ESTADÍSTICA BAYESIANA

TABLA DE CONTENIDO

MOTIVACIÓN DEL ENFOQUE BAYESIANO

La Estadística Bayesiana es un área específica dentro del campo de la Estadística, y hoy en día es la pieza central de muchas aplicaciones en Ciencia de Datos (Data Science) y Aprendizaje Automático (Machine Learning). Por ejemplo este enfoque es muy utilizado en Medicina, por ejemplo en muchas estimaciones relacionadas con afecciones, cáncer incluso virus como el coronavirus. También es muy utilizado en marketing online a la hora de comparar anuncios, páginas webs, ver rendimientos, tasas de clics o tasas de conversión, y sobre todo en aquellos problemas que se llaman “problemas online” que se caracterizan por estar “en línea” en el sentido que los datos no son estáticos sino que se van actualizando con el tiempo. Esto está relacionado con lo que se conoce como prueba A/B donde se busca decidir entre dos opciones A y B para elegir la opción que maximiza algún beneficio. Uno de los casos más relevantes es el de los anuncios publicitarios de Facebook, cuando quieren decidir qué anuncio mostrar a los usuarios para maximizar su propio beneficio y el de los anunciantes. Pero también lo utilizan los periódicos que buscan decidir entre diferentes titulares, los minoristas que buscan decidir entre diferentes embalajes, las farmacéuticas para evaluar diferentes tratamientos, las aerolíneas para decidir entre diferentes precios y, por supuesto, las plataformas publicitarias para decidir entre diferentes anuncios, en todos estos casos se utiliza mucho el enfoque bayesiano. Sus métodos se caracterizan porque la "evidencia" sobre lo "verdadero" se expresa en términos del grado de creencia o, más específicamente, en términos de probabilidades bayesianas. Todo se deriva de la interpretación del concepto de probabilidad.

Por ello es necesario comenzar por la parte más básica, por los fundamentos sobre los que se rige esta teoría o este enfoque. Así que comenzaremos dando respuesta a preguntas básicas pero muy profundas como por ejemplo qué significa la probabilidad. Vamos recordar todos los conceptos básicos necesarios antes de entrar en detalle y de lleno en el enfoque bayesiano, y vamos a ir razonando en el contexto de eventos o ejemplos sencillos como tiradas de dados, pero también eventos muy concretos y útiles como las carreras de caballos para entender luego el por qué del enfoque bayesiano. De esta manera vamos a poder tener la base necesaria para entender después los métodos más complejos de la inferencia bayesiana y del enfoque bayesiano del análisis de datos.

CONCEPTOS BÁSICOS

DISTRIBUCIÓN CONJUNTA, MARGINALES Y CONDICIONALES

Vamos a suponer que tenemos dos variables aleatorias A y B, las distribuciones marginales serían:

$$p(A),\ p(B)$$

La distribución conjunta la denotaremos como $p(A, B)$. La coma es una forma abreviada de decir “y”.

Y por último tenemos la distribución condicional, a la que nos referimos como $p(A|B)$ o $p(B|A)$.

La barra vertical es el símbolo que usamos para representar a una condición, condicionado a algo. Esto se lee P de A, dado B; y P de B, dado A. La distribución conjunta es la más general porque es a partir de esta distribución que podemos calcular todo lo demás.

Podemos calcular la distribución marginal si tenemos la distribución conjunta, de la siguiente manera:

$$p(A) = \sum_{B} p(A, B)$$

$$p(B) = \sum_{A} p(A, B)$$

También podemos calcular la distribución condicional usando la conjunta y la marginal:

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A,B)}{\sum_A p(A,B)}$$

$$p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)}$$

Como sabemos que la marginal puede calcularse con la conjunta realmente solo necesitamos la conjunta para calcular la condicional, como vemos en la última igualdad de ambas ecuaciones.

Una cosa interesante que podemos preguntar es si no tenemos la distribución conjunta, sino que solo tenemos la condicional y la marginal. ¿Podemos calcular la otra condicional? La respuesta es sí. Tengamos en cuenta que si tomamos la regla de las probabilidades condicionales que mencionamos anteriormente, simplemente podemos reorganizar los símbolos para mostrar que:

$$p(B|A) = \frac{p(A,B)}{p(A)} = \frac{p(A,B)}{\sum_B p(A,B)} = \frac{p(A|B)p(B)}{\sum_B p(A|B)p(B)}$$

Es decir, podemos calcular la distribución conjunta usando la condicional y la marginal. De hecho, acabamos de llegar al Teorema de Bayes pero considerando que esto son distribuciones.

Ahora bien, aquí hemos estado hablando de sumas, esto es porque estamos asumiendo que A y B son variables aleatorias discretas. Pero, ¿y si fueran variables aleatorias continuas? Llamemos a estas variables X e Y. En este caso, la p minúscula no se refiere a la probabilidad, sino a la función de densidad de probabilidad. Normalmente se usa f minúscula y todo se sigue cumpliendo igual excepto que en vez de sumas vamos a tener integrales:

Densidad conjunta: $f(x,y)$

Densidades marginales: $f(x),\ f(y)$

$$f(x) = \int f(x,y)dy$$

$$f(y) = \int f(x,y)dx$$

Condicional:

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x,y)}{\int f(x,y)dx}$$

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{f(x,y)}{\int f(x,y)dy}$$

Además, la regla de Bayes también se expresa de forma análoga:

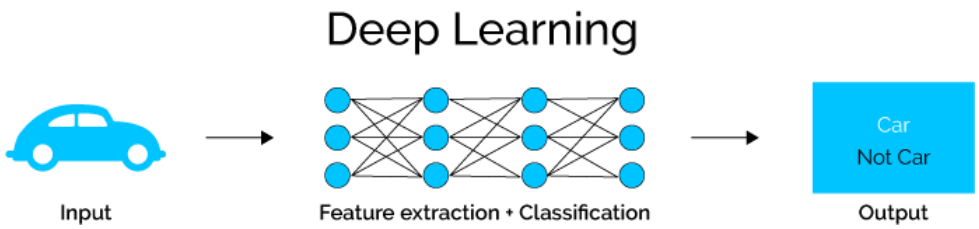
$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

$$f(y|x) = \frac{f(x|y)f(y)}{\int f(x|y)f(y)dy}$$

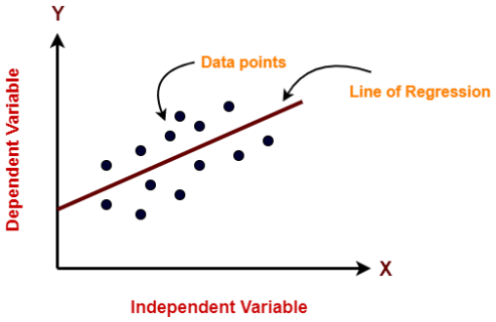
Simplemente reemplazamos la suma con una integral sobre la variable aleatoria relevante.

MÁXIMA VEROSIMILITUD: DISTRIBUCIÓN DE BERNOULLI

La estimación de máxima verosimilitud es una técnica de estadística que consiste en lo siguiente. Imaginemos que hemos recopilado un montón de datos de un experimento. Y que nos gustaría ajustar un modelo a esos datos. Por lo general, tal modelo está determinado por unos parámetros. Nuestro trabajo entonces es encontrar los mejores parámetros para que podamos modelar los datos que recopilamos lo más cerca posible de la realidad. Por ejemplo si tomamos cualquier modelo de Machine Learning, como un modelo para clasificar datos, o incluso las mismas redes neuronales, la parte de aprendizaje profundo es solo el acto de encontrar los pesos de la red neuronal que mejor explican el conjunto de datos.



O en un modelo de regresión lineal simple sería encontrar los valores de la pendiente y el intercepto de la recta que mejor representa la relación lineal entre los datos.



Por lo tanto, aprender es solo un término elegante para referirnos a lo que realmente se está haciendo que es el ajuste del modelo. Un modelo que depende de ciertos parámetros. Entonces, vamos a empezar desde algo sencillo, por ejemplo la distribución de Bernoulli. Uno de los modelos más fundamentales.

El ejemplo clásico para entender la distribución de Bernoulli es el ejemplo del lanzamiento de la moneda. Entonces, por ejemplo, podemos decir que intuitivamente la probabilidad de sacar cara es un 50% y de sacar cruz igual, es lo que en principio todos pensamos. Pero esto no tiene por qué ser así, puede ser que la moneda no sea perfecta y que estas probabilidades no sean iguales, por ejemplo podemos decir que la probabilidad de sacar cara es 0.6 y la probabilidad de sacar cruz es 0.4. Este experimento de lanzar la moneda sigue una distribución de Bernoulli.

Por supuesto, vamos a recordar cómo se define esta distribución de manera matemática. Es decir, ¿cuál es la ecuación que describe cómo varía la probabilidad asociada de una variable aleatoria con distribución de Bernoulli? Esta es una distribución discreta que por tanto describe a una variable aleatoria discreta. La ecuación o modelo que estamos buscando es la función de probabilidad. En el caso de la Bernoulli la ecuación que describe el comportamiento de la variable aleatoria, que también, como es discreta, se conoce como función de masa de probabilidad, esa ecuación es la siguiente:

$$p(x) = \theta^x(1-\theta)^{1-x}$$

En este caso, x solo puede tomar los valores cero o uno. Si estamos pensando en el ejemplo de lanzar una moneda, entonces generalmente diríamos que cero es cruz y uno es igual a cara. Theta (θ) se llama parámetro y es el único parámetro de esta distribución. Que casualmente, en esta distribución θ coincide con la probabilidad de que X sea igual a uno:

$$p(x = 1) = \theta^1(1-\theta)^{1-1} = \theta$$

Como ejercicio, puedes verificar que la probabilidad de que X sea igual a cero es igual a uno menos theta y, por lo tanto, la probabilidad de que X sea igual a uno más la probabilidad de que X sea igual a cero es igual a uno, como debería ser:

$$p(x = 0) = 1 - \theta$$

$$p(x = 0) + p(x = 1) = 1$$

El siguiente paso es adentrarnos en el problema de la máxima verosimilitud, o *maximum likelihood*.

Así que supongamos que hemos repetido este experimento de lanzar una moneda muchas veces:

$$data = \{\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4, \dots, \mathrm{x}_\mathrm{n}\}$$

Ahora queremos saber cuál es el mejor valor de Theta para describir estos datos que hemos recopilado. En estimación de máxima verosimilitud, comenzamos escribiendo esto como la función L(theta) que también a veces la puedes ver escrita como p(data|theta):

$$L(\theta) = p(data|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Además, esto va a ser igual al producto de las probabilidades de cada x_i en dependencia del parámetro θ. Donde cada x_i, cada elemento muestral, es un lanzamiento de la moneda, y tiene distribución de Bernoulli porque puede ser 0 o 1 (cruz o cara).

Entonces vamos a sustituir esas funciones de probabilidad en nuestra expresión:

$$L(\theta) = p(data|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

Esto no es solo una ecuación matemática. Tiene significado. Esta es la probabilidad de observar los datos que observamos, suponiendo que cada lanzamiento de la moneda fuera iid (independiente e idénticamente distribuido).

Es importante comprender bien la función de verosimilitud, y entender de qué realmente es una función y de qué no lo es. Por ejemplo, muchas veces las personas ven estas X y automáticamente piensan que como usualmente usamos X como variable, pues que X es la variable en este caso. Pero eso es incorrecto. En este caso, X no es una variable. Las x_i son los valores de nuestro experimento. Que como recordaremos, estos valores son o ceros o unos. En realidad, aquí la variable es θ que es el valor que estamos tratando de resolver o de encontrar, porque es el parámetro de nuestra distribución.

Entonces imagina que tenemos tres resultados del lanzamiento de la moneda, cara, cruz y cara:

$$\mathrm{x}_1 = 1, \mathrm{x}_2 = 0, \mathrm{x}_3 = 1$$

Luego los incluiríamos en la ecuación para la verosimilitud y obtendríamos lo que vemos aquí que es obviamente una función de theta, el parámetro:

$$L(\theta) = \prod_{i=1}^{3} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^1(1-\theta)^1\theta^1$$

Bien, ahora recordemos lo que estábamos haciendo, al principio hablamos del problema de la máxima verosimilitud, ¿por qué se llama así? Porque el propósito es que ahora queremos encontrar qué valor de theta maximizará a esta función de verosimilitud. Estamos preguntando qué valor de theta hace que los datos que recopilamos sean más probables. En otras palabras, ¿qué valor de θ maximiza la probabilidad?

Obviamente, si tenemos 100 caras y cero cruces, entonces decir que la probabilidad de obtener cara es del cinco por ciento no tendría mucho sentido. En cambio, es más probable que la probabilidad de cara sea más cercana al cien por ciento, intuitivamente. Entonces, la pregunta es, utilizando la función de verosimilitud, ¿podemos maximizarla para encontrar un valor de θ? Bien, para hacer esto lo que tenemos que ver es ¿cómo maximizamos una función? Para lo cual vamos a hacer uso del cálculo matemático.

Para encontrar el máximo de la función de verosimilitud $L(\theta)$ tenemos que encontrar la derivada de $L(\theta)$ con respecto a θ. Podemos resolver esto descubriendo qué valor de theta hace que esta derivada sea cero.

Antes de hacer esto, debemos mencionar que la mayoría de las veces, es mejor tomar el logaritmo de la verosimilitud en vez de derivar directamente la verosimilitud. Es decir primero aplicar logaritmo a $L(\theta)$, esto será entonces la log-verosimilitud: $l(\theta)$. Y a esta $l(\theta)$ es a la que le aplicamos la derivada e igualamos a cero y despejamos θ. Esto se hace así porque por lo general, conduce a una expresión más simple para la derivada que es más fácil de igualar a cero y resolver. ¿Por qué funciona esta transformación, por qué se puede aplicar? Funciona porque el logaritmo es una función monótona creciente. Lo que significa que si encuentro un valor de θ que maximiza el logaritmo de la verosimilitud $l(\theta)$, entonces este mismo valor de θ también maximizará a la función sin transformar $L(\theta)$.

Log-verosimilitud:

$$l(\theta) = \log\big(L(\theta)\big) = \log\left(\prod_{i=1}^{n} p(x_i|\theta)\right) = \sum_{i=1}^{n} (x_i \log(\theta) + (1 - x_i)\log(1 - \theta))$$

Derivamos $l(\theta)$:

$$\frac{dl(\theta)}{d\theta} = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1 - x_i)$$

Igualamos a cero la derivada anterior y despejamos θ:

$$\frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1 - x_i) \ = 0$$

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Entonces, ¿este resultado tiene sentido? Sí, de hecho, esto es exactamente lo mismo que decir el número de caras dividido por n, que, como saben, es como normalmente estimaríamos con la muestra la probabilidad de obtener una cara en el siguiente lanzamiento. ¿Por qué es esto? Porque obtener cara es el valor 1 y esta suma solo tendrá en cuenta a las caras (agregar un cero no modifica el resultado) esta es la suma de caras dividida entre el total de lanzamientos n.

EJEMPLO RATIO DE CLIQUEO

Un ejemplo de aplicación muy práctico es el ratio de cliqueo, muy conocido en áreas como publicidad en línea, análisis de páginas web, online marketing, e-mail marketing, etc.

El **ratio de cliqueo** (click through rate en Inglés), o también conocido como la tasa de clics puede considerarse como la probabilidad de que el usuario haga clic en un anuncio publicitario o en un enlace o en cualquier otra cosa en la que interesa que hagan clic.

Del mismo modo, la **tasa de conversión** puede considerarse como la probabilidad de que el usuario compre un producto, se suscriba al boletín de noticias o cualquier otra cosa que nos gustaría que hicieran.

Si lo piensan, estos eventos son binarios. O el usuario hace clic o el usuario no hace clic. O el usuario compra o el usuario no compra. Por lo tanto, estos se distribuyen como una Bernoulli. Sin embargo, es importante tener en cuenta que, por ejemplo en el caso de una moneda estamos acostumbrados a pensar que la probabilidad de éxito es de alrededor del 50 por ciento, pero la probabilidad de un clic o una conversión es mucho menor. Entonces, en estos ejemplos relacionados con publicidad en línea o medios digitales, veremos números muy pequeños. En promedio, un número esperable para las tasas de clics en anuncios será aproximadamente de 0.2 a 0.3 por ciento. Que es obviamente mucho menor que un 50%.

La tasa de clics se define como la cantidad de clics dividida por la cantidad de impresiones:

$$\text{CTR} = \% \frac{\text{clicks}}{\text{impresiones}}$$

Mientras que la tasa de conversión se define como la cantidad de personas que han realizado la acción deseada en la página web, dividida por la cantidad total de visitas a esa página web:

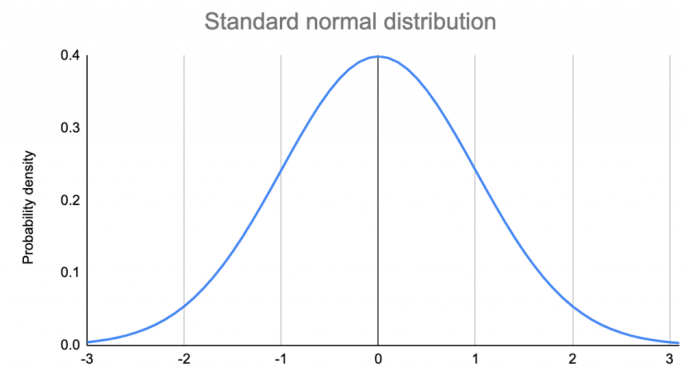
$$\text{Conversion Rate} = \frac{\text{Number of Conversions}}{\text{Total Visitors}} \times 100$$

Por supuesto, como ya hemos visto anteriormente en el tema de estimación de máxima verosimilitud, sabemos lo que estas tasas realmente se están estimando: la probabilidad de que los usuarios realicen estas acciones.

MÁXIMA VEROSIMILITUD: NORMAL

Anteriormente, analizamos la máxima verosimilitud en variables con distribución Bernoulli, variables aleatorias de tipo discreto que solo pueden tomar los valores cero y uno. Pero a veces tenemos variables aleatorias que pueden tomar más valores, o valores diferentes, continuos, por ejemplo, los números reales. Entonces, la pregunta es, ¿existe una distribución que sea apropiada para este contexto? ¿Y cuál

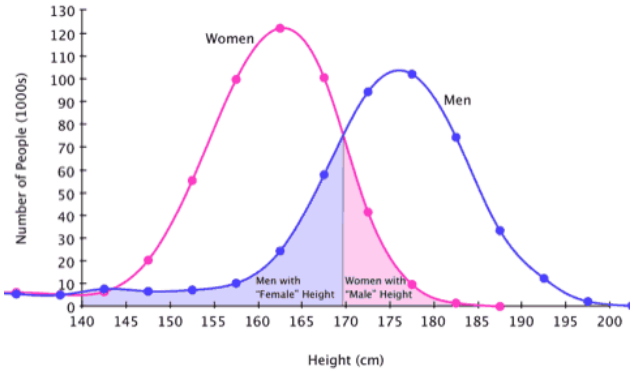
es esa distribución? Seguramente ya la conozcas porque es muy utilizada, se trata de la distribución Normal. Vamos a ver cómo usar el método de máxima verosimilitud en la distribución Normal, o Gaussiana (ambos nombres se refieren a la misma distribución). Antes de continuar, vamos a recalcar que la distribución Normal no es la única opción para variables aleatorias que toman valores reales. Otras distribuciones para variables aleatorias continuas muy conocidas también son la distribución t de Student, la distribución exponencial, la distribución Gamma, la chi-cuadrado, y muchas más. Por supuesto, siempre deberíamos elegir la distribución que creamos que mejor representa a los datos. Pero el proceso de resolución de problemas de máxima verosimilitud es el mismo siempre, sin importar la distribución. Entonces, si entendemos el procedimiento con la Normal, deberíamos ser capaces de aplicarlo después a otras distribuciones.



