Instituto Tecnológico Autónomo de México

Estadística Bayesiana

Una percepción Bayesiana ante el fútbol

Sebastián Serra Peña 180760

5 de diciembre de 2022

1. Introducción

En palabras de Lindley, la teoría y la práctica van de la mano. Sin práctica la teoría está sub-desarrollada y sin la teoría, la práctica es subjetiva e inconsistente es por eso que en este texto se hablará de un ejemplo práctico muy personal desde un punto de vista teórico basado enla estadística Bayesiana.

Para dar un contexto inicial es importante definir que es lo que sostiene a la estadística Bayesiana y este pilar es la idea de que la única forma de describir la incertidumbre es por medio de la probabilidad. La incertidumbre se encuentra en todos lados y la probabilidad nos ayuda a enfrentarnos ante esta dificultad, originalmente mediante ensayos y prubas repetidas una y otra vez. Sinn embargo con el paso del tiempo esta misma probabilidad trae consigo nociones matemáticas que permiten solucionar problemas relativamente abstractos de una manera muy formal e incluso basada en axiomas.

Ahora vale la pena meternos al mundo del fútbol, mundo que, como ha probado esta última copa del mundo (Qatar 2022), muestra una tremenda incertidumbre con respecto a los resultados de ciertos partidos. Si no me creen pregúnten a las selecciones de Japón, Arabia, Korea o Marruecos si la vieron venir. A pesar de la enorme cantidad de datos que ahora se tiene en nuestra disposición aún existe una gran incertidumbre ante lo desconocido y sería interesante plantear una descripción del conocimiento previo de uno sobre algún elemento desconocido del fútbol en presencia de los datos. Esta incertidumbre se ve reflejada mediante la probabilidad de lo desconocido, llamémosle θ dados los datos D y un conocimiento previo, un histórico H. Entérminos de probabilidad a esto lo escribiremos así:

$$P(\theta \mid D,\ H)$$

2. Problemas de Decisión

Al final determinar a esta θ implica un problema de decisión ante la elección de un posible enorme número de casos. Así que ahora en particular hablaremos del problema que se viene encaminado a nosotros. ¿Qué *theta* usamos? ó más aún ¿Qué *theta* decidimos usar? esto es un problema de decisión y para esto notamos que requerimos definir no sólo un árbol de posibles decisiones, sino también una buena cantidad de eventos inciertos qeu tendráncierta probabilidad de ocurrencia y cada uno de ellos traerá consigo una consecuencia a afrontar. Estás consecuencias han de ser evaluadas y puntuadas para ver que consecuencia preferimos más. Para lo anterior se requiere de coherencia, es decir; sentencias planteadas que no se contradicen entre ellas. Por el problema anterior es que se plantean los axiomas de coherencia en la estadística bayesiana, mismos axiomas que provienen de Ramsey y Savage por medio de la "Teoría de Decisiónz de Finetti por Reglas de Puntuación". No entraré en detalle a estos axiomas, pero cualquier lector de aquí es libre de buscarlos y empaparse un rato en ello.

La estadística Bayesiana trae consigo un método muy estructurado para tomar una decisión, esto es:

1. ¿Qué es incierto en el problema? Llámalo θ

2. ¿Qué sabes sobre θ? Llámalo D, específico y H en general.

3. Calcula $P(\theta \mid D,\ H)$

La pregunta es ahora ante el como calcular esa probabilidad y la respuesta es mediante el uso de reglas de probabilidad, pero para ser más claros diremos que se calcula con plena creatividad y amor al arte (rigurosamente claro está).

Para terminar esta sección diremos que toda decisión conlleva consecuencias tras tomarla, al valor de esta le llamaremos utilidad y nuestro propósito es calcular la utilidad esperada de cada decisión y aquella que sea más grande será la decisión a tomar.

