Cambios a la tesis *Un modelo Bayesiano y no paramétrico de regresión sobre cuantiles*

Omar Pardo

Febrero 2018

Generales

- Corrección de *typos*.
- Implementación de sugerencias para la redacción.

1. Introducción

- Exposición más clara de las limitaciones que pueden tener los modelos de regresión a la media, independientemente de si son lineales o no lineales, paramétricos o no paramétricos.
- Mayor contexto histórico de los modelos de regresión a la mediana.
- Aclaración que la definición expuesta de cuantil es informal, intentando evitar tecnicismos en esta sección.

2. Paradigma Bayesiano

- Énfasis en que la aleatoriedad dentro de este paradigma debe ser vista como una medida de la incertidumbre del modelador acerca del valor real.
- Correcciones de notación y hacer explícita la dimensión de distintos parámetros.

3. Modelos de regresión

- Aclaración que tanto el modelo de regresión a la media, como a la mediana, dan una especificación completa de $f(y|x)$ en el caso paramétrico, y en lo que difieren es en los parámetros a calcular.
- Cambiar el error Normal del modelo general de regresión a la media, por únicamente la restricción que $\mathbb{E}[\varepsilon] = 0$.
- Definición más general y formal de la función cuantil.
- Gráfica de la distribución asimétrica de Laplace y comentarios de cómo cambia, conforme varían los parámetros.

4. Especificación paramétrica

- Énfasis en que los ejemplos de la sección *Motivación* buscan por un lado mostrar que mediante transformaciones de variables se puede llegar al modelo tradicional, aún cuando en un principio pareciera tener una forma distinta. Pero por otro lado, esto abre la idea de que habrá relaciones que no podrán ser expresadas de manera lineal en los parámetros, y de ahí la importancia de los métodos no paramétricos.

• Cambio en la definición de *proceso Gaussiano*, para hacer explícito que es un proceso estocástico continuo, es decir, es un conjunto infinito de variables aleatorias. Cuando se toma un subconjunto de ellas, entonces su distribución es Normal.

• Aclaración de a qué se refiere la notación de cada elemento de la definición formal de los *procesos de Dirichlet*.

• Aclaración que G es una distribución fija, pero los *procesos de Dirichlet* son útiles para reflejar la incertidumbre acerca de su valor.

• Sustitución del término *con probabilidad igual a 1* por *casi seguramente*.

• Explicación que para el modelo de mezclas infinitas no se usa una distribución Multinomial, sino la distribución resultante de tomar el límite cuando una Multinomial tiende a tener infinitas categoriás.

5. Modelo GPDP

• Comentario acerca de que p también podría ser incorporado al modelo como un parámetro a estimar, lo cual probablemente mejoraría la aproximación de la distribución condicional $y|x$. Sin embargo, a pesar de que el modelo ajusta teóricamente toda la distribución, el parámetro de mayor interés para esta tesis es $f_p(x) = q_p(y|x)$, una vez que el modelador ya predefinió que el cuantil *p-ésimo* es aquel que le interesa entender.

• Descripción más detallada de cómo se obtiene la distribución condicional posterior de f_p, dentro del simulador de Gibbs. Se expone explícitamente la verosimilitud después de usar una variable aleatoria auxiliar, para aclarar por qué la distribución posterior es una Normal truncada, y el motivo por el que la varianza posterior es la misma que la inicial.

• Modificación en las heurísticas utilizadas para definir los hiper-parámetros iniciales, de forma que en ningún caso se requiera observar los datos.

• Comentario acerca de que alguna medida de bondad de ajuste estadísticamente robusta queda fuera del alcance de esta tesis.

6. Aplicaciones

• Reescritura completa de este capítulo.

• Estimación de la predicción de diversos cuantiles, usando tanto el modelo GPDP, como el modelo tradicional de regresión a la media. Se añadió un apéndice para explicar cómo se puede aproximar la distribución posterior predictiva de un cuantil en específico, utilizando el modelo tradicional de regresión a la media.

• Gráficas del ajuste de ambos modelos, cálculo del error cuadrático medio y correlación al cuadrado de la mediana de la predicción, respecto al valor real de cada cuantil; así como el cálculo de cuántos valores reales cayeron dentro del intervalo de probabilidad predictivo al 95 %.

• Simulación de un conjunto de datos que cumple los supuestos de la regresión tradicional a la media, así como otros conjuntos que violan alguno de dichos supuestos, para comparar el ajuste de ambos modelos. Discusión acerca de los resultados obtenidos.

• Comparación del tiempo que tardaron en correr ambos modelos, tanto en el ajuste, como en la predicción.

• Mención de las características de la cadena de Markov usada, por ejemplo, cuántas observaciones iniciales se *quemaron* y cuánto se *adelgazó*. Análisis de los resultados obtenidos al correrle los diagnósticos, una vez ajustado el modelo.

7. Conclusiones y trabajo futuro

• Reconocimiento de la complejidad para introducir el conocimiento previo del modelador en el modelo, aún cuando es Bayesiano, debido a la estructura jerárquica.

• Comentarios acerca de lo que representó desarrollar el paquete en R.

• Explicación de a qué me refiero cuando menciono que el modelo asigna la misma importancia a todas las variables, así como una mayor descripción del parámetro dinámico de rango para el cálculo de la correlación entre observaciones, que sugiero introducir.

• Exposición del área de oportunidad que existe para desarrollar una medida de bondad de ajuste estadísticamente robusta.