La distribución Normal tiene muy buenas propiedades, una de ellas es que si tenemos suficientes datos, en algunas ocasiones podemos aproximar algunas distribuciones a la Normal. Esto se debe al Teorema Central del Límite, que dice que las sumas de suficientes variables aleatorias independientes, sin importar la distribución que tengan, tienden a la Normal. Entonces, por ejemplo, si estamos midiendo algo que es el resultado de la suma de muchas otras cosas aleatorias, el resultado estará normalmente distribuido.

Otro punto que vale la pena recordar es la aplicación. Anteriormente, vimos que la distribución de Bernoulli podría aplicarse a resultados binarios. Como cuando un usuario hace clic o no hace clic, o compra o no compra. Podemos aplicar esto a cualquier acción deseable que el usuario pueda o no pueda hacer. Pero hay situaciones en las que también deseamos medir las recompensas en un valor real. Un ejemplo simple de esto son las calificaciones. Por ejemplo, podríamos pedirles a los usuarios calificaciones de cero a diez. De manera bastante intuitiva, una calificación más cercana a diez es mejor y una calificación más cercana a cero es peor. Obviamente, estas calificaciones pueden usarse para recomendar artículos, productos o páginas más deseables para los usuarios. Otro ejemplo es el tiempo que el usuario pasan en la página web, podríamos pensar en esto como un indicador de participación. Obviamente, cuanto más tiempo pase el usuario en nuestra página web, mejor. Por supuesto, estos tiempos pueden no tener una distribución gaussiana, pero usualmente es una buena aproximación.

Vamos a ver un ejemplo concreto. Supongamos que hemos recopilado un montón de datos de alturas de estudiantes en una clase.



En este caso esos datos de altura serán una muestra de una variable aleatoria de tipo continuo, no es igual que la Bernoulli de antes que solo tomaba valores cero o uno. Como es continua, entonces su probabilidad estará caracterizada por una función continua que se llama función de densidad f. No como la discreta que la caracterizaba una función llamada función de probabilidad o masa. Entonces, si escribimos la función de verosimilitud será el producto de funciones continuas:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Si consideramos que estas alturas siguen una distribución Normal, tendríamos que sustituir la fórmula de la función de densidad de la Normal:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$\theta = \{\mu, \sigma^2\}$$

Veamos que esta expresión es simétrica cuadrática alrededor de μ. Es por eso que esta distribución es simétrica en el gráfico, centrada en μ. Por eso parece una curva de campana, que es igual a ambos lados.

Bien, entonces, ¿cuál es el siguiente paso? Como recordaremos, estamos haciendo el método de máxima verosimilitud, ¿qué significa eso? Nuevamente, eso significa que, dados los datos, queremos encontrar los parámetros que mejor describan a esos datos. En otras palabras, ¿qué valores de μ y σ tienen más sentido dados los datos que recopilamos? Estos valores maximizarán la probabilidad.

Para ello tendríamos que hacer lo mismo que antes, aplicar logaritmo y trabajar con la log-verosimilitud. Luego derivar e igualar a cero. Y eso nos daría los dos parámetros que maximizan la verosimilitud. Aquí tendremos que aplicar propiedades del logaritmo que son por ejemplo que el logaritmo del producto es la suma de los logaritmos.

$$Función\ de\ log-verosimilitud:$$

$$\begin{aligned} l(\theta) = \log L(\theta) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\right) \\ &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n} -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 = -\frac{1}{2}\left[log(2\pi\sigma^2)\sum_{i=1}^{n} 1 - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \\ &= -\frac{1}{2}\left[n\ log(2\pi\sigma^2) - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \end{aligned}$$

Luego derivamos con respecto a μ, la primera parte no depende de μ, y la segunda es cuadrática, al derivar queda:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n} \left(\frac{x_i-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Esto hay que igualarlo a cero y despejar a μ:

$$0 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right) \frac{1}{\sigma}$$

El resultado sería el siguiente, y podemos ponerle un sombrero a μ para mostrar que esto sería un estimador:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Esto de aquí es bastante familiar, es lo que se conoce como **media muestral**, sumar todos los datos de la muestra y dividir entre la cantidad total de datos. Esto tiene mucho sentido ya que μ es la media poblacional de la distribución Gaussiana. El siguiente paso será un poco más difícil. Porque habría que derivar ahora con respecto a σ^2, igualar a cero y despejar. Con ello hallaremos el estimador máximo verosímil para el parámetro σ^2 de la Normal, que coincide con su varianza. Entonces, comencemos expresando la probabilidad que vimos anteriormente de una manera un poco más simple, sabemos que el parámetro es sigma cuadrado, que nunca aparece por sí sola, siempre está al cuadrado. Entonces como esto puede parecer un poco confuso vamos a reemplazar a sigma al cuadrado por una nueva variable v, es decir $v = \sigma^2$ y vamos a meterla en la fórmula donde aparezca sigma al cuadrado:

$$l(v) = -\frac{1}{2}\left[n \log(2\pi v) - \frac{1}{v} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$

Entonces, en el primer término, sigma al cuadrado se convierte en v. En el segundo término, sigma al cuadrado aparece en el denominador, que podemos reemplazar con v si nos deshacemos del cuadrado. Tengamos en cuenta que esto lo podemos hacer porque v no depende del índice i del sumatorio. Por eso se puede incluso sacar fuera porque es como una constante.

El siguiente paso es diferenciar con respecto a v y después igualar el resultado a cero:

$$\frac{\partial l}{\partial v} = -\frac{1}{2}\left[n \frac{1}{2\pi v} 2\pi - \frac{1}{v^2} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$

$$0 = -\frac{1}{2}\left[n \frac{1}{v} - \frac{1}{v^2} \sum_{i=1}^{n} (x_i - \mu)^2\right]$$

Y el último paso sería despejar v:

$$\hat{v} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Obtenemos que v es igual a la expresión habitual para la varianza muestral. Es la suma de todos los elementos muestrales menos μ, todo eso al cuadrado, dividido por n. Bien, en este punto, hemos resuelto la estimación de máxima verosimilitud tanto para la media como para la varianza con la distribución Gaussiana. Pero aquí notamos algo interesante en nuestra expresión para la varianza muestral. Si vemos con atención, veremos que depende de μ, la media como parámetro poblacional. Sin embargo, en realidad **no conocemos este valor.** Lo que tenemos es un estimador para ese parámetro que nos dio la media muestral que hallamos al principio. Entonces lo que podemos hacer es reemplazar el parámetro verdadero μ que es desconocido, con nuestra estimación, el estimador máximo verosímil para μ, que es la media muestral:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

DISTRIBUCIÓN DE LOS ESTIMADORES

Anteriormente hallamos los estimadores máximo verosímiles de los dos parámetros de la distribución Normal, su media y su varianza. Ahora vamos a ver más detalles sobre ellos, o sobre en realidad cualquier estimador. Primero vamos a recordar un teorema que nos dice que las funciones de variables aleatorias también son variables aleatorias. ¿Por qué mencionamos esto ahora? Pues porque nuestros estimadores son funciones de variables aleatorias por lo tanto son variables aleatorias. Y, ¿cuál es la relevancia de esto? Bueno dado que $\hat{\mu}$ y $\hat{\sigma}^2$ son variables aleatorias, podemos hacer las cosas habituales que hacemos con variables aleatorias. Por ejemplo, podemos preguntar cuál es su distribución de probabilidad y cuál es su valor esperado. Y aquí suceden algunas cosas interesantes, por ejemplo, si partimos de una distribución Normal, es decir, si nuestros datos son Normales, la distribución de $\hat{\mu}$ es también una distribución Normal, porque la suma de Normales es Normal. Por otro lado, si no partimos de una distribución Normal, como la expresión de $\hat{\mu}$ es una suma de variables iid, entonces si tenemos suficientes elementos en esa suma, por el Teorema Central del Límite también sería Normal la distribución de $\hat{\mu}$.

Si quisiéramos hallar el valor esperado de $\hat{\mu}$ como variable aleatoria, en otras palabras su media, el resultado coincide con μ, la media poblacional, es decir, el valor esperado de la variable original. Esto es una noticia estupenda, porque dado que $\hat{\mu}$ es un estimador, cuando cambiemos de muestra, su valor puntual que se llama estimación también cambiará, pero en general es bueno saber que el valor que esperaríamos coincide con el parámetro que realmente está estimando, que es μ. Esta propiedad se llama insesgadez. Un estimador es insesgado si su valor esperado coincide con el parámetro que estima. En este caso el estimador máximo verosímil de la Normal para la media poblacional, $\hat{\mu}$, que es la media muestral, es un estimador insesgado porque su valor esperado coincide con el parámetro que estima: μ. La media de la media muestral es la media poblacional. Y su varianza es sigma al cuadrado sobre n.

Sin embargo, si quisiéramos hallar el valor esperado del estimador de sigma al cuadrado, $\hat{\sigma}^2$, no coincide con sigma al cuadrado poblacional. De hecho, el valor esperado de sigma al cuadrado es igual pero dividido por n-1 en vez de por n. Entonces, ¿cuál es la razón de que, por lo general, los estadísticos suelan dividir por n menos 1 en lugar de n? Porque esto nos devuelve un estimador insesgado: la cuasivarianza. En Machine Learning (Aprendizaje Automático), en realidad no tienden a preocuparse demasiado por esta distinción, ya que los conjunto de datos son tan grandes que dividir entre n o n-1 realmente no hace una gran diferencia.

Entonces en resumen, para $\hat{\mu}$ como variable aleatoria, su distribución, valor esperado y varianza son:

$$\hat{\mu} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

$$E(\hat{\mu}) = \mu$$

$$Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

Para $\hat{\sigma}^2$ como variable aleatoria, valor esperado no coincide con el parámetro que estima, por lo que no es insesgado:

$$E(\hat{\sigma}^2) \neq \sigma^2$$

Pero este estimador sí lo es, la Cuasivarianza muestral:

$$\hat{\sigma}_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

$$E(\hat{\sigma}_c^2) = \sigma^2$$

FUNCIÓN DE DISTRIBUCIÓN

Vamos a hablar ahora de una función muy importante que caracteriza a la probabilidad asociada a una variable aleatoria. Es la función de distribución. También se llama función de distribución acumulada. En ingles Cumulative Distribution Function (CDF). Entonces, para entender la CDF, probablemente sea más intuitivo comenzar con el caso discreto, suponiendo que X es una variable aleatoria discreta. La definición es la siguiente:

$$F(x) = P(X \leq x) = \sum_{k=-\infty}^{x} p(k)$$

Entonces, por ejemplo, F evaluado en tres sería igual a la probabilidad de que la variable aleatoria pueda tomar un valor menor o igual a tres. Bien, entonces, ¿cómo podemos encontrar este valor? Como X en este caso es discreta:

$$X: variable\ aleatoria\ discreta\ que\ toma\ valores\ \{0,1,2,3,\dots\}$$

Supongamos que tenemos su función de masa de probabilidad PMF, que es la probabilidad de que la variable aleatoria, X pueda tomar exactamente el valor x minúscula:

$$p(k) = P(X = k), \qquad k = \{0,1,2,3,\dots\}$$

Si queremos saber la probabilidad de que X sea menor o igual a tres, simplemente sumamos todas las probabilidades individuales, menores o iguales que 3:

$$F(3) = P(X \leq 3) = \sum_{k=-\infty}^{3} p(k) = p(0) + p(1) + p(2) + p(3)$$

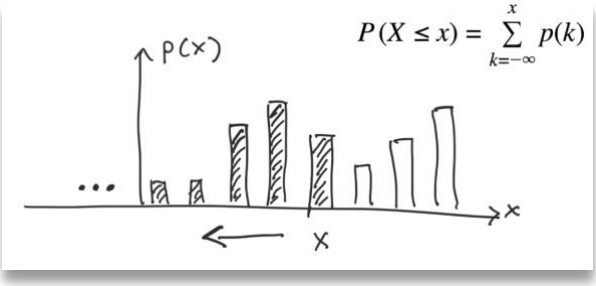
Entonces, la probabilidad de que pueda ser tres más la probabilidad de que pueda ser dos, más la probabilidad de que pueda ser uno, más la probabilidad de que pueda ser cero y seguimos hacia abajo hasta llegar al valor más pequeño que tome nuestra variable. En este caso el valor más pequeño es cero así que paramos ahí. Al sumar todas estas probabilidades individuales, obtendremos la probabilidad total de que X sea menor o igual a tres.

Ahora, pensemos en cómo se define la CDF para el caso continuo. Como saben, cuando se trabaja con variables aleatorias continuas, las sumas se convierten en integrales:

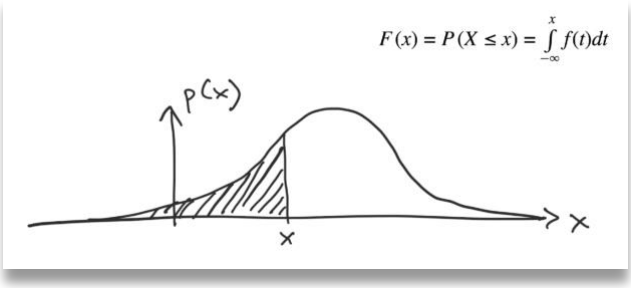
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Tengamos en cuenta que la t dentro de la integral se llama variable ficticia. Así que no importa qué letra usemos aquí, puede ser s, u, v, la letra que sea, mientras no sea x para no confundirnos con la otra x. No importa realmente porque al final esta variable desaparece después de tomar la integral. La CDF se halla con la integral de la PDF en el caso continuo. Y si tenemos la CDF, la PDF se puede hallar con la derivada de la CDF.

Si queremos entender esto de manera intuitiva por ejemplo en términos de imágenes, para el caso discreto, podemos imaginar la distribución como un montón de barras. Si queremos conocer la CDF en algún valor x, sumamos todas las barras hasta la barra en x inclusive.



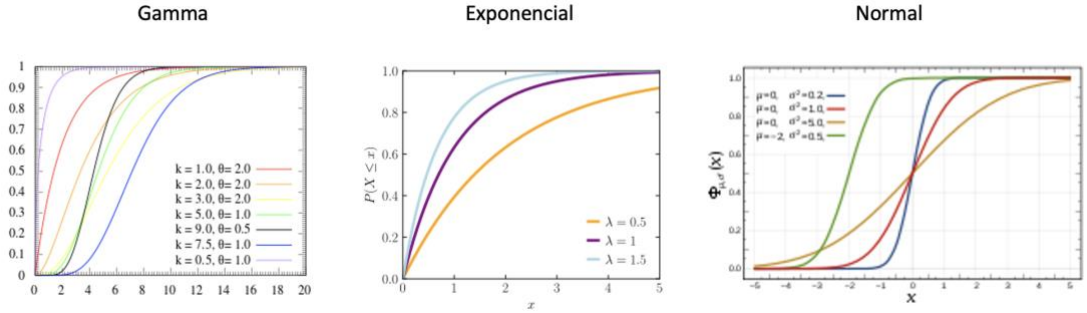
Para el caso continuo, la función de densidad es una función continua, y la CDF, o sea la integral de ella, es el área bajo esta curva.

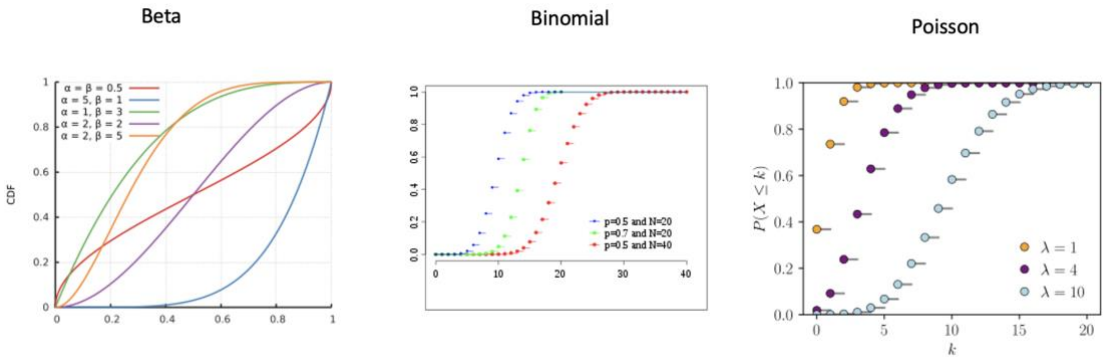


Por lo tanto, si queremos conocer la CDF en algún valor x, encontraremos el área bajo la curva desde menos infinito hasta x.

Una consecuencia de cómo definimos la CDF es que los valores en los extremos son siempre cero y uno. La CDF evaluada en el extremo inferior sería evaluarla en menos infinito y esto debe ser cero porque es la probabilidad de que la variable aleatoria pueda ser menor o igual a menos infinito. Pero esto no incluye ningún valor y, por lo tanto, esa probabilidad es cero. La CDF en más infinito debe ser uno porque es la probabilidad de que la variable aleatoria pueda ser menor o igual que infinito positivo. Por supuesto, esto incluye todos los valores posibles de la variable y, por lo tanto, esta probabilidad es uno. Además, dado que tanto la función de probabilidad como la función de densidad (caso discreto y continuo) son siempre mayores o iguales a cero, esto trae como consecuencia que la CDF siempre sea no decreciente.

Por ejemplo, si vemos las siguientes CDF de las distribuciones Gama, Exponencial, Normal, Beta, Binomial y Poisson, vemos que todas tienen esta forma sigmoidea.



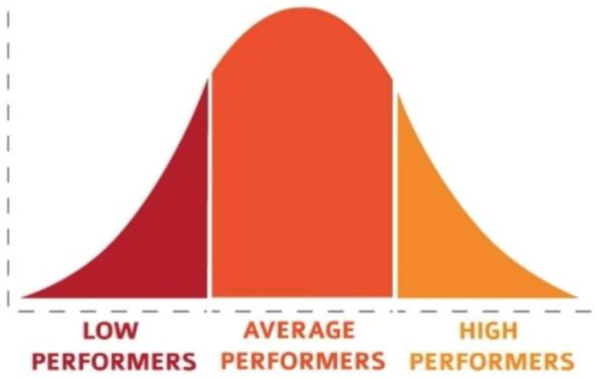


Esto es solo una consecuencia del hecho de que la CDF siempre comienza en cero, siempre termina en uno, y siempre es no decreciente.

PERCENTILES

Una función importante en estadística es la inversa de la CDF. Esto también se conoce como función de punto porcentual (en inglés Percent point function: percentile) por eso también se le relaciona con la noción de percentil. Entonces, ¿qué mide esta función inversa? Básicamente, vamos en la dirección opuesta en comparación con la CDF, en lugar de preguntar cuál es la probabilidad de que la variable aleatoria sea menor o igual a algún valor, hacemos la pregunta opuesta. Dada alguna probabilidad de que la variable aleatoria sea menor o igual a algún valor, ¿cuál es el valor donde sucede esto? Consideremos ejemplos más concretos.

Un ejemplo es considerar las calificaciones de estudiantes en un examen. Como sabemos, podemos crear curvas de campana a partir de las calificaciones de los alumnos.