3. Caso fútbol

Existen una infinidad de cosas inciertas en el fútbol, pero en este caso plantearé el problema del análisis de un rival cuyos únicos datos conocidos son sus jugadores y las estadísticas de ellos. El análisis táctico de un rival se lleva mediante un entendimiento de lo que se denomina parado táctico. Este parado es muy geométrico y a pesar de que la mayoría de los espectadores casuales están familiarizados con alineaciones como el 4-3-3 o el 4-4-2 a esto no me refiero. Estas alineaciones son un mero posicionamiento inicial, pero al iniciar el juego cada jugador tiene sus roles y en gran parte de las situaciones este dibujo táctico cambia drásticamente de acuerdo a como se llevan entre los jugadores o más particularmente, como maximizan su trayectoria para el gol o su presión para robar unh balón, entre mil opciones más. Si nos centramos en el poder mantener la mayor posesión del balón, entonces los datos importantes a resaltar son la probabilidad con las que un pase es acertado entre compañeros.

Si tras un partido tu calculas todos los pases dados y dibujas una línea cuando se hicieron estos pases obtienes algo parecido a esto:

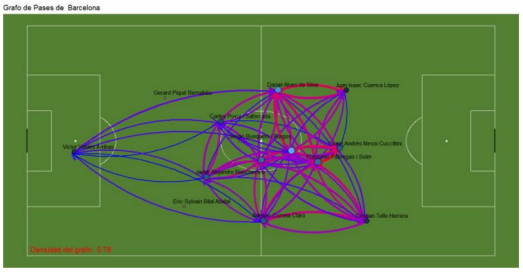


Fig. 3: Grafo de pases de Barcelona (1) [3]

Lamentablemente esto no lo tenemos previo a un partido, sería ilógico, entonces habría que inferir las conexiones, he ahí nuestro desconocido θ. Mediante datos D queson las estadísticas de pase de cada jugador y aparte el conocimiento previo H de alguien que conozca de fútbol, no se, tal vez un entrenador. La duda está entonces en la creación de un modelo basado en estadística Bayesiana para puntuar la probabilidad de que exista una conexión entre dos jugadores, castigando acorde a sus estadísticas. de aquí otro problema y es la cantidad de datos, que si bien puede aparentar ser enorme, es complicado asegurar una

cantidad óptima de datos que actualizarán la probabilidad tras añadirlos, pues si bien Messi jugó muchos años con el Barcelona, no son los mismos jugadores en vestuario, ni es el mismo Messi (Nótese que soy hincha del Barcelona). ¿Cómo se la cantidad óptima de datos a tomar en una muestra para maximizar mi utilidad? Esta respuesta nos la da Jose M. Bernardo en su artículo "Statistical Inference as a Decision Problem: The Choice of Sample Size".

4. El tamaño de muestra

Sea $d \epsilon D$ una decisión posible a tomar cuyas consecuencias dependen de $\theta \epsilon \Theta \subset \mathbb{R}$ (desconocido) con una probabilidad de ocurrencia (a priori) $P(\theta \mid H)$ que por facilidad escribiremos $P(\theta)$. Si se toma una muestra de tamaño n con $X_{(n)} = \{x_1, ..., X_n\}$ que se relacionan con *theta* mediante $P(X \mid \theta)$ y la utilidad del experimento si al escoger a d está dada por:

$$u(X_{(n)},\ d,\ \theta) = g(d,\ \theta) - c(n,\ X_{(n)})$$

Siendo la función g la ganancia y c la función costo (ambas con mismas unidades). Entonces la coherencia implica un tamaño de muestra óptimo de n que maximiza a la utilidad de la siguiente forma:

$$u^*(n) = \int_{X^n} P(X_{(n)}) sup\{\int_\Theta g(d,\ \theta) P(\theta \mid X_{(n)})\ d\theta\} - c^*(n)$$

El resultado al que llega Bernardo está basado en la utilidad esperada por un tamaño de muestra n fijo dado por:

$$u^*(n) = v\ I(n) - c^*(n)$$

La función $I(n)$ es:

$$\int_{X^n} P(X_{(n)}) \int_\Theta P(\theta \mid X_{(n)})\ log_2(\frac{P(\theta \mid X_{(n)})}{P(\theta)})\ d\theta\ dX_{(n)}$$

