidatos para $\lambda_0(u)$ es infinito-dimensional. Este es usualmente el problema que vuelve difícil el análisis y estimación de esta distribución. Para corregir dicha situación usualmente se extrae información de un modelo teŕico u otros estudios periféricos para obtener características de la función $\lambda_0(u)$ para facilitar estimación de $F(t)$.

2.1.4 Análisis envolvente de datos

- Análisis envolvente de datos (DEA, por sigla en inglés) es un método no-paramétrico en investigación de operaciones y economía para estimar las fronteras de producción. Es utilizado empíricamente para medir la eficiencia de productividad sobre las unidades de decisión (DMU: decisión making units).

  A pesar de que DEA tiene una fuerte ligadura hacia la teoría de producciones en economía, la herramienta puede ser utilizada para sondeo de eficiencia en el administración de operaciones, donde un conjunto de medidas/indicadores es seleccionado como un punto de referencia del rendimiento de la manufactura y servicio de operaciones.

  En cuando a establecer un punto de referencia, los DMUs eficientes, tal y como se definen en DEA, puede que no sean de la forma de una frontera de producción, si no que se acerquen ás a una frontera de óptima operación (Cook, Tone y Zhu, 2014).

Algunas ventajas de DEA son:

1. No hay necesidad de específicar previamente la forma matemática de la función de producción.
2. Está verificado de ser útil y efectivo en descubrir relaciones entre parámetros que no son fácilmente descubiertas por otras metodologías.
3. Capaz de manejar múltiples variables de entrada y variables de salida.
4. Capaz de ser utilizado con cualquier tipo de medición; es decir, es independiente de cómo se obtuvieron los datos de salida y entrada.
5. Las posibles fuentes de ineficiencia pueden ser analizadas y cuantificadas individualmente respecto a cada unidad evaluada.

Algunas desventajas de DEA son:

1. Los resultados son altamente sensibles a la selección de variables de entrada y variables de salida. (Berg, 2010)
2. No es posible verificar si estamos utilizando el conjunto óptimo de parámetros. (Berg, 2010)

2.1.5 Métodos libres de distribución

Los métodos no-paramétricos libres de distribución son métodos para realizar pruebas de hipótesis estadísticas que, a diferencia de los métodos paramétricos, no necesitan realiar ninguna asunción sobre la forma o estructura de las distribuciones involucradas, si aun caso se hace mención a ellas.

Ejemplos de ellas son:

- Análisis de similitudes
- Test de Anderson–Darling: Prueba si una muestra pertenece o no a una distribución dada con la cual se compara mediante una "distancia" en el espacio de todas las distribuciones.
- Mpetodos de bootstrap estdísticos: Estiman la exactitud y la distribución muestral de un estadístico.
- Test Q de Cochran: Prueba/verifica si k tratamientos en un diseño aleatorio de bloques con posibles salidas binarias 0/1 tienen efectos idénticos.
- Kappa de Cohen: Mide la concordancia entre elementos categóricos.
- Análisis de doble vía de Friendman para la varianza por ranks: Verifica si k tratamientos en un diseño aleatorio de bloques tienen efectos idénticos
- Kaplan–Meier: Estima la función de supervivencia en datos sobre tiempo de vida. Modela además el senso.
- Tau de Kendall: Mide la dependencia estadística entre dos variables.
- W de Kendall: Mide interrelación de dos variables de tipo binaria 0/1.
- Test de Kolmogorov–Smirnov: Verifica si dos distribuciones son iguales utilizando ranks.
- Análisis de una vía para la varianza por ranks de Kruskal–Wallis: Verifica si dos muestras independientes fueron extraidas de una misma distribución pre-especificada.