Bien, digamos que creamos un modelo gaussiano para las calificaciones del examen. Ya que no queremos confundir las puntuaciones de los exámenes con probabilidades, vamos a decir por ejemplo que la calificación máxima que se puede obtener en el examen es de doscientos. Entonces, 200 de 200 es una puntuación perfecta. Ahora, vamos a hacer una pregunta, ¿cuál es la probabilidad de que el estudiante logre una puntuación de 170 o menos? Para ello tendríamos que usar la CDF. Vamos a suponer que esto nos dio un 95%:

$$F(170) = P(X \leq 170) = 0.95$$

Y si queremos la probabilidad de que cualquier estudiante de la clase obtenga una puntuación por encima de 170? Sería la parte contraria entonces esto es 1-la probabilidad anterior:

$$P(X > 170) = 1 - P(X \leq 170) = 1 - 0.95 = 0.05$$

Entonces, la probabilidad de que un estudiante supere los 170 puntos de 200 en el examen es del cinco por ciento. Eso es de esperar, ya que es una calificación muy alta.

Bien, pues ahora volvamos a la inversa de la CDF. Digamos que quiero preguntar, ¿cuál es la puntuación máxima que logró alcanzar el 95 por ciento (inferior) de la clase? Por supuesto, la respuesta es la CDF inversa de 0.95, como cabría esperar:

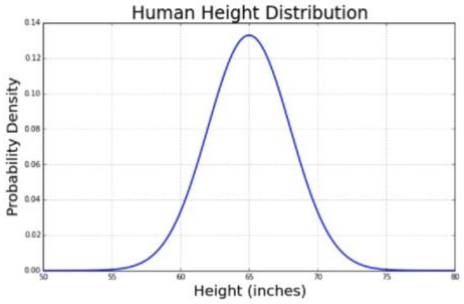
$$F^{-1}(0.95)$$

$$P(X \leq \ ?) = 0.95$$

La respuesta es: 170

Es decir, si estás en el 95 por ciento inferior de los estudiantes, obtienes una puntuación de 170 o menos. Ahora, cuando decimos que algún alumno está en el percentil 95, queremos decir que lo está haciendo mejor que el 95 por ciento inferior.

Bien, veamos otro ejemplo, digamos que hemos medido todas las alturas de nuestros amigos, y nuevamente, vamos a modelar estas alturas como una distribución Gaussiana, supongamos que la altura promedio es de 170 centímetros y la desviación estándar es de 7 centímetros.



Bien, ahora digamos que un amigo nos dice que su estatura está en el percentil 95. ¿Podríamos determinar qué tan alto es? La respuesta es sí, simplemente calculando la CDF inversa del 95%. La respuesta es 181.5 centímetros.

$$F^{-1}(0.95) = 181.5\text{cm}$$

Ahora, digamos que medimos la altura de otro amigo y mide 160 centímetros, ¿en qué percentil está?

$$F(160) = 0.08$$

En este caso, la respuesta es aproximadamente el 8%. Por lo tanto, ese amigo se encuentra en el octavo percentil. En otras palabras, el ocho por ciento de las personas son más bajas que él. Por el contrario, el 92 por ciento de las personas son más altas que él.

RESUMEN

Vamos a resumir todo lo que hemos cubierto en esta sección. Hasta ahora hemos estado revisando conceptos básicos de probabilidad. La razón es que esta es la base de todo lo que vamos a ver después, tanto de la parte de Inferencia Bayesiana como la parte de modelos de Machine Learning. Bien, entonces, ¿qué hemos cubierto hasta ahora? Hemos revisado las reglas básicas de probabilidad, incluida la regla de Bayes. Aprendimos sobre distribuciones de probabilidad. Aprendimos sobre distribuciones conjuntas, distribuciones condicionales y distribuciones marginales. Aprendimos sobre el importante concepto de

estimación de máxima verosimilitud. Una idea muy importante que tenemos que mencionar antes de seguir es que cuando estimamos parámetros como por ejemplo la media, estas eran estimaciones puntuales o precisas, un número. Eso significa que, dados los datos, siempre terminamos con un solo valor. Hay algunas preguntas importantes que aún no hemos considerado sobre este valor único. Por ejemplo, ¿qué confianza tenemos en esta estimación? ¿Es una buena estimación o es una mala estimación? Esto nos lleva al concepto de intervalos de confianza, que estudiaremos más adelante. Y que forma parte de la Inferencia Estadística. Otra pregunta que podemos hacernos es esta. Supongamos que tenemos dos estimaciones. Por ejemplo, estamos tratando de comparar la tasa de clics entre el anuncio A y el anuncio B, y queremos saber si A es mejor que B o viceversa. Obviamente, gana el que tenga la tasa de clics más alta. Pero, ¿cómo podemos estar seguros de que A es mejor que B? Esto se remonta a la idea de confianza. Si ni siquiera estamos seguros de nuestras estimaciones, tampoco estamos seguros de que A sea mejor que B. Esa es otra pregunta que discutiremos a continuación. Por supuesto, el objetivo es elevar nuestra comprensión al nivel de la perspectiva bayesiana. Reconocemos que las estimaciones puntuales son limitadas. Los bayesianos dirían entonces ¿qué pasa si en lugar de una estimación puntual, puedo encontrar una distribución? Es decir, en vez de encontrar el estimador máximo verosímil de la media, ¿qué pasa si puedo encontrar la distribución de la media como variable aleatoria?

Esto es mucho más poderoso ya que, como podemos ver, nos da directamente algún tipo de intervalo de confianza, aunque en el enfoque bayesiano no se llaman intervalos de confianza. Y como tenemos distribuciones, podemos hacer preguntas como ¿cuál es la probabilidad de que A sea mejor que B? Entonces, el enfoque bayesiano es tratar todo como si fuera aleatorio. El enfoque clásico, también llamado enfoque de frecuencias, o frecuentista, es tratar los parámetros como si fueran fijos pero desconocidos. Y las estimaciones puntuales son nuestra mejor suposición en este enfoque. Mientras que en el enfoque bayesiano, decimos que esos parámetros no son fijos sino que también son variables aleatorias. Y debido a que son aleatorias, tienen distribuciones que pueden encontrarse utilizando los datos que recopilamos.

INTRODUCCIÓN A LA ESTADÍSTICA BAYESIANA

En esta sección vamos a explorar la estadística bayesiana desde cero. Para que pueda ser útil y resulte sencillo de entender para todo el mundo, creo que es necesario comenzar por la parte más básica, por los fundamentos sobre los que se rige esta teoría o este enfoque. Así que comenzaremos dando respuesta a preguntas básicas pero muy profundas como por ejemplo qué significa la probabilidad. En eso nos vamos a enfocar en esta sección introductoria, y vamos a razonarlo en el contexto de eventos o ejemplos sencillos como tiradas de dados, pero también eventos muy concretos y útiles para entender el por qué del enfoque bayesiano, como las carreras de caballos. Vamos a repasar el concepto de la probabilidad condicional y vamos a ver cómo puede ayudarnos a resolver algunos problemas poco intuitivos. También hablaremos sobre el Teorema de Bayes y prometo que vamos a tratar de mantener las cosas lo más visuales e intuitivas posible en todo momento. Finalmente, echaremos un vistazo al rompecabezas que inició todo esto hace 250 años cuando Thomas Bayes propuso un enigma que involucraba una mesa de billar. De esta manera vamos a poder tener la base necesaria para entender después los métodos más complejos de la inferencia bayesiana y del enfoque bayesiano del análisis de datos

PROBABILIDAD FRECUENTISTA VS BAYESIANA

Vamos a comenzar con esta pregunta: ¿Qué es la probabilidad?

Primeramente vamos a recordar que la probabilidad es en realidad un concepto relativamente nuevo (entre comillas) en matemáticas. Porque solo se empezó a pensar realmente de la forma en que lo pensamos ahora, en el siglo XVII. Entonces, vamos a ver cómo podemos dar sentido a lo que significa ese

concepto, la probabilidad, a través de un ejemplo. Seguramente sepas que la probabilidad de tirar un dado y sacar un 3 es 1/6, y todos los demás números del dado tienen exactamente la misma probabilidad de salir. Esto es uno de los ejemplos más clásicos que se suelen poner a la hora de aprender el concepto de la probabilidad, así que seguramente te suena mucho y estarás muy probablemente convencido de que efectivamente 1/6 es la probabilidad de sacar cualquiera de los números del dado. Así que lo que vamos a hacer es, vamos a pensar en dos formas diferentes en las que podríamos caracterizar este ejemplo.

La primera forma en que podríamos pensar en esto es desde un punto de vista objetivo. Entonces, ¿por qué, por ejemplo, decimos que la probabilidad de sacar un 3 en un dado es un sexto? Esto es resultado de un punto de vista objetivo, porque lo que se hace para calcular esa probabilidad es tener en cuenta un enfoque que se llama frecuentista. Vamos a ver esto qué significa. En el enfoque frecuentista, lo que vamos a hacer es tirar el dado no solo una vez, sino dos, tres, cuatro, millones de veces, o incluso infinitas veces. Entonces, cuando hacemos esto lo siguiente es tomar una muestra aleatoria finita, o incluso tal vez infinita, de esas tiradas, donde miraremos los resultados. Y qué nos interesa, pues queremos saber qué proporción de esos dados tirados salió con un número 3. Y sabemos que cuando el número de tiradas tiende a infinito, esa proporción tenderá a un sexto. Entonces, una forma de explicar cuál es esa probabilidad es en términos de frecuencias. Diríamos que si el evento sucediera infinitas veces, ¿cuál sería la proporción de lo que nos interesa?

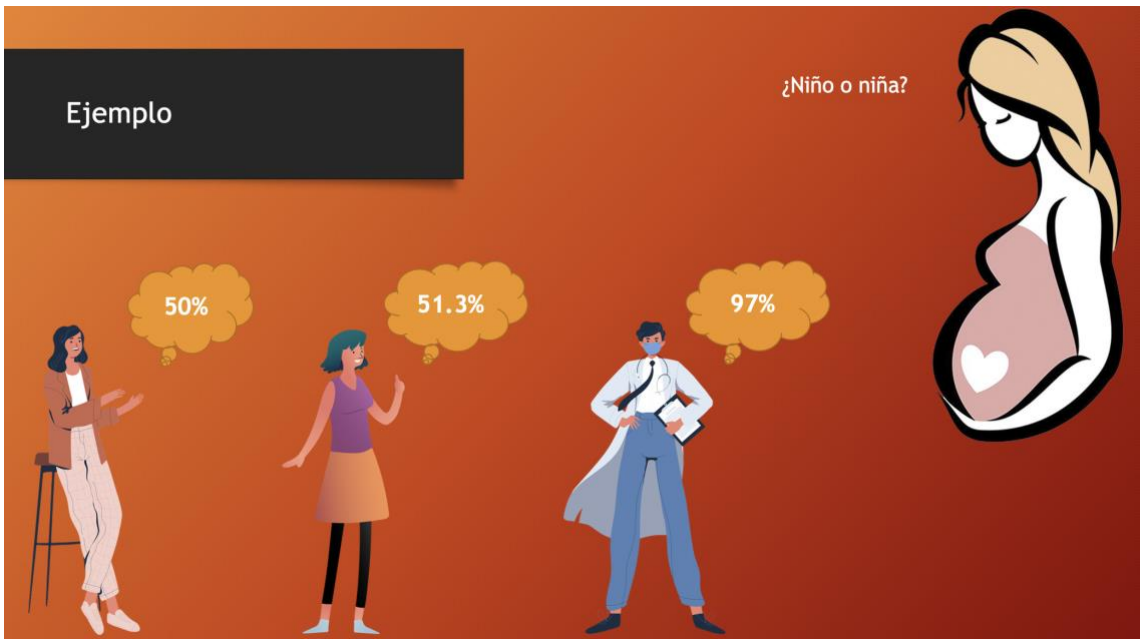
Ahora bien, ¿cuál podría ser un punto de vista diferente? Bueno, una forma diferente de pensarlo es un enfoque subjetivo, es decir, cada persona tiene sus propias razones para creer en una cierta probabilidad. Entonces, esta postura subjetiva es la que formará la base de los modelos bayesianos. Ahora bien, una cuestión interesante es decidir, ¿cuál es el resultado correcto? ¿Lo que nos dice el enfoque frecuentista o lo que nos dice el enfoque bayesiano? Lo que pasa a menudo es que la intuición de mucha gente al principio es decir que el enfoque subjetivo parece demasiado débil. Pero lo que realmente pasa es que el enfoque no es subjetivo en el sentido de que puedas pensar lo que tú quieras, sino subjetivo en el sentido de que cada persona puede tener una respuesta diferente a la pregunta. Ahora, vamos a ver dos razones para pensar de esta manera.

La primera razón se basa en que el enfoque frecuentista no funciona en todos los casos, y la segunda razón tiene que ver con una situación diferente, donde vamos a ver una gran razón para pensar que el modelo subjetivo bayesiano es realmente bueno.

Entonces primero, el enfoque frecuentista es bastante bueno y encaja a la perfección cuando pensamos en dados, o en algo que nos podemos imaginar repitiendo muchas veces como, por ejemplo, el tirar el dado, el tirar una moneda, o algo similar. Pero ¿qué pasa cuando tenemos un evento específico que no se puede repetir, que solo puede suceder una vez? Tomemos, por ejemplo, una carrera de caballos. Entonces nos podríamos preguntar, ¿cuál es la probabilidad de que gane nuestro caballo? Es el tipo de cosas que todo el mundo quiere saber, todo el que va a apostar en las carreras de caballos. Bien, pues hay algunas razones bastante buenas para pensar que el enfoque frecuentista realmente no funciona en esta situación. Porque en el enfoque frecuentista ¿qué haríamos?, tendríamos que plantear la hipótesis de un conjunto infinito de carreras de caballos. Entonces tendríamos que imaginar esta única carrera de caballos sucediendo infinitas veces. Y a partir de eso, tendríamos que extraer una muestra aleatoria de estas carreras de caballos, al igual que hicimos con el ejemplo de los dados. Y tal vez incluso esta muestra sea infinita. Y luego lo que vamos a hacer con esto es contar cuántas veces ganó nuestro caballo. Y vamos a dividir eso por la cantidad de carreras que hubo. Y sea cual sea el número, eso nos dirá la probabilidad de que nuestro caballo gane.

Mucha gente piensa que hay muchos errores con este modelo por dos razones. La primera razón es, ¿qué significa imaginar un conjunto infinito de un solo evento como es una carrera de caballos? No es algo que pueda suceder más de una vez, el mismo evento. Es pura imaginación. Y además, ¿qué significa eso de sacar una muestra aleatoria de un conjunto hipotético? Entonces, parece que hay algunas razones bastante buenas para pensar que el enfoque frecuentista no es bueno en situaciones que son eventos únicos.

Ahora veamos algunas razones por las que el modelo bayesiano, el enfoque subjetivo podría ser realmente bueno en este tipo de situaciones. Vamos a hablar de un ejemplo diferente que además es muy fácil de entender, sobre todo la parte de la subjetividad. Supongamos que mi mejor amiga esperando un bebé y no sabemos si va a ser un niña o niño. Vamos a imaginar que aquí está mi madre María. Ella piensa que la probabilidad de que el bebé sea hembra es del 50%. No tiene ninguna razón para decantarse por uno o por otro así que piensa que esa probabilidad es la mitad para cada uno. ¿Podríamos decir que esta probabilidad es correcta? Creo que mucha gente lo haría. Pero vamos a suponer que ahora entra en escena nuestra amiga Giselle. Giselle es una investigadora, y Giselle acaba de escribir un artículo para la Organización Mundial de la Salud que demuestra que en mi área, el 51.3% de todos los bebés nacidos son hembras. Entonces, ¿la opinión de Giselle es más correcta que la de mi madre? Giselle ciertamente tiene más información. ¿Pero es justo decir que Giselle tiene razón y mi madre está equivocada? Sería injusto decir eso porque quizás si mi madre se lee el artículo de Giselle, y tiene más información, podría cambiar de opinión. Pero tal y como están las cosas, sería irracional que mi madre dijera 51.3% porque ella no tiene ninguna razón para pensar que la probabilidad sea superior al 50%. Ahora, vamos a incluir a otro personaje, aquí está Alberto. Alberto es ginecólogo, y aunque mi amiga no tiene mucho tiempo todavía, le han hecho una exploración y aunque no es un 100% exacto, Alberto cree que la probabilidad de que sea niña es del 97%. Entonces, ¿qué decimos ahora? ¿Es correcto decir que la probabilidad de Alberto es la correcta y que mi madre y Giselle están equivocadas? La intuición de la mayoría de las personas va a ser que Alberto ciertamente tiene más información. Así que por ejemplo que mi madre diga de repente que hay un 95% de probabilidad que sea niña, sin tener ninguna información, eso sería irracional.

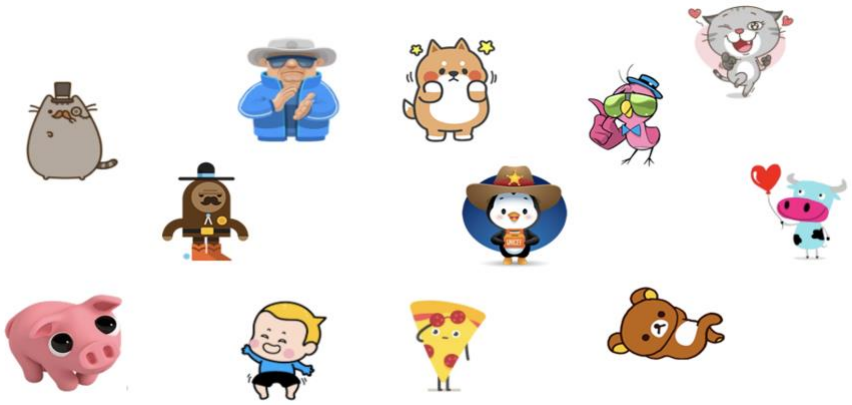


Entonces, el enfoque bayesiano consiste en decir que en realidad las tres opiniones podrían ser correctas. Porque realmente las probabilidades representan el grado de creencia de cada individuo, es decir, una medida de su incertidumbre. Ese es el sentido en el que la probabilidad es subjetiva. No es que puedas pensar lo que quieras, es decir, Alberto no puede mirar el escáner y decir que todavía cree que es del 50%. Eso sería irracional. Mi madre no puede en ausencia de cualquier otra evidencia, decir que cree que hay un 97% de probabilidades de que sea niña. Eso también sería irracional. Lo racional es que cada persona tiene una cierta cantidad de evidencia que forma ese grado de creencia o esa medida de su incertidumbre. Los grados de creencia formarán la base del enfoque bayesiano.

PROBABILIDAD CONDICIONAL

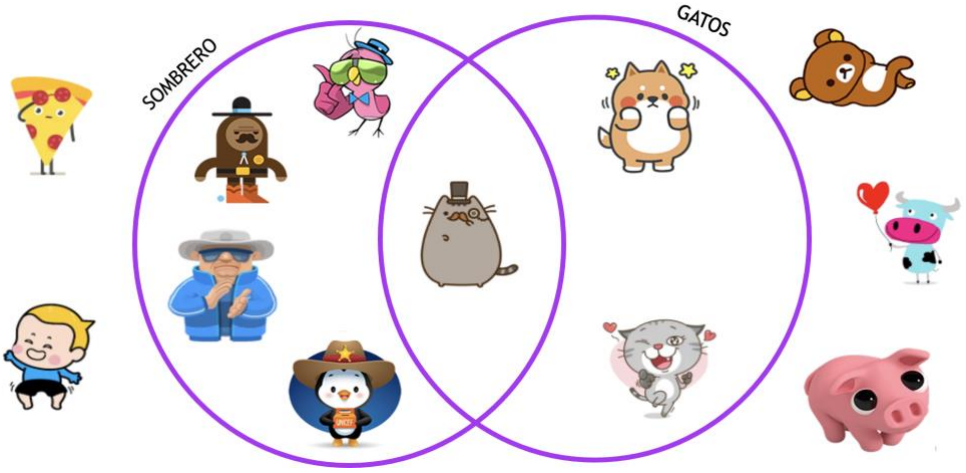
Uno de los conceptos más importantes para entender la estadística bayesiana es la probabilidad condicional. Para entenderlo vamos a ver un ejemplo muy sencillo que nos va a ayudar a construir esa

intuición que necesitaremos para entender el enfoque bayesiano. Vamos a considerar que tenemos estos 12 personajes de aquí:



Podríamos plantear diferentes preguntas. Una de las cosas que podríamos hacer con ellos es clasificarlos, y aunque tenemos muchas formas diferentes para hacerlo, vamos a clasificarlos por ejemplo según si son gatos y llevan sombrero.