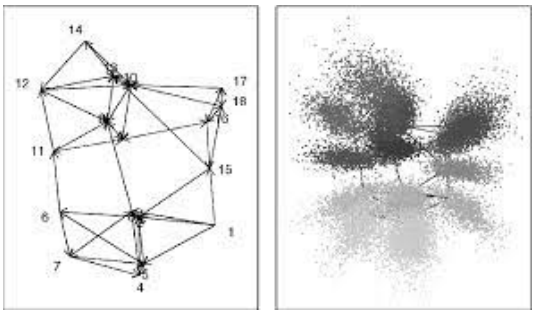
A su vez la variable v es un tanto menos caótica. Esta es el valor esperado del tomador de decisiones dela respuesta ante la pregunta sobre θ, es decir el valor esperado de una pequeña información sobre lo desconocido.

de esta forma se sabe que si se espera tener cierta probabilidad de conocer lo desconocido entonces al dar los datos aquí se arroja una n que es el tamaño de muestra necesario para cumplir con lo requerido.

5. El Modelo

Para definir el modelo mediante el cual decidiremos si existe una conexión o no entre jugadores requerimos de un mapeo de puntos, es decir, tomar la muestra de datos que tenemos y usarlos en una función tal que los arrojé en un plano euclidiano (coordenado x,

y). Particularmente la idea es que mientras más cerca se encuentren los puntos (jugadores), más probable es que exista una conexión entre ellos y de esta forma infieres el dibujo táctico del rival, sin haber jugado con ellos aún.



En este punto vale la pena hablar sobre los modelos, pues un Bayesiano ve al modelo como una forma de especificar las probabilidades esenciales a su método de estudio de la incertidumbre. Sin embargo, muchas veces el trabajo de un estadístico es el de estudiar los datos en sí, la muestra. El modeloplantea una solución pero mediante la muestra es que definiremos a las consecuencias, o como ya les llamamos, utilidades. Para entrenar a una persona informada capaz de definir a estas probabilidades y más aún a la utilidad, es necesario el uso de reglas de puntaje que ayudarán a asociar una utilidad a las consecuencias que trae consigo lo desconocido.

6. Conclusión

¿Cómo que la conclusión? ¿Dónde está el modelo de fútbol? ¿Tanto para nada? Pues sí, esto es lo divertido, esto no quiero que se tome como una iniciación a la estadística Bayesiano, pero si como una motivación para emprender este camino tan bello y confuso. No doy el modelo, pues es mi trabajo de tésis y no es que no quiera compartirlo, sino que no lo he terminado. Sin embargo, creo que esto permite dar ideas y ampliar la visión sobre la conjunción entre la práctica y la teoría bayesiana. Mi pasión es el fútbol y las matemáticas y de ahi esta unión, pero esto es aplicable en miles de lugares, desde medicina hasta leyes, desde finanzas hasta ingeniería. Al final hay que entender que incertidumbre, la duda está en todos lados de la vida, desde una ruptura dolosa hasta ver quién será nuestro próximo presidente (favor de checar el conteo rápido de Manuel Mendoza y Luis E. Nieto, lo dejo en referencias). Esto es tal y como dijo Lindley, mismo que ya cité unas mil millones de veces aqui:

"**El paradigma Bayesiano trata la incertidumbre y su única herramienta es la coherencia expresada a través de leyes de probabilidad.**" *(D. Lindley, Theory and Practice of Bayesian Statistics, junio 1983)*

7. Referencias

Lindley, D. (1983). Theory and Practice of Bayesian Statistics. The Statistician, volumen(32).

Bernardo, J. (1997). Statistical Inference as a Decision Problem: The Choice of Sample Size. The Statistician, volumen(46).

Mendoza, M., Nieto-Barajas, L. (2013). Quick counts in the Mexican preidential elections: A Bayesian approach. Elsevier, volumen(43).

Serra, S. (2022, 12 de septiembre). Fútbol y Grafos/Estadística. Tiro al Ángulo. https://www.tiroalangulo.com/post/futbolygrafos