- Test de Kuiper: Prueba si dos muestras fueron sacadas de cierta distribución pre-especificada, sensible a variaciones cíclicas como días o semanas.
- Test de Logrank
- Test de ranking de suma de Mann–Whitney U o Wilcoxon: Verifica si dos muestras fueron extraidas de una misma población en base a una comparación con hipótesis alternativa
- Prueba de McNemar: Prueba si, en tablas de contingencia de 2×2 con una característica docotómica y un par de sujetos, las filas y columnas marginales son iguales.
- Test de la media: Verifica si la media de dos muestras es igual.
- Test de permutación de Pitman: Una prueba de significancia estadística que provee el valor-p exacto al examinar todas las etiquetas de nuestros datos.
- Productos de ranks: Detecta diferencias expresadas en los genes en experimentos de microarreglos replicados.
- Test de Siegel–Tukey: Se utiliza para probar diferencia de escalar entre dos muestras.
- Prueba de signo: Prueba si un par ligado de muestras son extraidas de poblaciones con la misma media.
- Coeficiente de correlación de ranks de Spearman: Mide la dependencia estadística entre dos variables utilizando una función monotónica.
- Test de cuadrados de rank: Prueba si hay igualdad de varianzas entre dos o más muestras.
- Test de Tukey–Duckworth: Prueba la igualdad de dos distribuciones utilizando ranks.
- Test de Wald–Wolfowitz: Verifica si elementos de una sucesión son mutuamente independientes/aleatorios.
- Test de signos de Wilcoxon: Verifica si dos muestras ligadas fueron extraidas de poblaciones con diferente ranking de medias.

Capítulo 3

SECCIÓN DE REFERENCIAS

"La poesía está por todas partes, pero llevarla al papel es, por desgracia, más complicado que verla."

-Vincent van Gogh

Referencias

- Stuart A., Ord J.K, Arnold S. (1999), Kendall's Advanced Theory of Statistics: Volume 2A—Classical Inference and the Linear Model, sixth edition, §20.2–20.3 (Arnold).
- Conover, W.J. (1999), Çhapter 3.4: The Sign Test", Practical Nonparametric Statistics (Third ed.), Wiley, pp. 157–176, ISBN 0-471-16068-7
- Sprent, P. (1989), Applied Nonparametric Statistical Methods (Second ed.), Chapman & Hall, ISBN 0-412-44980-3
- Scott, David W. (1992). Multivariate Density Estimation: Theory, Practice, and Visualization. New York: John Wiley.
- Sturges, H. A. (1926). "The choice of a class interval". Journal of the American Statistical Association. 21 (153): 65–66. doi:10.1080/01621459.1926.10502161. JSTOR 2965501.
- e.g. § 5.6 "Density Estimation", W. N. Venables and B. D. Ripley, Modern Applied Statistics with S (2002), Springer, 4th edition. ISBN 0-387-95457-0.
- A. Colin Cameron, Dept. of Economics, Univ. of Calif. - Davis, .ᴱˣCEL Univariate: Histogram".
- Online Statistics Education: A Multimedia Course of Study. Project Leader: David M. Lane, Rice University
- Doane DP (1976) Aesthetic frequency classification. American Statistician, 30: 181–183
- Scott, David W. (1979). "On optimal and data-based histograms". Biometrika. 66 (3): 605–610. doi:10.1093/biomet/66.3.605.
- Freedman, David; Diaconis, P. (1981). .ᴼn the histogram as a density estimator: L2 theory"(PDF). Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 57 (4): 453–476. CiteSeerX 10.1.1.650.2473. doi:10.1007/BF01025868.
- Wasserman, Larry (2004). All of Statistics. New York: Springer. p. 310. ISBN 978-1-4419-2322-6.
- Stone, Charles J. (1984). "An asymptotically optimal histogram selection rule"(PDF). Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer.
- Shimazaki, H.; Shinomoto, S. (2007). .ᴬ method for selecting the bin size of a time histogram". Neural Computation. 19 (6): 1503–1527. CiteSeerX 10.1.1.304.6404.
- Jack Prins; Don McCormack; Di Michelson; Karen Horrell. "Chi-square goodness-of-fit test". NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH. p. 7.2.1.1. Retrieved 29 March 2019.
- Moore, David (1986). "3". In D'Agostino, Ralph; Stephens, Michael (eds.). Goodness-of-Fit Techniques. New York, NY, USA: Marcel Dekker Inc. p. 70. ISBN 0-8247-7487-6.