Si los organizamos vemos que en el medio queda solo un personaje, que cumple las dos condiciones a la vez, de ser gato y llevar sombrero:



Tenemos algunos que son gatos pero no llevan sombrero, y otros que llevan sombrero pero no son gatos. Y otros que no cumplen ninguna de las dos condiciones. Ahora que tenemos esto visualizado, podemos hacer algunas preguntas de probabilidad al respecto. Por ejemplo, **¿cuál es la probabilidad de que un personaje seleccionado al azar sea un gato?** Para lo cual vamos a contar cuántos gatos hay. Hay tres gatos de 12 personajes, por lo que esa probabilidad sería:

$$P(G) = \frac{3}{12}$$

Otra pregunta, un poco más compleja, sería **¿cuál es la probabilidad de que un personaje seleccionado al azar sea un gato dado que usa sombrero?** Y para esto, debemos pensar en la condicionalidad. Entonces, la condición aquí es que el personaje use sombrero, lo que significa que nos vamos a limitar a pensar solo en los cinco personajes que usan sombrero. Y la pregunta es ¿cuántos de ellos son gatos? Solo 1 cumple esa condición. Así que la respuesta es:

$$P(G|S) = \frac{1}{5}$$

Entonces, uno de cada cinco es la probabilidad de que un personaje sea un gato dado que usa sombrero.

Bien, hasta ahora esto sería un ejemplo muy sencillo de ver la probabilidad condicional. Pero ¿cómo podemos formalizar lo que acabamos de hacer y establecer una regla general? Podemos decir que en el numerador de la probabilidad condicional anterior, el 1 lo que representa es la cantidad de personajes que son de tipo gato y a la vez usan sombrero, así que eso puede representar la probabilidad de la intersección de ambas condiciones. Y el conjunto mayor del que se extrae sería la probabilidad de que un personaje use sombrero, ya que esta fue la condición que nos pusieron desde un principio:

$$P(G|S) = \frac{P(G \cap S)}{P(S)}$$

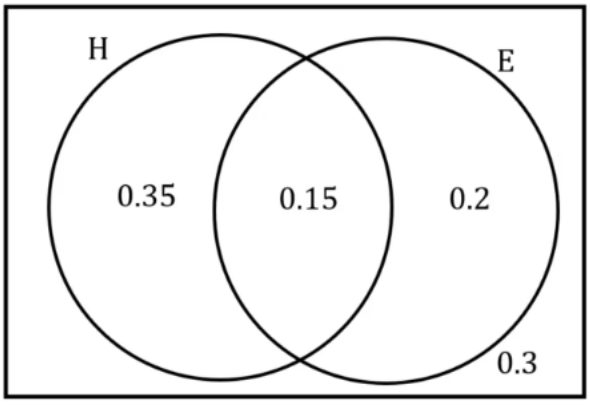
Esto de aquí no es más que la definición más simple del Teorema de Bayes, que es la base de la Estadística Bayesiana.

DIAGRAMAS DE VENN

Ahora que hemos derivado el Teorema de Bayes en su definición simple, vamos a aplicarlo a otro ejemplo que utiliza probabilidades donde no solo tengamos que contar simplemente, como en el ejemplo anterior. Nos piden la probabilidad del teorema que es la probabilidad de H dado E. Y que sabemos que es igual a la probabilidad de la intersección de H con E dividido entre la probabilidad total de E.

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

Nos dan el siguiente diagrama. Vamos a analizarlo y resolverlo paso a paso.



Lo más fácil es comenzar con la probabilidad de E. ¿Cuál es la probabilidad de que E ocurra? Bueno, no es solo 0.2 sino toda la burbuja E, entonces sería:

$$P(E) = 0.2 + 0.15 = 0.35$$

Y la probabilidad de H intersección E sería solo la parte que queda entre los dos, que es:

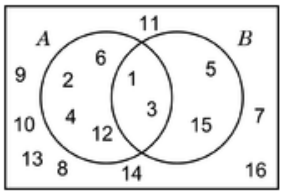
$$P(H \cap E) = 0.15$$

Entonces:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{0.15}{0.35} = 0.43$$

En gráfico de arriba se llama Diagrama de Venn. Los Diagramas de Venn representan relaciones topológicas de unión, inclusión y disyunción entre dos conjuntos. Por ejemplo:

$A = \{1; 2; 3; 4; 6; 12\}$
$B = \{1; 3; 5; 15\}$
$U = \{1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16\}$



https://es.wikipedia.org/wiki/Diagrama_de_Venn

Vamos a ver ahora otro ejemplo donde la información que se nos presenta la podemos representar nosotros en un diagrama de Venn y luego hallar probabilidades. Es decir, vamos a dibujar un diagrama de Venn por nosotros mismos.

El ejemplo nos dice, José y María a veces llegan tarde a clase. El setenta por ciento (70%) de las veces, ninguno de ellos llega tarde. José llega tarde el 20% de las veces y María el 25% de las veces. Por ejemplo, el lunes pasado, María llegó tarde. **¿Cuál es la probabilidad de que José haya llegado tarde?**

Entonces, lo primero que debemos hacer es asignar una terminología para formalizar qué es lo que realmente nos han dicho y lo que nos han pedido. Primero ¿cuál es nuestra evidencia? Nos dicen que María llegó tarde así que vamos a llamar al evento de que María llega tarde "E". Y ¿cuál es la hipótesis?, lo que nos interesa probar o encontrar es la probabilidad de que José haya llegado tarde. Entonces "H" va a ser José llega tarde:

Notación:

- H: José llega tarde
- E: María llega tarde

Entonces, vamos a intentar ahora escribir con esta notación la parte de la información que se nos ha dado.

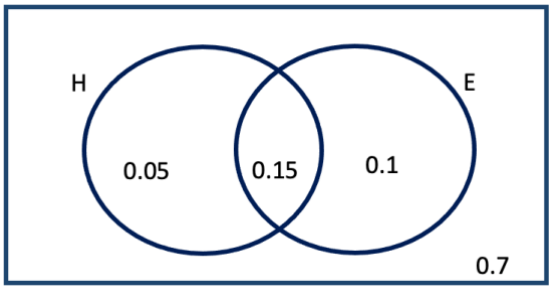
Información que nos dan:

a) Que ninguno llega tarde el 70% de las veces: $P(H^C \cap E^C) = 0.7$

b) Que José llega tarde el 20% de las veces: $P(H) = 0.2$

c) Que María llega tarde el 25% de las veces: $P(E) = 0.25$

Como hemos nombrado a H y E como eventos donde ellos sí llegan tarde, cuadno nos hablan de lo contrario lo denotamos como el complemento de esos eventos o esos conjuntos. Hay varias formas de denotar al complemento (con ', con C, etc.) Aquí estamos usando la C. Ahora, ¿cómo podemos representar todo esto en un Diagrama de Venn? Primero dibujamos la caja, los dos conjuntos o burbujas H y E.



Fuera de los conjuntos podemos poner 0.7 porque es la probabilidad de la intersección de ambos complementos. A continuación, sabemos que la totalidad de la burbuja H es 0.2 porque la probabilidad de que suceda H es 0.2. Y sabemos que lo exterior es 0.7, y que 0.2+0.7 sería 0.9, y como en total debe sumar 1, lo que falta que es la parte que forma parte de E y que no es la intersección, que resulta ser 0.1. Además como sabemos que la P(E)=0.25 y hay un trozo que es 0.1, el otro trozo tiene que ser 0.15, que justo es la intersección entre H y E. Y como ese trozo también forma parte de H, y P(H)=0.2, el otro trozo de H que no es la intersección, sería 0.05. Así que ese sería el Diagrama de Venn que representa este problema y ahora ya estaríamos en condiciones de ir a por la pregunta principal que nos formularon al principio, que es: **¿cuál es la probabilidad de que José llegue tarde, sabiendo que María ha llegado tarde?**

Si usamos el Teorema de Bayes:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{0.15}{0.25} = 0.6 = 60\%$$

Con lo cual podemos decir que el hecho de que María llegue tarde es realmente una evidencia nueva que influye de alguna manera en que José llegue tarde, porque normalmente él tenía una probabilidad de llegar tarde un poco baja, del 20%, pero si ahora nosotros tenemos una nueva evidencia, de que María llega tarde, esta probabilidad se actualiza, y aumenta a un 60%. Y es este tipo de pensamiento el que nos va a ayudar a entender el enfoque bayesiano que veremos más adelante, esta idea de actualizar nuestra probabilidad, dada alguna nueva evidencia.

DIAGRAMA DE ÁRBOL

Vamos a ver otro ejemplo de cómo aplicar la fórmula de la probabilidad condicional, usando una nueva herramienta de ayuda que se llama Diagramas de Árbol. En el ejercicio nos dicen que a Pedro le encanta jugar al Tenis. Pero especialmente cuando hace buen tiempo. Cuando hace sol, la probabilidad de que juegue al tenis es del 80%. Cuando no hace sol, la probabilidad es solo del 35%. Y también nos dicen que hay un 60% de probabilidad de que haga sol en un día determinado. De antemano sabemos que Pedro el sábado pasado jugó al tenis. **Entonces, sabiendo esto, ¿cuál es la probabilidad de que haya habido sol el sábado pasado?** Es decir que haya sido un día soleado. Bien, entonces obviamente podemos pensar que hay un 60% de probabilidad de que haya habido sol porque nos dan esa probabilidad para un día cualquiera, nos dicen que un día cualquiera será soleado con un 60% de probabilidad, normalmente. Pero nosotros tenemos ahora una nueva evidencia, que es que Pedro jugó al tenis ese día y eso será relevante, y podría llevarnos a actualizar nuestra opinión sobre cuál es realmente esa la probabilidad. Vamos a intentar modelar esto.

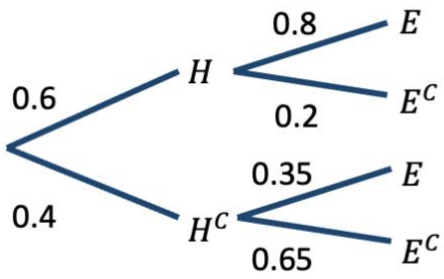
Decidamos rápidamente cuál es nuestra hipótesis y cuál es nuestra evidencia. La hipótesis, lo que estamos investigando, es si hizo sol el sábado pasado o no. Nuestra evidencia es que Pedro jugó al tenis el sábado pasado.

H: Fue un día soleado el sábado pasado

E: Pedro jugó al tenis el sábado pasado

Entonces, ¿cómo podemos representar esto en un diagrama de árbol?

Bueno, primero vamos a trazar dos ramas que representan las dos posibilidades de la hipótesis, que es que el sábado pasado fue un día soleado o no lo fue (su complemento).



Ahora, la probabilidad de que un día sea soleado (H) sabemos que es del 60 por ciento. Antes de que nosotros conozcamos alguna evidencia. Por lo tanto, hay un 40 por ciento de probabilidad o 0.4 de que no haya habido sol el sábado pasado (complemento de H).

Ahora, si fue un día soleado el sábado pasado (si estamos en la rama H), sabemos que Pedro podría haber jugado tenis. Recordando que esto sucede con probabilidad, 80 por ciento. Porque sabemos que cuando hace sol, hay un 80 por ciento de posibilidades de que juegue al tenis. Y por otro lado esto deja un 20 por ciento de posibilidades de que no juegue al tenis en un día soleado, porque es lo que falta para llegar al total, al 100% (es el complemento).

Luego, también podríamos ver qué pasa cuando no hace sol (si estamos en la rama complemento de H). Entonces, en la rama de abajo, cuando no es un día soleado, Pedro también podría jugar al tenis, pero la probabilidad de que eso suceda es del 35 por ciento. Y eso significa que la probabilidad de que no haya jugado al tenis en un día malo, es del 65 por ciento.

Bien, entonces ya tenemos nuestro diagrama de árbol completo con todas las probabilidades y ahora podemos aplicar la fórmula.

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{0.48}{0.62} \approx 77\%$$

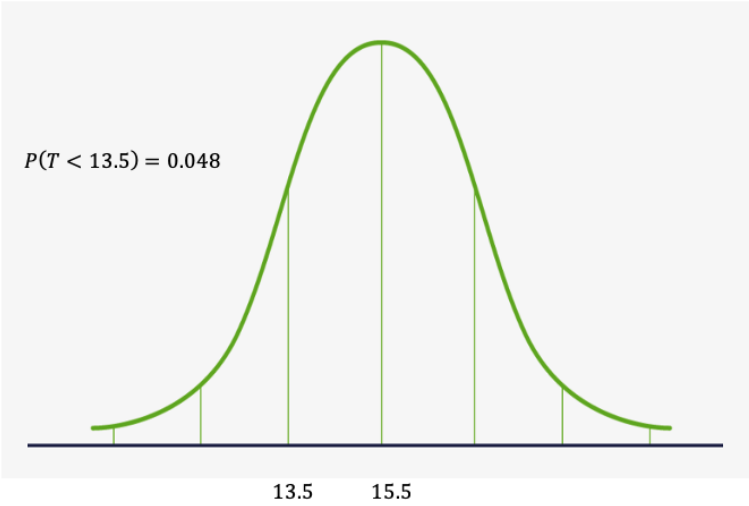
Entonces la P(H|E) es igual a la probabilidad de la intersección de H con E, la probabilidad de que haya sido un día soleado y de que Pedro juegue al tenis. Para ello, multiplicamos las ramas de la probabilidad de que hubiese sol y la probabilidad de que jugara al tenis y eso da 0.48. Ahora, ¿qué pasa con el denominador? ¿Cuál es la probabilidad de nuestra evidencia en general? La evidencia de que haya jugado al tenis. Esto es más complicado porque ese hecho depende de si el día fue soleado o no. Entonces, si el día fue soleado la probabilidad de que jugara al tenis es la que acabamos de hallar (0.48). Y si no lo fue, es decir, si fue un día nublado, la probabilidad es la de la otra rama 0.35 x 0.4 y la tenemos que agregar. Entonces podemos sumarlas y obtenemos 0.62 como nuestra probabilidad de que Pedro haya jugado al tenis, es decir, P(E). Ahora tenemos el numerador y el denominador. Dividimos y obtenemos que hay un 77 por ciento de probabilidades de que hiciera sol dada la evidencia de que Pedro ha jugado al tenis. Una vez más, podemos ver que este teorema nos ha ayudado a actualizar nuestra idea de cuál es la probabilidad. Nosotros creíamos que, en general, había un 60 por ciento de posibilidades de que hiciera sol. Pero si llega una nueva evidencia, que es que Pedro jugó al tenis en ese día, esto afecta directamente a nuestra idea sobre cuál es la probabilidad. Ahora hay un 77 por ciento de probabilidades de que haya habido sol ese día, porque sabemos que Pedro jugó al tenis ese día. Así que esto es lo que realmente hace el Teorema de Bayes, nos ayuda a actualizar nuestra idea sobre la probabilidad, dada alguna nueva evidencia que nos permite actualizar y modificar nuestro propio criterio.

PROBABILIDAD CONDICIONAL CON LA NORMAL

Algunas veces, las probabilidades que necesitamos para hallar, por ejemplo, una condicional, no nos las dan de antemano, como en los ejemplos anteriores, sino que lo que nos dan es una distribución de probabilidad, que es una función que caracteriza la probabilidad de una variable aleatoria, y con ella es que vamos a encontrar las probabilidades que antes ya nos daban calculadas o en porcentajes. Vamos a aprender cómo se aplica a las distribuciones Normales lo que hemos aprendido sobre la probabilidad condicional. Vamos a considerar que tenemos a una niña a la que le encanta correr que se llama Sofía. Y digamos que para las niñas de su edad, el tiempo que se tarda en correr una carrera de 100 metros, ese tiempo, sigue una Distribución Normal. Recordemos que a la distribución Normal la caracterizan dos parámetros que son su media y su desviación típica. Entonces sobre esta Normal sabemos que la media es de 15.5 y la desviación típica es de 1.2. ¿Qué quiere decir esto? Que el tiempo promedio que se tarda en correr una carrera rápida de 100 metros para las niñas de la edad de Sofía, es de 15.5 segundos. Y el 1.2 significa que la mayoría de las niñas completarán la carrera dentro del rango de 1.2 segundos alrededor de ese valor promedio. Ahora, la primera pregunta que queremos responder aquí es **¿cuál es la probabilidad de que esta chica corra 100 metros en menos de 13.5 segundos?**

Bueno, para responder a eso, vamos a tener que calcular la probabilidad usando la distribución Normal, que recordemos que es un área debajo de la curva.

$$T \sim N(\mu = 15.5, \sigma = 1.2)$$



Esta probabilidad la podemos calcular usando la tabla de la Normal. Vamos a pasar directamente a su resultado, que es:

$$P(T < 13.5) = 0.048$$

Pero ahora supongamos que tuviéramos más información. Supongamos que también sabemos que para entrar en el equipo de corredores de la escuela, es necesario poder correr 100 metros en menos de 14 segundos. Porque solo aceptan a los niños que ya tienen una cierta preparación. Y digamos que también sabemos que esta chica llamada Sofía está en el equipo de corredores de la escuela. Así que ahora tenemos una nueva evidencia para respaldar.

Sabemos que, en general, la probabilidad de que pueda correr 100 metros en menos de 13.5 segundos es de aproximadamente 4.8% o 0.048. Nuestra hipótesis es que Sofía corre 100 metros en menos de 13.5

segundos. Y nuestra evidencia es que ella está en el equipo de corredores de la escuela. Así que veamos cómo esta nueva evidencia cambia lo que sabemos de nuestra hipótesis.

Entonces, estamos buscando la probabilidad de H dado E. La probabilidad de que nuestra hipótesis sea cierta, dada la evidencia, y tenemos nuestra fórmula para esto:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P\big((T < 13.5) \cap (T < 14)\big)}{P(T < 14)}$$

En el numerador nos queda la intersección de H con E. H es nuestra hipótesis, el hecho de que Sofía completa los 100m en menos de 13.5 segundos. Y E es la evidencia, de que Sofía forma parte del equipo de la escuela que esto es equivalente a que el tiempo en que se completan esos 100m de carrera es inferior a 14 segundos, ya que este era el requisito para formar parte de este equipo. Y dividimos ese numerador por la probabilidad de que esté en el equipo de la escuela, que significa que hace la carrera en menos de 14 segundos.

Vamos a trabajar el numerador. ¿Qué significa que su tiempo, el tiempo en el que completa los 100m, sea inferior a trece punto cinco? ¿Y también sea menos de 14 segundos? Bueno, si su tiempo es menos de 13.5 y menos de 14, esto es lo mismo que decir que es simplemente menos de 13.5. Porque si es menos de trece coma cinco, entonces ya sabemos que es menos de 14. Entonces queda una probabilidad que ya sabemos a qué es igual:

$$P\big((T < 13.5) \cap (T < 14)\big) = P(T < 13.5) = 0.048$$

Y ahora hay que dividir por la probabilidad de que su tiempo sea inferior a 14:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P\big((T < 13.5) \cap (T < 14)\big)}{P(T < 14)} = \frac{P(T < 13.5)}{P(T < 14)} = \frac{0.048}{0.106} = 45\%$$

Esta probabilidad la podemos hallar con la tabla de la Normal y es igual a: 0.106. Cuando hacemos la división obtenemos un 45%.

Entonces, nuevamente, la evidencia nos hizo actualizar significativamente nuestros pensamientos sobre si Sofía puede correr 100 metros en menos de trece coma cinco segundos. Pensábamos que estaba alrededor del 4.8% antes, pero ahora que sabemos que ella está en el equipo de corredores de la escuela, tenemos evidencias que nos obligan a actualizar nuestra visión. Y ahora tenemos que decir que la probabilidad de que pueda correr en menos de trece coma cinco segundos es del 45 por ciento.

CONTRADICCIONES CON LA NORMAL

Vamos a echar un vistazo a algunos resultados que van a terminar siendo contradictorios aunque podrían derivarse de una buena comprensión de la probabilidad condicional. Vamos a considerar el coeficiente intelectual (intelectual coeficient IQ). Ahora bien, se sabe que el coeficiente intelectual se distribuye normalmente, tiene una distribución Normal con una media de 100 que es el coeficiente intelectual promedio, y una desviación estándar de 15. Lo que significa que la mayoría de la gente tiene un coeficiente intelectual dentro del rango de los 15 puntos para arriba y para abajo alrededor de 100. Ahora, para ser considerado un genio, necesitas tener un coeficiente intelectual de más de 171. Supongamos que hay una píldora que podemos tomar y esta píldora, todo lo que hace es cambiar el promedio ligeramente en la distribución del IQ. Aquellos que toman la píldora tienen una distribución diferente, con una media que es 2 puntos mayor que la otra.

$$IQ \sim N(\mu = 100,\ \sigma = 15^2)$$

$$IQ^+ \sim N(\mu = 102,\ \sigma = 15^2)$$

Y ahora nos hacen esta pregunta: ¿Cuánto más probable es que tu hijo sea un genio si toma esta píldora?

Primero vamos a pensar qué nos dice la intuición. Pues a la mayoría de la gente lo que la intuición le dice es que no va a tener un efecto particularmente grande en la probabilidad. Que agregar dos puntos en el promedio no va a tener un efecto significativo en hacer a la persona más propenso a ser un genio. Vamos a investigar esta afirmación. Vamos a imaginarnos realizando el siguiente experimento. Cogemos a un grupo grande de niños y le daremos la pastilla a la mitad de ellos. Esta píldora aumentará dos puntos el coeficiente intelectual promedio de estos niños que la han tomado. Los dejaremos crecer y luego veremos cuáles de ellos resultaron ser genios. Este experimento mental nos ayudará a responder a la pregunta. Porque esto realmente es una pregunta sobre una probabilidad condicional. ¿Cuál es la probabilidad de que hayan tomado la píldora dado que son genios? En realidad, esa es la pregunta que queremos responder. Dado que son genios, es decir, miramos a todos los niños que resultan ser genios y entonces preguntamos, ¿cuántos de ellos realmente tomaron esta pastilla roja? Nuestra hipótesis es que tomaron la píldora roja cuando eran niños. Eso es lo que nos interesa. Y nuestra evidencia es que son genios.

H: Tomaron la píldora cuando eran niños.

E: Son genios.

Entonces, sabemos que lo que debemos hacer primero es encontrar la probabilidad de H y E. Que es la probabilidad de que tomaron la píldora roja y se convirtieron en genios y luego lo dividiremos por la probabilidad de que sean genios en general.

$$P(H|E) = \frac{P(H \cap E)}{P(E)}$$

Entonces, ¿cuál es la probabilidad de la intersección, de que tomaran la píldora roja y fueran genios? Bueno, a la mitad de los niños se les dio esa pastilla roja. Por tanto ponemos 0.5 multiplicado por la probabilidad de que, de ese grupo de los que recibieron la pastilla roja, se conviertan en genios. Bueno, recuerda la distribución para los que tomaron la pastilla: $IQ^+ \sim N(\mu = 102,\ \sigma = 15^2)$

Entonces, nuestro tipo de IQ plus tenía una distribución Normal con una media de 102 y una desviación estándar de 15. Y realmente lo que estamos haciendo es responder a la pregunta, ¿cuál es la probabilidad de que este tipo de IQ plus (de este grupo de personas que tomó la pastilla) es mayor que ciento setenta y uno (condición para ser genio)?

$$P(IQ^+ > 171) = 0.000000211$$

Y eso resulta ser bastante pequeño. Pensemos ahora en cuál es la probabilidad de que, en general, un niño sea un genio. Bueno, depende de si tomaron la pastilla o no. Si tomaron la píldora, es lo mismo que tenemos arriba que acabamos de hallar. Y si no la tomaron sería un 50% multiplicado por la probabilidad de que sean genios si no han tomado la pastilla, para lo cual usamos la otra distribución con media 100: $IQ \sim N(\mu = 100,\ \sigma = 15^2)$. Y eso es igual a: $P(IQ > 171) = 0.\ 00000011$.

Entonces quedaría:

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{0.5\ \times 0.000000211}{0.5\ \times 0.000000211 + 0.5\ \times 0.\ 00000011} = 0.66 = 66\%$$

En total esto da 66%. Esto quiere decir que el 66% de los niños que son genios, han tomado la pastilla. Es decir la probabilidad de que hayan tomado la pastilla dado que se sabe que son genios es del 66%.

Pensemos en esto por un momento. Lo que esto significa es que de todos esos genios en nuestro experimento, el número de personas que tomaron la pastilla roja es aproximadamente el doble del número de personas que no lo hicieron, que no se la tomaron. Sin embargo, esta píldora aumenta el coeficiente intelectual promedio ¡solamente en dos puntos! Y para ser un genio, necesitas un coeficiente intelectual muy superior, mayor que 171. Sin embargo, este simple y pequeño aumento en el promedio,

duplica las posibilidades de ser un genio. Aunque esto sea un ejemplo un poco artificial, lo hemos visto para darnos cuenta de que el uso de la probabilidad condicional puede ayudarnos a comprender algunos resultados realmente contra intuitivos.

Ahora, imaginemos que hubiera una píldora diferente. Y esta píldora, en lugar de aumentar las posibilidades de un niño tenga en promedio coeficiente intelectual de 100 a 102, vamos a cambiar la desviación estándar de 15 a 20, lo que significa que la mayoría de los niños tienen un coeficiente intelectual dentro del rango de 20 puntos alrededor de 100:

$$IQ^{+} \sim N(\mu = 100,\ \sigma = 20^{2})$$

Ahora bien, ¿qué efecto tiene eso? Bueno, este caso, en realidad es bastante asombroso. No vamos a pasar por todos los cálculos esta vez, pero siguiendo exactamente las mismas líneas que antes, veremos que al tomar esta píldora, que no hace más que modificar la densidad alrededor del promedio, el resultado es un asombroso 99.4%:

$$P(H|E) = 99.4\%$$

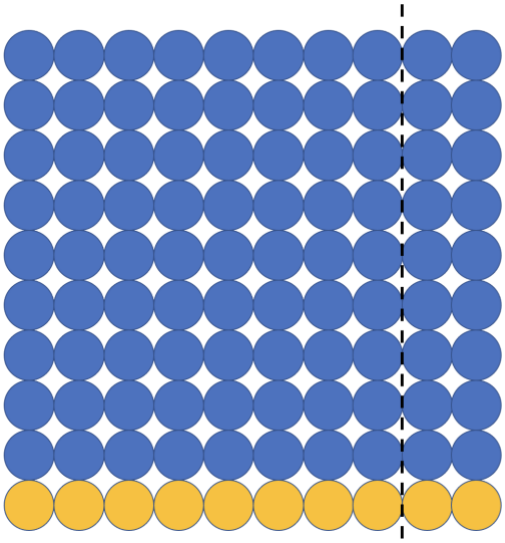
Ahora bien, esto no significa que si tomamos la píldora tenemos un 99.4% de posibilidades de ser un genio. Esto significa que si miramos a todos los genios de nuestro estudio y luego preguntamos, ¿cuál es la probabilidad de que un genio seleccionado al azar, esa persona haya tomado esta pastilla? Eso es del 99.4%, un porcentaje altísimo.

TEOREMA DE BAYES

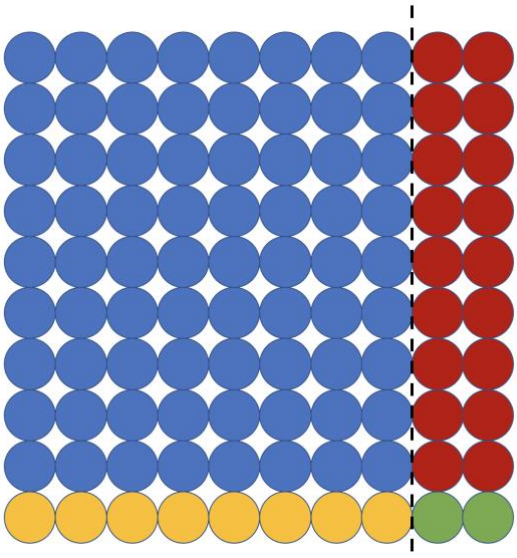
APLICACIÓN A ENFERMEDADES

Ahora tenemos suficiente información y hemos desarrollado nuestra intuición como para abordar el Teorema de Bayes en detalle. Y lo haremos mirando uno de los ejemplos más famosos. Supongamos que una enfermedad X afecta al 10 por ciento de las personas donde nosotros vivimos. Ahora supongamos también que hemos experimentado algunos síntomas. Vamos al médico, nos realizan una prueba y la prueba dice que tenemos la enfermedad. Ahora, esta prueba que nos hacen para detectar la enfermedad, en realidad no es del 100 por ciento fiable, se puede equivocar y dar un falso positivo. Así que en realidad la prueba da el resultado correcto el 80 por ciento de las veces. Queremos saber ¿cuál es la probabilidad de que realmente tengamos la enfermedad? La intuición de la mayoría de la gente será que esta probabilidad es del 80 por ciento. Antes pensábamos que era el 10 por ciento, pero eso es la probabilidad de que una persona aleatoria tenga la enfermedad, es un pensamiento previo. Ahora nos han hecho la prueba que dice que tenemos la enfermedad. Y esa prueba es un 80 por ciento fiable. La mayoría de las personas seguramente actualizarían su criterio y dirían que ahora hay un 80% de probabilidades de tenerla. Vamos a comprobar si esto es correcto o no.

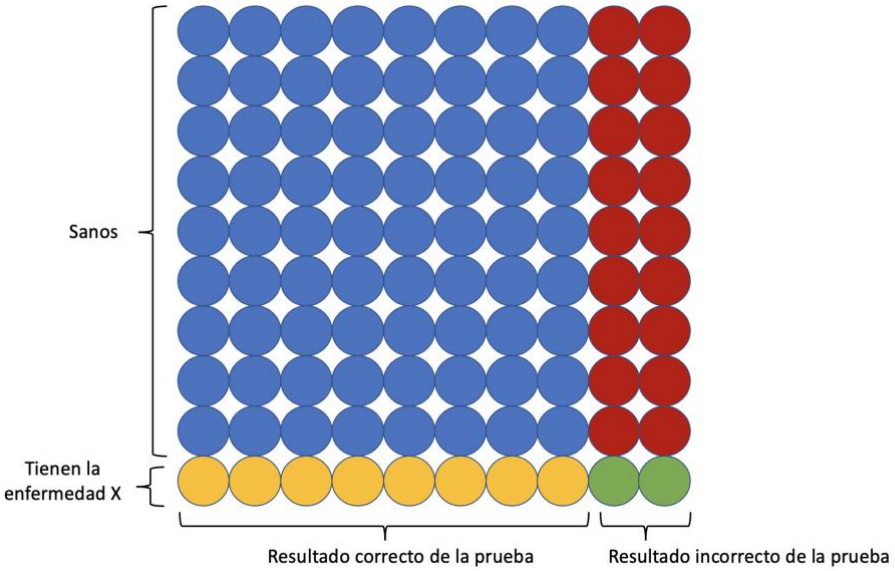
Lo primero que vamos a hacer es imaginar una muestra representativa de la sociedad.



Ahora, sabemos que el 10 por ciento de las personas tienen la enfermedad. La bolas amarillas que están en la última fila, representan a esas 10 personas de 100 (10%). En realidad están distribuidas de manera aleatoria entre la población pero vamos a ponerlas abajo para que podamos tener una buena imagen de lo que está sucediendo. También sabemos que al 20% de las personas examinadas se les dice la respuesta incorrecta (porque el 80% obtenía una respuesta correcta en la prueba). Vamos a dibujar una línea discontinua para representar esa separación:



A las personas a la derecha de esa línea, que es el 20 por ciento de nuestra muestra, se les diagnostica incorrectamente. Vamos a ver realmente cuáles son estos diferentes grupos. Las bolas azules son personas que están sanas y les diagnosticaron correctamente. Las bolas amarillas de abajo son personas que están enfermas y también les diagnostican correctamente, les dicen que efectivamente tienen la enfermedad. Las bolas rojas son otras personas que están sanas, pero les dijeron incorrectamente que tenían la enfermedad, el test se ha equivocado. Y las dos bolas verdes son los que tienen la enfermedad, pero desafortunadamente, la prueba les dijo que no la tienen, se equivoca también:



Entonces podemos clasificarlos en estos grupos, los sanos, los que tienen la enfermedad, las personas en las que el test no se equivoca y las personas en las que el test sí se equivoca.

La prueba que nos han realizado a nosotros nos ha dicho que tenemos la enfermedad. Entonces, ¿en cuál de los grupos de personas podríamos estar? Bueno, no somos del grupo azul porque ellos están sanos y les dijeron que están sanos. Entonces no somos uno de ellos. Tampoco somos los verdes, porque ellos tienen la enfermedad y le dicen que no la tienen, y a nosotros el test nos ha dado positivo, no negativo. Por lo cual no somos de ese grupo. Eso significa que estaremos entre los grupos amarillo o rojo. Entonces la pregunta es, ¿en qué grupo estoy?

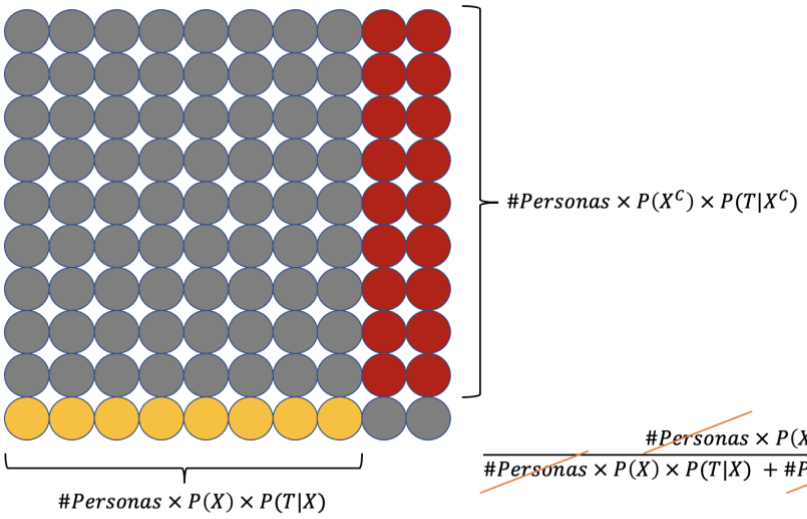
Bueno, hay ocho personas que tienen la enfermedad y les dicen que la tienen. Y hay 18 personas que no tienen la enfermedad a quienes les dicen que la tienen. Entonces, ¿cuál es la probabilidad de que tenga la enfermedad? Bueno, son 8 (porque hay 8 que la tienen) sobre el total:

$$\frac{8}{8+18} = \frac{8}{26} = 31\%$$

Así que volvamos a la pregunta original. El médico hace la prueba. Me dice que tengo la enfermedad. Sabemos que esta enfermedad afecta al 10 por ciento de las personas donde nosotros vivimos. Sabemos que esta prueba es un 80 por ciento fiable. Pero, solo hay un 31 por ciento de probabilidades de que yo realmente tenga la enfermedad. Nuestra comprensión previa de tener la enfermedad, que era del 10 por ciento, se actualizó mediante esta prueba que mayoritariamente es confiable. Y ahora sabemos que la probabilidad es del 31 por ciento. Es decir que probablemente no tenga la enfermedad. Porque la probabilidad de lo contrario, que es que no la tenga, es del 69%. Bastante mayor.

Vamos a juntar ahora todas cosas de nuevo y ver si podemos formalizar esto a través de fórmulas.

Así que, lo primero que hicimos fue tomamos el número de personas de nuestra muestra (100) y lo multiplicamos por la probabilidad de tener la enfermedad, que era del 10%. Entonces esto redujo las cosas a las 10 personas de abajo. Luego se introduce la probabilidad de tener una prueba correcta, que es 80% lo que se reduce a 8 personas. Y P(T|X) es la probabilidad de que el test sea positivo dado que tenemos la enfermedad. Entonces estos son los círculos en amarillo.



$$\frac{\#Personas \times P(X) \times P(T|X)}{\#Personas \times P(X) \times P(T|X) + \#Personas \times P(X^C) \times P(T|X^C)}$$

Teorema de Bayes → $$\frac{P(H) \times P(E|H)}{P(H) \times P(E|H) + P(H^C) \times P(E|H^C)}$$

Ahora, ¿qué pasa con la gente roja? ¿De dónde vienen?

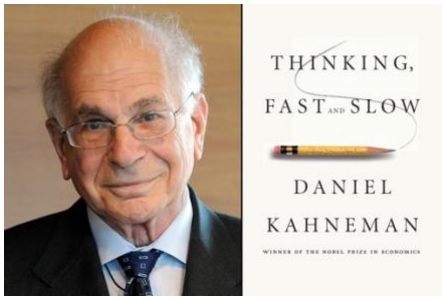
Bueno, para calcularlos, tomamos el número de personas de nuestra muestra. Multiplicamos por la probabilidad de no tener la enfermedad. Que eso fue 0.9. Porque hay un 10% que sí la tiene, por tanto un 90% que no la tiene. Luego multiplicamos por la probabilidad de obtener un resultado positivo en la prueba, dado que no tienes la enfermedad, que fue 0.2 (el % de equivocación de la prueba). Esto es lo que nos dio ese grupo de 18 bolas rojas.

Así que si ahora sustituimos todo esto tenemos una expresión donde podemos simplificar al # de personas. Por lo que vemos que realmente no es relevante cuántas personas estemos considerando en esa muestra. Y entonces la expresión se queda reducida a esto:

$$\frac{P(H) \times P(E|H)}{P(H) \times P(E|H) + P(H^C) \times P(E|H^C)}$$

La evidencia es sobre el test que nos han hecho y ha dado positivo. Y la hipótesis que queremos averiguar es que estemos realmente enfermos, es decir, que tengamos la enfermedad. Y esto es el Teorema de Bayes.

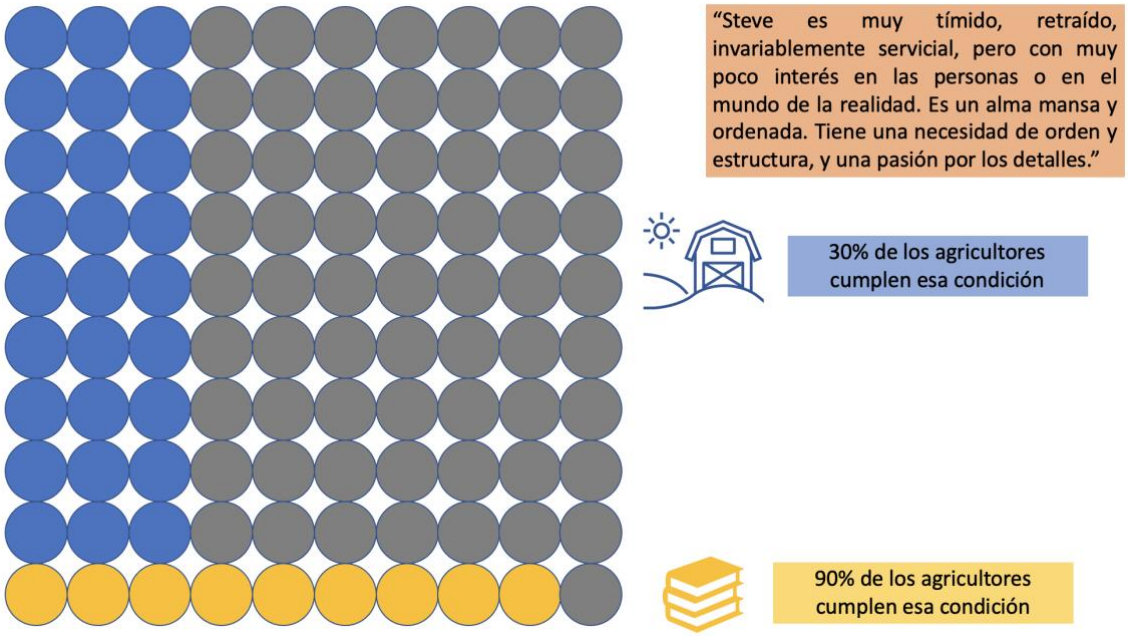
EJEMPLO DANIEL KAHNEMAN



Vamos a ver un ejemplo que nos ayudará a entender mucho mejor el Teorema de Bayes. El ejemplo que vamos a ver es uno que se hizo famoso por Daniel Kahneman que habló de ello en su libro, Thinking Fast and Slow (Pensar rápido pensar despacio).

En el ejemplo tenemos a Steve. Nos dicen que “Steve es muy tímido, retraído, invariablemente servicial, pero con muy poco interés en las personas o en el mundo de la realidad. Es un alma mansa y ordenada. Tiene una necesidad de orden y estructura, y una pasión por los detalles.”

Ahora, esta es la pregunta que Kahneman no propone: ¿crees que es más probable que Steve sea agricultor o bibliotecario? Bueno, la intuición de todos dirá que es mucho más probable que sea bibliotecario. Veamos si podemos usar el Teorema de Bayes para socavar esto, para descubrir realmente qué está pasando con más detalle. Entonces, lo que todos nos olvidamos de hacer es considerar cuántos agricultores y cuántos bibliotecarios realmente hay. En realidad los números exactos no importan mucho pero sí en comparación, hay muchos más agricultores en el mundo que bibliotecarios. De hecho, en casi todos los países, porque Steve puede ser de cualquier lugar, probablemente haya más agricultores que bibliotecarios. En este ejemplo vamos a suponer que hay 10 veces más. Por otro lado tenemos la evidencia sobre Steve, que es tímido y retraído. Le gusta ayudar, le gusta el orden y la estructura. Entonces podríamos preguntar ¿qué fracción de agricultores cumplirían estas características? Bueno, vamos a decir que el 30 por ciento. Es un porcentaje suficientemente bajo de agricultores que cumplen con este criterio. Sin embargo estas características obviamente, serán mucho más probables entre los bibliotecarios, así que digamos que el 90 por ciento de los bibliotecarios cumplen esta condición.



Bueno, Steve debe ser uno de los granjeros azules o uno de los bibliotecarios amarillos. Porque esas son las personas que cumplen esa condición. ¿Cuál es la proporción de agricultores mansos, ordenados y tímidos con respecto a los bibliotecarios mansos y ordenados, y tímidos? Sería una proporción de 27 contra 9.

En términos del Teorema de Bayes:

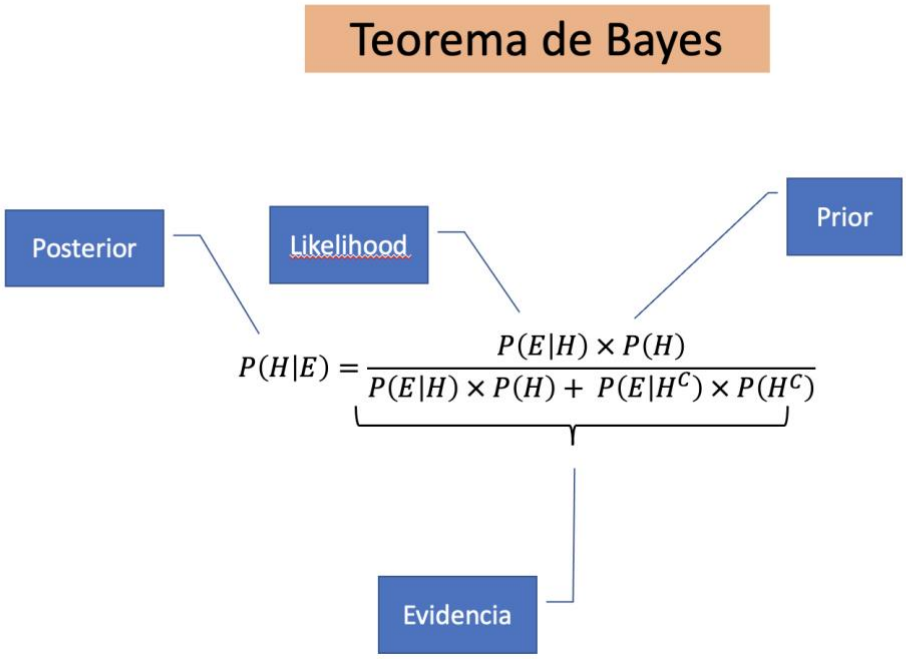
H: Steve es un bibliotecario

E: Steve es tímido, ordenado, etc…

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E|H) \times P(H) + P(E|H^C) \times P(H^C)} = \frac{\frac{9}{10} \times \frac{10}{100}}{\frac{9}{10} \times \frac{10}{100} + \frac{27}{90} \times \frac{90}{100}} = 23\%$$

Entonces, ¿es más probable que sea agricultor o bibliotecario? Es mucho más probable que sea agricultor (77%) que bibliotecario (23%). Y esto va en contra de nuestras intuiciones, porque lo que olvidamos es cuántos agricultores hay, en comparación con el otro grupo. De hecho aquí veríamos que en realidad Steve tiene tres veces más probabilidades de ser agricultor que bibliotecario.

En el caso de nuestros bibliotecarios y agricultores, pensaríamos, ¿cuál es la probabilidad de que una persona seleccionada al azar sea bibliotecaria? Bueno, en nuestra muestra, los bibliotecarios eran 10 de 100. Eso es lo que creemos que es la probabilidad "a priori", antes de cualquier evidencia. Ahora, en contraste con eso, la que estamos hallando se llama probabilidad "a posteriori", después de conocer la evidencia. Y hay otra parte que se le llama likelihood en inglés, o verosimilitud en español. Es la probabilidad de ver la evidencia, dado que la hipótesis es cierta. Y en el caso de nuestros bibliotecarios y agricultores, esta es esencialmente la fracción de bibliotecarios que son tímidos, ordenados, etc., que fue del 90 por ciento en ese caso. Finalmente, al denominador se le llama probabilidad de la evidencia. A veces también se le llama verosimilitud marginal, o probabilidad total. Es como la probabilidad general de ver la evidencia que hemos visto. Y ese denominador simplemente se descompone en la posibilidad de ver la evidencia si la hipótesis es cierta y la probabilidad de ver la evidencia si la hipótesis no es cierta.

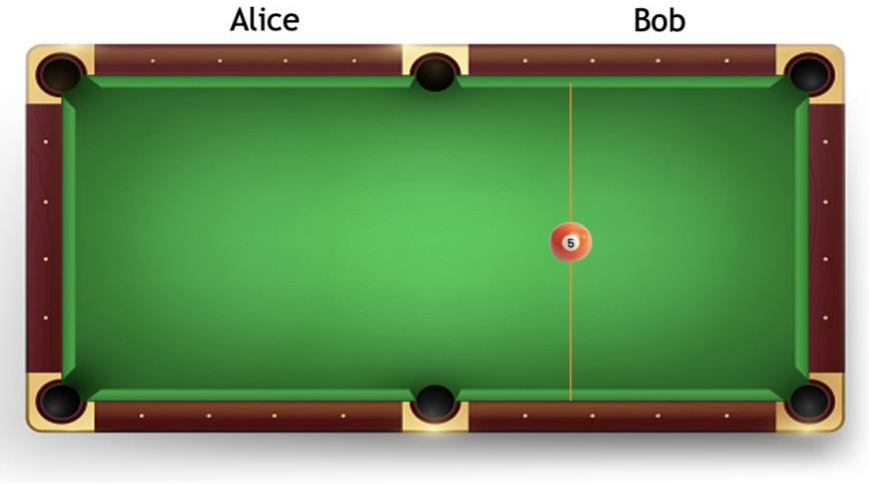


A menudo es más fácil escribir el denominador en forma abreviada de la probabilidad de E. Pero hay situaciones y casos en los que la única forma de calcular la probabilidad de E es dividirla en esas dos posibilidades. Por lo tanto, a veces lo podremos ver de una forma y otras veces de otra, a veces con la P(E) y otras veces dividida en partes como aquí.

PUZZLE DE BAYES DEL JUEGO DE BILLAR

Vamos a ver un ejemplo sustancialmente más complicado, el ejemplo dado por el mismo Bayes en sus escritos sobre esto cuando estaba desarrollando por primera vez esta forma de pensar. Vamos a adaptarlo un poco para hacerlo más simple y más sencillo de entender. Esta adaptación fue realizada por S. R. Eddie en su artículo *What is Bayesian Statistics*, Nature Biotechnology, Volume 22, No. 9. 2004.

Lo que vamos a hacer es imaginar un juego entre Alice y Bob. El juego comienza con esto: una bola rueda hacia una posición aleatoria en la mesa. Esto divide la tabla en dos secciones: A y B.



Alice y Bob juegan entonces al siguiente juego. Las bolas se van a hacer rodar sobre la mesa al azar. Si aterrizan del lado de Alice, ella obtiene un punto. Si aterrizan del lado de Bob, él obtiene un punto. Y el primero que llegue a 6 puntos gana. Pero hay un detalle importante, tenemos que asumir que Alice y Bob, no pueden ver la mesa. Y eso sería también fingir que nosotros tampoco podemos ver la mesa. Así que no sabemos cómo se dividió en primer lugar.

Entonces, digamos que comienza el juego y Alice obtiene un punto rápido. Luego la bola vuelve a caer de su lado y Alice obtiene otro punto porque la pelota aterriza allí. Luego imaginemos que en el siguiente paso le toca a Bob, y gana un punto, y así continúa el juego, y luego otro punto para Alice, y luego otro para Bob. Es decir que todo lo que sabemos son los puntos que se están obteniendo, no dónde exactamente aterrizan las bolas. Así que vamos a suponer que el juego llega a un punto, en el que **Alice tiene 5 puntos, y Bob tiene 3**. Hagamos una pausa en este momento y ahora vamos a preguntarnos: **¿Cuál es la probabilidad de que Bob gane este juego?**

Todo lo que tenemos son los resultados. Entonces, ¿cómo diablos podemos resolver este problema?

Bueno, lo primero sobre lo que vamos a reflexionar es: ¿cuál es la probabilidad de que Bob obtenga un punto en cualquier intento? Este será un punto de partida muy útil para que pensemos. Entonces, vamos a comenzar pensando en este problema desde un punto de vista frecuentista. Si recuerdas al principio observamos la diferencia entre los métodos frecuentistas y los métodos bayesianos. Y este es un gran ejemplo para ver ambos enfoques y ver cómo el método bayesiano puede resolver el problema de manera más elegante y precisa. Así que vamos a ver la solución que plantearían las pruebas frecuentistas. La pregunta es la siguiente: ¿Cuál es la probabilidad de que Bob obtenga un punto? Bueno, Bob tiene tres de los puntos hasta ahora jugados, es decir, de los ocho puntos que se han ganado. Así que la probabilidad de que Bob gane un punto sería:

$$\frac{3}{8}$$

Ahora, para ganar el juego, necesitará ganar los siguientes tres puntos, porque el juego se gana cuando se tienen 6 puntos. Entonces, la probabilidad de que eso suceda es de tres octavos elevados al cubo, según el **enfoque frecuentista**:

$$\left(\frac{3}{8}\right)^3 = 5.3\%$$

Veamos ahora el enfoque bayesiano:

¿Cuál es la probabilidad de que Bob gane? Vamos a usar la notación que ya nos es familiar. Nuestra hipótesis H es que Bob gana el juego. Y nuestra evidencia E es que la puntuación actual es cinco puntos para Alice y tres para Bob. Intentemos aplicar el Teorema de Bayes a esto.

- H: Bob gana
- E: a=5, b=3

¿Entonces que sabemos? Bueno, sabemos que la probabilidad de que Bob gane, dado que Alice tiene cinco puntos y Bob tres, es una aplicación directa del Teorema de Bayes. Es la probabilidad de ver una puntuación como esta dado que Bob gana multiplicado por la probabilidad de que Bob gane, dividido por nuestra evidencia, que es la probabilidad de encontrar una puntuación de cinco para Alice y para Bob de tres:

$$P(Bob\ gana|a=5,\ b=3\ ) = \frac{P(a=5,\ b=3\ |Bob\ gana) \times P(Bob\ gana)}{P(a=5,\ b=3)}$$

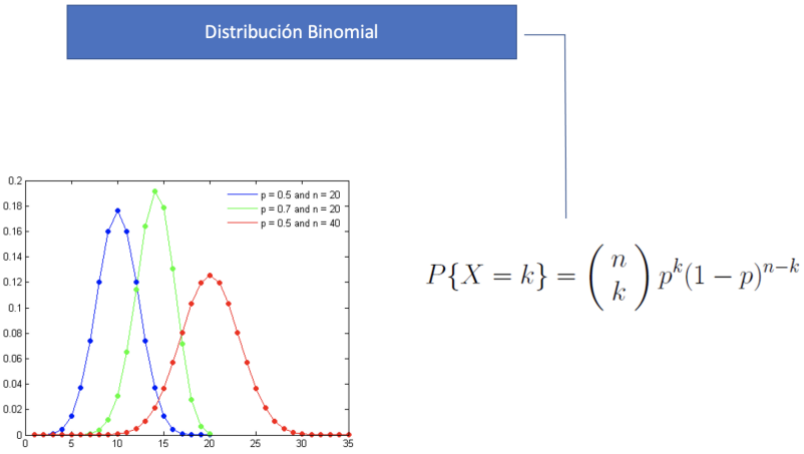
Comencemos entonces con el denominador: ¿cuál es la probabilidad de ver una puntuación de cinco puntos para Alice y tres para Bob?

Ahora, recordemos que no sabemos dónde está la línea de separación. No sabemos cuál es la probabilidad de que Bob gane un juego determinado. Esto depende de esa línea porque si el espacio de Bob es mucho más grande es más probable que gane él. Pero esa probabilidad de que Bob por ejemplo gane un punto cualquiera en general, podría ser cualquier cosa desde cero hasta uno, y nosotros no la conocemos. Entonces, a la luz de eso, tendremos que pensar, ¿cuál es la probabilidad de ver una puntuación de cinco versus tres a favor de Alice, en general, y teniendo en cuenta cualquier posibilidad y probabilidad de que Bob gane un punto. Este es el denominador de la fórmula.

Vamos a nombrar a la probabilidad de que Bob gane un punto como x, y a la probabilidad de ver una puntuación de cinco versus tres a favor de Alice bajo cualquier valor de la probabilidad de que Bob gane un punto, cualquier valor de x, la vamos a llamar y. Esto según la distribución Binomial sería esta fórmula de aquí que depende de x, porque no hemos fijado un valor para esa probabilidad de que Bob gane un punto sino que queremos considerar cualquier posible valor para ella, entre 0 y 1 claro.

$$y = \binom{8}{3} x^3(1-x)^5$$

$\binom{8}{3}$ es la cantidad de posibilidades o combinaciones de 3 en 8, porque Bob ha ganado 3 puntos de los 8 puntos en total (Alice tiene los otros 5).



Ahora queremos saber cuál es la suma de todas estas posibilidades diferentes para todos los valores de x de cero a uno. Si nosotros tenemos una función que representa las probabilidades de una variable en este caso los valores que varían son los de x, y queremos saber cuál es la suma de todos los resultados posibles para un número infinito de valores, podemos hacerlo mediante la integral de esa función, entre cero y uno que es donde están acotados los posibles valores de x. Esta integral da como resultado:

$$P(a = 5,\ b = 3) = \int_0^1 \binom{8}{3} x^3(1-x)^5 dx = \frac{1}{9}$$

Entonces, lo que esto nos dice es que, en general, hay 1/9 posibilidades de jugar este juego y que en algún momento del juego, nos encontráramos con esa posición de que Alice tiene cinco puntos y Bob tres. Bien, entonces ya tenemos el valor del denominador: 1/9. Ahora sigamos y echemos un vistazo al numerador.

Vamos a simplificar algunas cosas para poder ayudar un poco a la intuición en esta parte. Entonces, la primera parte es la probabilidad de que A sea igual a cinco y B sea igual a tres, dado que Bob gana, ¿podemos idear una función que pueda capturar eso? Bueno, sabemos que la función que describe que Alice tiene cinco puntos y Bob tres es esta expresión binomial. Si sabemos que Bob gana el juego (están condicionando a eso), entonces Bob necesita ganar tres puntos porque se gana cuando se tienen 6 puntos. Entonces habría que agregar una x al cubo que representaría el suceso de que Bob gana un punto, y el siguiente y el siguiente (3 seguidos):

$$\binom{8}{3} x^3(1-x)^5 \ \ x^3 = \binom{8}{3} x^6(1-x)^5$$

Ahora, como antes, vamos a usar la integral de esta función sobre todos los posibles valores de x entre 0 y 1:

$$P(a = 5,\ b = 3\ |\textit{Bob gana}) \ \times\ P(\textit{Bob gana}) = \int_0^1 \binom{8}{3} x^6(1-x)^5 dx = \frac{1}{99}$$

Ya tenemos numerador y denominador:

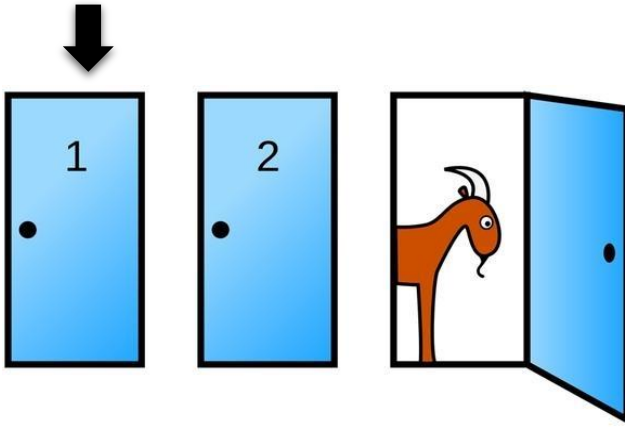
$$P(\textit{Bob gana}|a = 5,\ b = 3\ ) = \frac{P(a = 5,\ b = 3\ |\textit{Bob gana}) \ \times\ P(\textit{Bob gana})}{P(a = 5,\ b = 3)} = \frac{\frac{1}{99}}{\frac{1}{9}} = \frac{1}{11} = 0.09 = 9\%$$

Entonces usando un enfoque frecuentista obteníamos que era 5 %, mientras que el resultado con el enfoque bayesiano es casi el doble, un 9%. Ahora, ¿quién tiene razón? Bueno, durante las clases prácticas del curso vimos que cuando hacíamos simulaciones de este ejemplo, el resultado era que el enfoque bayesiano es el acertado.

MONTY HALL

El problema de Monty Hall, es un acertijo estadístico cuya resolución no es tan intuitiva si no está muy familiarizado con la teoría de la probabilidad o incluso a veces cuando sí lo estamos. El problema consiste en el siguiente juego. Imagina que tú eres el jugador, y estás frente a tres puertas cerradas. Detrás de dos de esas puertas hay una cabra, (una para cada puerta), mientras que detrás de la tercera puerta hay un coche.

Lo que se te pide es que elijas una puerta para ver si te puedes llevar lo que estaba detrás de esa puerta (lo que a todos nos gustaría es elegir la puerta que tiene el coche). Entonces, como primer paso, se te pide que elijas una puerta entre las tres puertas. Pero imagina que esto es un juego de televisión y hay un presentador que sabe lo que hay detrás de todas las puertas. Y entonces una vez que has elegido la puerta, en vez de abrirla, el presentador te abre otra de las dos puertas restantes que tú no has elegido. Como en esas dos puertas el presentador sabe lo que hay detrás, y siempre va a haber al menos una cabra, él no te va a abrir la puerta que tiene el coche obviamente, si es que la hay. Sino que te abre la puerta que tiene una cabra.



Entonces, después de abrirte una puerta y descubrir una cabra, el presentador te da dos opciones: puedes plantarte y seguir con la puerta que habías elegido antes, o cambiar y elegir la otra puerta que ha quedado. Ahora la pregunta es: una vez que sepas que una de las dos puertas que no habías elegido, escondía una cabra. ¿Qué harías? ¿Quedarte con la misma puerta o cambiar? El pensamiento intuitivo de mucha gente es que es mejor quedarnos con la que habíamos elegido. Pero ¿habrá más o menos probabilidad de que, cambiando de decisión, esa otra puerta sea la que esconde el coche? La respuesta es que es mejor cambiar de puerta. Porque si cambiamos, tenemos una probabilidad mayor de ganar. Veamos por qué.

La explicación formal de esto sería usar el Teorema de Bayes. Pero veamos primero el enfoque desde un razonamiento más intuitivo. Al comienzo del juego, hay 1/3 de probabilidad de abrir la puerta que esconde el coche (todas las puertas tienen 1/3 de probabilidad de esconder un coche). Así que escogemos una puerta y, enseguida, el presentador nos abre una de las otras dos puertas restantes y vemos que

escondía una cabra. Ahora, todos los posibles escenarios según si te quedas con la puerta elegida antes, o cambias, son los siguientes:

Si no cambiamos:

- Escogiste la puerta escondiendo el auto: **ganas**
- Escogiste una de las puertas que escondían la cabra: pierdes
- Escogiste la otra puerta que esconde la cabra: pierdes

Si cambiamos:

- Escogiste la puerta escondiendo el auto: pierdes
- Escogiste una de las puertas que escondían la cabra: **ganas**
- Escogiste la otra puerta que escondía la cabra: **ganas**

Es decir que, si cambias, de tres posibilidades tienes 2 de ganar. La victoria proviene del hecho de que si elegiste la puerta que escondía a la cabra y la otra que escondía a la otra cabra se abre, entonces al cambiar ¡revelarás la puerta que esconde el coche! Con el primer escenario, en el que no cambiamos de puerta, la probabilidad de ganar es 1/3, mientras que el escenario cambiamos de puerta, ganaríamos con probabilidad 2/3. Por lo tanto, si cambiamos de puerta, se dobla la probabilidad de ganar inicial, es decir, ¡es dos veces más probable que ganemos!

En base a los cálculos con el Teorema de Bayes:

Vamos a asumir que elegimos la puerta 1 y el presentador abre la puerta 3 que tiene una cabra.

- H: Ganar quedándonos con la puerta 1 (coche en puerta 1).
- E: abren la puerta 3

Probabilidad a priori $P(H)$ que el coche esté en la puerta 1:

$$P(H) = 1/3$$

Likelihood o verosimilitud $P(E|H)$:

- Si el coche está en la puerta 1, abrirán la puerta 2 o 3, así que la probabilidad de que abran cualquiera de las dos puertas es un 50%.
- Si el coche está en la puerta 2, no la van a abrir, así que sólo abriría la puerta 3 porque la 1 la hemos abierto nosotros.

Es decir:

- Si la puerta 1 tiene el coche:

$$P(E|H) = 1/2$$

- Si la puerta 1 NO tiene el coche:

$$P(E|H^C) = 1$$

Entonces:

$$P(E) = P(E|H)P(H) + P(E|H^C)P(H^C) = \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1 = \frac{1}{2}$$

Y finalmente:

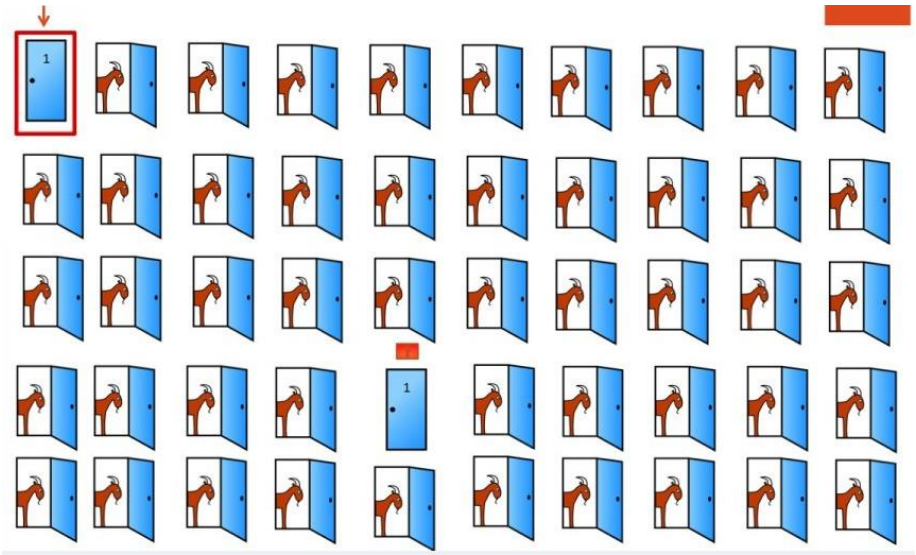
- Probabilidad de ganar si nos quedamos con puerta 1:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} = \frac{\frac{1}{3} \times \frac{1}{2}}{1/2} = 1/3$$

- Probabilidad de perder si nos quedamos con puerta 1:

$$P(H^C|E) = \frac{P(E|H^C) \times P(H^C)}{P(E)} = \frac{\frac{1}{3} \times 1}{1/2} = 2/3$$

Una forma más clara de verlo es replantear el problema. Si en lugar de haber sólo tres puertas hubiese 100 puertas, y tras la elección original el presentador abriese 98 de las restantes para mostrar que tras todas ellas hay cabras, si no cambias tu elección ganaría el coche sólo si lo has escogido bien (1 de 100 posibilidades), mientras que si la cambias, ganarías si precisamente no has escogido bien (lo que puede suceder en un 99 de 100 posibilidades), ¡99 de cada 100 veces ganaríamos! ¿No es obvia la elección?



EJEMPLO DE CORREOS SPAM

El **spam** (o correo basura) es uno de los grandes problemas tanto en empresas como en individuos. El uso de filtros para intentar controlar el tráfico de estos correos es muy importante. Uno de los filtros más eficientes que se conocen son los **filtros bayesianos**. Este filtro está basado en el Teorema de Bayes para determinar un correo electrónico como spam o no. La probabilidad de que un correo electrónico sea spam, considerando que haya ciertas palabras en él, es igual a la probabilidad de encontrar esas ciertas palabras en un correo basura por la probabilidad de que algún correo sea spam, dividido entre la probabilidad de encontrar esas palabras en algún correo.

El filtro bayesiano necesita una base de datos que contenga palabras y otros criterios (direcciones IP, hosts,…), para calcular la probabilidad de que un correo determinado sea spam, sacados de un ejemplo de correo basura y de correo válido. A cada palabra se le establece un valor de probabilidad basado en la frecuencia de aparición de dicha palabra en un correo basura frente a un correo válido. Estas asignaciones se realizan a través de un proceso de análisis del correo. Por ejemplo, si la palabra *viagra* aparece en 600 de 2000 correos de spam y en 3 de 200 correos válidos, la probabilidad de ser spam es 0.9524. De este modo el filtro bayesiano se adapta al usuario, pues si se trata de una empresa de software tratará con mayor probabilidad de spam la palabra *viagra* que por ejemplo una empresa de productos farmacéuticos, que posiblemente la frecuencia de aparición de *viagra* en sus correo válidos sea mayor, por lo que la probabilidad de que sea spam disminuye.

Disponiendo de la base de datos el filtro podrá actuar. Cuando se recibe un nuevo correo, el análisis consiste en descomponer el texto en palabras y se seleccionan las más relevantes, las cuales el filtro bayesiano procesará calculando la probabilidad de que el correo que hemos recibido sea spam o no. Si la probabilidad supera un umbral establecido se considerará spam. Si estás interesado, aquí hay un artículo en PDF http://cs.wellesley.edu/~anderson/writing/naive-bayes.pdf que explica un par de enfoques para combinar los resultados de varias palabras.

EJEMPLO DEL FALLO DE LA ALARMA

La probabilidad de que haya un accidente en una fábrica que dispone de alarma es 0.1. La probabilidad de que esta suene si se ha producido algún incidente es de 0.97 y la de que suene si no se ha producido ningún indicente es de 0.02. En el supuesto de que haya funcionado la alarma, ¿cuál es la probabilidad de que no haya habido ningún incidente?

Sean los sucesos:

I = Producirse incidente

A = Sonar la alarma

Los datos que tenemos son:

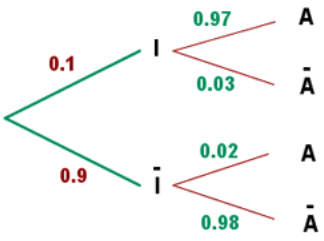
$$P(\bar{I}) = 0.9$$

$$P(I) = 0.1$$

$$P(A|\bar{I}) = 0.02$$

$$P(A|I) = 0.97$$

El diagrama de árbol sería:



La probabilidad que nos piden se calcula:

$$P(\bar{I}|A) = \frac{P(A|\bar{I})P(\bar{I})}{P(A)} = \frac{P(A|\bar{I})P(\bar{I})}{P(A|\bar{I})P(\bar{I}) + P(A|I)P(I)} = \frac{0.02 * 0.9}{0.02 * 0.9 + 0.97 * 0.1} = 0.157$$

La probabilidad de que no haya habido accidente aunque funcione la alarma es aproximadamente del 16%.

EJEMPLO DEL CÁNCER

¿Cuál es la probabilidad de que una mujer tenga cáncer si tiene un resultado positivo en la mamografía?

- El 1% de las mujeres mayores de 50 años tiene cáncer de mama.
- El 92% de las mujeres que tienen cáncer de mama dan positivo en las mamografías.
- El 8% de las mujeres serán falsos positivos.

¿Qué es lo que queremos averiguar?

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

- E: el resultado de la mamografía es positivo a cáncer.
- H: la mujer tiene cáncer.

Los datos que tenemos:

$$P(H) = 1\% = 0.01$$

$$P(H^C) = 99\% = 0.99$$

$$P(E|H) = 92\% = 0.92$$

$$P(E|H^C) = 8\% = 0.08$$

La probabilidad que nos piden:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^C)P(H^C)} = \frac{0.92 * 0.01}{0.92 * 0.01 + 0.08 * 0.99} = \frac{0.0092}{0.0884} = 0.1041$$

La probabilidad de que una mujer que da positivo en el test, tenga cáncer es aproximadamente del 10%.

PRUEBAS A/B FRECUENTISTAS

El término prueba o test A/B se utiliza en el ámbito del Marketing Digital y la Analítica web para describir experimentos aleatorios con dos variantes, A y B, siendo una la de control y la otra la variante. En el campo del diseño de páginas web (especialmente, cuando se diseñan experiencias de usuario), el

objetivo es identificar los cambios que incrementan o maximizan un resultado determinado (por ejemplo, la proporción de clics que recibe un anuncio publicitario).

Como el propio término indica, se comparan dos versiones (A y B), que son idénticas salvo por una variación que puede afectar al comportamiento del usuario. La versión A puede ser la que se esté utilizando en un momento determinado (control), mientras que la versión B se modifica en algún aspecto concreto (variante). Por ejemplo, en una página web de comercio electrónico, el proceso de compra es normalmente un buen candidato para realizar un test A/B, dado que, incluso mejoras marginales en la tasa de abandono, pueden implicar incrementos significativos en las ventas. Igualmente, se pueden observar mejoras cuando se modifican elementos como el texto, la disposición de elementos (layout), imágenes o los colores.

Por ejemplo, una empresa con una base de datos de 2.000 clientes crea una campaña de correo electrónico con un código de descuento. Su objetivo es generar ventas a través de su sitio web. Se crea dos versiones del mismo correo electrónico pero con diferentes intenciones. Una anima a realizar una compra, y la otra contiene el código promocional.

- A 1.000 personas se les envió un correo electrónico diciendo: "¡La oferta finaliza este sábado! Use el código A1"
- A otras 1.000 personas se les envió un correo electrónico diciendo: "¡La oferta termina pronto! Use el código B1".

Todos los demás elementos del correo incluido el diseño, eran idénticos. La campaña finalizó con una alta tasa de éxito mediante el análisis de la utilización de los códigos promocionales. El correo electrónico con el código A1 obtuvo un 5% de tasa de respuesta (50 de las 1.000 personas a las que se le envió el correo electrónico), y el correo electrónico con el código B1 obtuvo una tasa de respuesta del 3% (30 de las 1.000 personas a las que se le envió el correo electrónico). La empresa determinó que, en este caso, la primera llamada a la acción era más eficaz y la utilizaría en las futuras ventas.

El propósito de la prueba era determinar cuál es el camino más eficaz para alentar a los clientes a realizar una compra. Sin embargo, si el objetivo de la prueba era para ver cuál de los correos electrónicos que generaba el mayor CTR (es decir, el número de personas que realmente hacen clic en el sitio web después de recibir el correo electrónico) los resultados podrían haber sido diferentes. Muchos de los clientes que recibieron el código B1, a los cuales no se les enseñaba la fecha final de la promoción para acceder al sitio web, no sintieron la necesidad de hacer una compra inmediata. Sin embargo, si el propósito de la prueba hubiera sido ver qué correo electrónico traería más tráfico al sitio web, el correo electrónico con el código B1 bien podría haber tenido más éxito. Una prueba de A/B debe tener un resultado definido que sea medible como el número de ventas realizadas, porcentaje de clics, o el número de personas que se registra, etc.

Uno de los ejemplos más prácticos de estos es el algoritmo de Facebook de anuncios publicitarios.

Aunque son fáciles de entender, las pruebas A / B en las que el ganador se lo lleva todo pueden resultar en una pérdida de dinero para los anunciantes. Hay 3 desafíos clave con las pruebas A / B clásicas frecuentistas:

1. Las preferencias de los consumidores evolucionan con el tiempo. Si un solo anuncio se declara ganador, eso implica que siempre debe mostrarse. Este no es siempre el caso: la prueba A / B puede haberse realizado durante una temporada específica o en días específicos (por ejemplo: una aplicación de entrega de alimentos puede tener una ensalada como anuncio ganador en verano y el chocolate caliente como ganador en invierno). Un "ganador" puede no ser siempre un ganador, y un "perdedor" puede no ser siempre un perdedor. Puede haber falsos negativos y positivos.
2. Todos los anuncios son (a veces) efectivos. La vida real es probabilística: un anuncio ganador no es el mejor para el 100% de los usuarios / impresiones (como a veces implica un paradigma de pruebas A / B). El hecho de que un anuncio esté "ganando" solo significa que tiene un mejor rendimiento la mayor parte del tiempo. Puede haber un 10% de las impresiones / audiencias para las que el anuncio "perdedor" es mejor. Algunas impresiones en anuncios "malos" generan compras y algunas impresiones en anuncios "buenos" no generan compras.
3. La asignación equitativa de impresiones da como resultado un gasto inútil. Si un anuncio es mejor el 90% del tiempo y otro es mejor el 10% del tiempo, entonces distribuir el 50-50 entre ellos puede resultar en una pérdida de inversión, la empresa pierde dinero. Puede resultar costoso ejecutar el anuncio ganador con más frecuencia de la que se debería.

EL PARADIGMA DE LOS BANDIDOS BAYESIANOS (BAYESIAN BANDITS): EL CASO DE FACEBOOK

El algoritmo de Facebook utiliza un enfoque bayesiano probabilístico para abordar los problemas anteriores. Si bien Bayesian Bandits es un nombre genial (para un algoritmo), ¿por qué es mejor que los algoritmos menos complejos? En un paradigma bayesiano, se utiliza información que ya se conoce (a priori) para hacer predicciones sobre algo que se desea saber. El término "bandidos" proviene de una clase de problemas de probabilidad que tratan con variables que tienen "muchos brazos", como una fila de máquinas tragamonedas en el piso de un casino.

Los "bandidos" parecen idénticos pero tienen diferentes comportamientos. En el paradigma de los Bandidos Bayesianos, nuestro jugador sabe con qué frecuencia las máquinas tragamonedas resultan en "victoria" y se enfrenta al problema de tomar decisiones sobre en qué máquinas jugar en el futuro.

¿Cómo se aplica todo esto a los anuncios de Facebook?

Al igual que un jugador se enfrenta al problema de qué máquina tragamonedas jugar para maximizar las ganancias, el "juego" de Facebook es seleccionar y priorizar los anuncios que puede mostrar a un usuario, para maximizar los ingresos para sí mismo y para el anunciante.

Facebook podría usar las pruebas A / B y mostrar diferentes anuncios a un número igual de usuarios antes de cerrar los anuncios con peor rendimiento. Dichas pruebas A / B en las que el ganador se lo lleva todo tiene las desventajas de tiempo, efectividad marginal y gasto inicial desperdiciado, como se mencionó anteriormente. Por eso Facebook utiliza el paradigma Bayesian Bandits en lugar de las pruebas A / B clásicas frecuentistas para abordar estos problemas.

Más importante aún, el enfoque bayesiano no es exclusivo de los anuncios de Facebook. Los enfoques bayesianos han experimentado un crecimiento masivo en las últimas dos décadas. El crecimiento de Internet ha creado una medición ubicua del comportamiento y las interacciones de los usuarios. El enfoque bayesiano aprovecha de forma única este enorme almacén de datos. Lo que una vez fue una técnica estadística desatendida se volvió más poderosa cuando las enormes cantidades de información que requiere el algoritmo estuvieron disponibles, y durante la última década, los enfoques bayesianos han sido catapultados de un campo matemático recóndito a un área que puede aprovechar directamente la ubicuidad de los datos que el Internet ha habilitado.

Los periódicos que buscan decidir entre diferentes titulares, los minoristas que buscan decidir entre diferentes embalajes, las farmacéuticas que buscan decidir entre diferentes tratamientos, las aerolíneas que buscan decidir entre diferentes precios y, por supuesto, las plataformas publicitarias que buscan decidir entre diferentes anuncios, todos usan algo del enfoque bayesiano.

INFERENCIA ESTADÍSTICA

Vamos a empezar a hablar de Inferencia Estadística. Seguramente muchos de ustedes probablemente ya hayan escuchado hablar o hayan estudiado los temas de Inferencia que son Estimación, Intervalos de confianza y Contrastes de Hipótesis. Pero incluso si lo han hecho, vamos a repasarlos desde cero. Vamos a construir todo desde un lugar muy intuitivo.

Bien, anteriormente aprendimos cómo hacer una estimación de máxima verosimilitud, uno de los parámetros que aprendimos a calcular fue la estimación de máxima verosimilitud de la media. Recordemos que esto se reduce a la media muestral, tanto en el caso de la Bernoulli como en el de la Normal. Bien, supongamos que hemos recopilado un pequeño conjunto de datos que contiene los

números 1, 3 y 4. Ahora, supongamos que queremos la media muestral de este conjunto de datos. Muy claramente la respuesta es 3. Esto tiene sentido porque 3 está en el medio. Bien, pero ahora consideremos un conjunto de datos diferente, digamos que esta vez nuestro conjunto de datos contiene los números 2.9, 3.0 y 3.1. En este caso, cuando calculamos la media nuevamente, también obtenemos 3. Así que ahora consideremos la pregunta es, ¿en cuál caso tenemos más confianza? ¿Estamos más seguros de que el verdadero valor de la media es 3 en el caso del primer dataset o en el caso del segundo? Bien, creo que la mayoría de nosotros estaríamos de acuerdo en que tenemos más confianza en el segundo caso porque los números están más juntos, cabe menos duda.

¿Por qué se tiene más confianza cuando los números están más juntos? Veamos un ejemplo más extremo de esto. Consideremos que tenemos una muestra con solo dos valores pero muy separados, el 0 y 10 millones. La media sería 5 millones. Pero probablemente no tengamos mucha confianza en esa estimación en comparación con la media de 3 cuando nuestras muestras estaban bastante más unidas. En este caso, la media podría ser de cinco millones y uno, o cinco millones y dos, y no tendríamos suficiente evidencia para decir una cosa o la otra.

Ahora, consideremos otro escenario diferente. Supongamos que hemos recopilado dos conjuntos de datos sobre el lanzamiento de una moneda. Y lo que queremos saber es la probabilidad de que salga cara, calculando la media. Cuando calculo la media en mi primer conjunto de datos, tengo 10 muestras. Eso significa que lancé la moneda 10 veces. La media es 0.6.

En mi segundo conjunto de datos, tengo 1000 muestras. Nuevamente, calculo la media y la media ahora es 0.65.

- Dataset 1: 10 lanzamientos: $\bar{x} = 0.6$
- Dataset 2: 1000 lanzamientos: $\bar{x} = 0.65$

¿En cuál de estas estimaciones tenemos más confianza? ¿Tenemos más confianza cuando hemos recopilado solo 10 muestras, o cuando hemos recopilado 1000 muestras?

Bien, espero que todos lleguemos a la misma respuesta, que es que tenemos más confianza cuando recopilamos más muestras porque de manera intuitiva esto nos proporciona más seguridad, menos posibilidad de error. Esto es muy importante cuando hacemos un experimento médico, por ejemplo, para probar un fármaco, necesitamos tener un cierto umbral de participantes antes de estar seguros de que funciona. Por ejemplo en el caso de una vacuna como la de la COVID19 para estudiar si es efectiva se requiere un número bastante grande de personas en el estudio. No se puede simplemente probar el medicamento o la vacuna en una persona y decir, ajá, en esta persona ha funcionado, por lo tanto, mi medicamento funciona bien. Porque pueden haber otros factores o efectos aleatorios por detrás, y lo que sucede es que esperamos que estos efectos aleatorios se promedien cuando probamos el medicamento en un número mucho mayor de personas. Por eso tenemos que recolectar más muestras.

Bien, entonces en resumen lo que hemos encontrado hasta ahora, es que sabemos que hay dos cosas que afectan nuestra confianza en una estimación. Uno, la dispersión de las muestras. Si están más dispersas tenemos menos confianza en la cantidad de muestras que hemos recopilado. Y también parece importar el número de muestras, nos volvemos más seguros cuando tenemos más muestras.

NIVEL DE CONFIANZA

Vamos a hablar ahora de Intervalos de confianza (IC). No vamos a discutir la derivación detallada o los por qué y cómo. Eso lo haremos luego. De momento vamos a hablar sobre el intervalo de confianza Z, que es la conocido IC para la media de una Normal. Lo que tenemos es lo siguiente.

Supongamos que hemos recolectado algunas muestras de una variable que me interesa. Llamemos a eso X, entonces tengo mis n elementos muestrales: $\{x_1, \dots, x_n\}$. La media muestral, la cuasi varianza y la cuasi desviación típica serían:

- Media muestral: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Cuasi varianza muestral: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Cuasi desviación típica: $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

De esta manera, recordemos que la cuasi varianza se diferencia de la varianza en que la cuasi varianza está dividida entre n-1 y la varianza entre n, pero escogemos la cuasi varianza porque es un estimador insesgado y es mejor de cara a esa confianza que tendremos en esta estimación. Finalmente, el intervalo de confianza se puede definir así específicamente, con esta fórmula, para un 95% de nivel de confianza:

$$IC_{95\%} = \left[\hat{\mu} - 1.96\frac{\hat{\sigma}}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}\right]$$

Como es un intervalo, tiene un extremo superior y uno inferior. Los intervalos de confianza los podemos calcular para otros porcentajes de confianza distintos del 95%, lo cual causaría un efecto en el 1.96 de la ecuación, sería otro número distinto, por lo tanto si cambiamos el nivel de confianza esto haría que cambiara también la longitud del intervalo.

¿Qué significado tiene el nivel de confianza? Significa que esa proporción de intervalos contendrá al parámetro verdadero que estemos estimando y para el que estemos calculando ese intervalo, en nuestro ejemplo, la media.

Ahora recordemos cuáles son las dos cosas que vimos antes que afectaban a la confianza y por tanto al IC. Primero, tenemos la varianza en los datos. Si los datos están más dispersos o, en otras palabras, tienen una mayor varianza, entonces nuestra confianza es menor. En segundo lugar, tenemos la cantidad de muestras que hemos recolectado cuando recolectamos más muestras nuestra confianza es mayor.

Puedes ver que ambas intuiciones se reflejan en la fórmula del el intervalo de confianza. Podemos ver que el ancho del intervalo es proporcional a sigma. Es decir, si aumenta la varianza, también lo hace el ancho del intervalo de confianza porque el intervalo es más amplio. Y si es un intervalo más amplio esto hace que en realidad tengamos menos precisión en nuestra estimación. En segundo lugar, podemos ver que el ancho del intervalo es inversamente proporcional a la raíz cuadrada del número de muestras n, es decir, cuando el número de muestras aumenta, el ancho del intervalo de confianza se vuelve más pequeño. Debido a que el ancho del intervalo se vuelve más pequeño, tenemos más precisión en nuestra estimación. Así que una mayor varianza conduce a menos confianza, mientras que más muestras conduce a más confianza o precisión.

DISTRIBUCIÓN DE UN ESTIMADOR

Ahora vamos a cambiar a una perspectiva más matemática sobre este tema. Supongamos que sabemos que nuestros datos provienen de una distribución Normal. Digamos que cada punto de datos que hemos recopilado es iid (independiente e idénticamente distribuido). Entonces cada muestra que hemos recolectado proviene de la misma distribución y son independientes entre sí. Por ejemplo, si estoy midiendo alturas, sé que la altura de John es independiente de la altura de Mary. Sabemos que la media muestral es la suma de todas nuestras muestras divididas por n, y n es solo un número, no es aleatorio, pero los xi son aleatorios. Como recordarán, las funciones de variables aleatorias también son variables aleatorias. En este caso, la suma de variables aleatorias también es una variable aleatoria intuitivamente.

Por ejemplo, considere X_1 y X_2, que son el resultado de dos lanzamientos de moneda. Entonces sea $Y = X_1 + X_2$. Como X_1 y X_2 toman valores 0 o 1, esto implica que la Y que es la suma de las dos puede tomar valores, 0, 1 o 2, según las posibles combinaciones de salida. Y, por supuesto, como X_1 y X_2 son variables aleatorias, también lo es Y. Eso significa que puedo calcular la probabilidad de que Y sea cero, Y sea uno e Y sea dos. Es decir, Y tiene una distribución, y decir que tiene una distribución es lo mismo que decir que es aleatorio. Entonces las sumas de variables aleatorias son aleatorias y, en general, las funciones de variables aleatorias son aleatorias. Ahora bien, ¿qué sabemos acerca de las sumas de variables aleatorias distribuidas normalmente? Bueno, resulta que también se distribuyen normalmente.

Si tengo dos variables aleatorias independientes $X_1 \sim N(\mu_1, \sigma_1^2)$ y $X_2 \sim N(\mu_2, \sigma_2^2)$, entonces ¿cuál es la distribución de la suma?

$$Y = X_1 + X_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2)$$

Lo siguiente que podemos hacer es extender esta idea a la suma de n variables aleatorias normales iid, es decir independientes y con exactamente la misma distribución Normal:

$$X_1 + \cdots + X_n \sim N(n\mu,\ n\sigma^2)$$

¿Para qué vamos a usar esto?

Sabemos que el estimador media muestral de nuestros datos es una función de variables aleatorias iid:

$$\hat{\mu} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Entonces si nos preguntamos cuál es la distribución de $\hat{\mu}$, dado que es la suma de n variables iids Normales dividido por una constante n, solo tenemos que aplicar lo anterior lo que nos da que tendría esta distribución:

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Vamos a demostrarlo de forma correcta, usando el valor esperado y la definición de varianza basada en él, aplicando las propiedades del valor esperado:

$$E(\hat{\mu}) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2$$

Bien, entonces juntemos todo esto, acabamos de encontrar la distribución que tiene nuestro estimador $\hat{\mu}$. Consideremos ahora por qué esto tiene sentido. Ya hemos analizado la media, que creo que es bastante intuitiva. Que el valor esperado del estimador sea la media que es lo que está estimando es algo bueno, es lo que hemos llamado propiedad de insesgadez. Esto tiene sentido porque eso es lo que estamos tratando de estimar. Su valor esperado debería ser μ.

Pero ¿qué pasa con la varianza? Como puedes ver, depende de dos cosas. Primero, depende de sigma cuadrado, que es la varianza de la variable original, la varianza de nuestros datos originales. En segundo lugar, depende de n, la cantidad de muestras que hemos recolectado. Entonces, ¿por qué eso tiene sentido? Bueno, como puedes ver, una de estas cosas hace que aumente, mientras que la otra la reduce. Cuando σ^2, la varianza de los datos originales, es más grande, la varianza de $\hat{\mu}$ se vuelve más grande. Eso tiene sentido. Si X está más dispersa entonces $\hat{\mu}$ también estará más dispersa. Es decir, μ es más difícil de

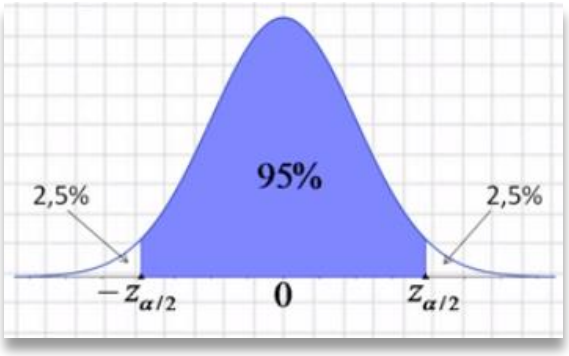
estimar si las muestras de X varían mucho. También podemos ver que a medida que recolectamos más y más muestras, es decir, a medida que la muestra se hace más y más grande (n crece), la varianza de $\hat{\mu}$ se hace más pequeña. Eso también tiene sentido ya que a medida que recopilamos más y más datos, la varianza de nuestra estimación disminuye.

Por ejemplo, supongamos que estamos lanzando una moneda para intentar medir la probabilidad de que salga cara. En las primeras etapas del experimento, la estimación puede variar mucho porque no tenemos tantas muestras. Pero una vez que hayamos recolectado cien o mil muestras, la estimación estará muy cerca de 0.5. Y los nuevos lanzamientos de monedas no afectarán tanto a la estimación.

Bien, entonces más muestras significa que la varianza de nuestra estimación disminuye. De acuerdo, esta es una idea muy intuitiva detrás del intervalo de confianza, lo que estamos tratando de hacer es crear una distribución de la media muestral, u otro parámetro en otro caso. Entonces el intervalo de confianza es solo un intervalo sobre esta distribución. Y cuando la varianza de nuestros datos originales es mayor, esto hace que nuestro intervalo de confianza sea mayor, o sea, menos preciso. Sin embargo, podemos reducir nuestro intervalo de confianza y aumentar esa precisión, al recopilar más datos. Es decir, podemos tener más confianza en nuestra estimación si recopilamos más muestras. Bien, entonces más varianza significa intervalos más grandes o menos precisos, es lo mismo, y más muestras significa intervalos más pequeños y más precisos.

RAZONANDO LA FÓRMULA

Anteriormente, aprendimos que cuando nuestro estimador media muestral tiene una distribución Normal con media mu y varianza sigma al cuadrado sobre n, entonces el intervalo de confianza se puede definir de la siguiente manera. Para el intervalo de confianza del 95 por ciento, simplemente seria el intervalo que cubre el 95 por ciento medio del área de esta distribución.



Vamos a descubrir cómo podemos determinar los valores numéricos reales para este intervalo. Es decir, ¿qué cálculos tenemos que hacer para determinar el extremo inferior y superior de este intervalo?

Además, en general, no es necesario utilizar el 95 por ciento como nivel de confianza. Otros valores comunes son el 90 por ciento o el 99 por ciento. En general, si llamamos γ al nivel de confianza el intervalo de confianza γ es el intervalo que cubre la porción γ del área debajo de la PDF (función de densidad).

Entonces, tenemos una distribución que es Normal, y queremos encontrar el extremo inferior y superior para que el área en el medio sea el 95 por ciento. Recordemos que esto se puede hacer usando la CDF, la función de distribución acumulativa. Supongamos que tenemos una variable aleatoria Z que tiene una distribución Normal estándar, es decir, media cero y varianza uno. Entonces, debido a que esta distribución es simétrica, si queremos cubrir el noventa y cinco por ciento en el medio, tendremos 2.5% a la izquierda y en el lado derecho 2.5% también. Consideremos el lado izquierdo.

¿Qué valor de Z necesitamos para que el área a la izquierda de Z sea 2.5 por ciento? A esto lo llamaremos Valor Z izquierdo. Esta es solo la CDF inversa de Z izquierda, el área desde menos infinito hasta Z izquierda queremos que sea 0.025 (2.5%). Podemos calcular los valores usando una tabla de probabilidades de la Normal, o con Python.

Después de hacerlo, deberíamos obtener la respuesta $-z_{\alpha/2} = -z_{0.025} = -1.96$.

Ahora, debido a la simetría alrededor de cero (Z es Normal estándar con media cero), podemos concluir que la respuesta para la parte derecha es ese mismo valor pero positivo $z_{\alpha/2} = z_{0.025} = 1.96$.

Bien, entonces, esto es para un caso sencillo que es la Normal estándar, pero ¿qué debemos hacer cuando los datos no son Normal estándar? Si tenemos una variable aleatoria X que es normal pero no con media cero y varianza 1. Pues lo usual es estandarizar la variable X, es decir, transformarla para llevarla a una distribución Normal estándar. Y eso se hace restando la media y dividiendo por la desviación estándar. A esto se le llama estandarización. Es decir, si tienes alguna variable aleatoria Normal con media mu y varianza sigma cuadrado, puedes transformarla en una variable aleatoria estandarizada restando mu y dividiendo por sigma. A menudo usamos la letra Z para denotar esta variable aleatoria estandarizada.

Es decir, si: $X \sim N(\mu, \sigma^2)$:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Entonces, ¿qué sabemos sobre Z?

Bueno, sabemos que si queremos los extremos inferior y superior del intervalo de confianza del 95 por ciento, estos son los que acabamos de calcular, -1.96 y 1.96. Así que ahora tenemos un intervalo de confianza para Z, y podemos expresarlo de la siguiente manera $-1.96 \leq Z \leq 1.96$, donde decimos que Z está entre esos dos valores. Pero por supuesto, no queremos saber sobre Z. Queremos saber sobre nuestro parámetro mu. Llamémosle Z a la estandarización de $\hat{\mu}$ considerando cuál es su media y su desviación típica:

$$Z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Entonces

$$-1.96 \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \leq 1.96$$

$$-1.96\frac{\sigma}{\sqrt{n}} \leq \hat{\mu} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}$$

$$-1.96\frac{\sigma}{\sqrt{n}} - \hat{\mu} \leq -\mu \leq 1.96\frac{\sigma}{\sqrt{n}} - \hat{\mu}$$

$$\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}$$

Transformando a cada lado obtenemos un intervalo de confianza para μ. Ya logramos el objetivo de aislar el parámetro que me interesa que es μ, y obtenemos estos dos extremos superior e inferior que serian un intervalo de confianza para μ al 95%.

$$\mathrm{IC}_{95\%} = \left[\hat{\mu} - 1.96\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

De manera más general, si quisiéramos el intervalo de confianza con un nivel de confianza $\gamma = 1 - \alpha$, sería:

$$\mathrm{IC}_{1-\alpha\%} = \left[\hat{\mu} - \mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + \mathrm{z}_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

A esto lo llamaremos el intervalo de confianza Z, ya que se basa en la Normal estándar.

Pero es importante destacar aquí que esto no se interpreta como la probabilidad de que μ esté entre estos valores. En cambio, recordemos que estamos usando la interpretación frecuentista, entonces lo que esto significa es que no se trata de una probabilidad, sino de una frecuencia relativa. En otras palabras, significa que si hago algún experimento muchas veces, el 95 por ciento de esas veces la μ verdadera estará contenida en este intervalo de confianza:

$$0.95 \approx \frac{\#\ de\ experimentos\ donde\ \mu\ está\ contenida\ en\ el\ IC_{95\%}}{\#\ total\ de\ experimentos}$$

Ahora, hay un pequeño problema con la derivación del IC, si miramos el intervalo de confianza con cuidado, debemos reconocer que hay un valor que no conocemos. Este valor es σ. Recordemos que en la práctica solo tenemos los datos, no conocemos los valores reales de los parámetros poblaciones de su distribución, solo conocemos estimaciones muestrales de esos parámetros. Afortunadamente, esto se puede aproximar simplemente usando la desviación típica muestral. Esto no es preciso, pero es una aproximación muy común para usar en la práctica. Además, si tenemos muchos datos, o en otras palabras, nuestra muestra es suficientemente grande esto se vuelve más preciso. Normalmente, se utiliza la estimación insesgada de la varianza, la cuasi varianza, que es dividido por n menos uno.

$$\mathrm{IC}_{1-\alpha\%} = \left[\hat{\mu} - \mathrm{t}_{n-1,\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}},\ \hat{\mu} + \mathrm{t}_{n-1,\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}}\right]$$

Sin embargo si no conocemos el valor real de sigma y tenemos que estimarlo, esto ya no va a tener una distribución Normal sino una distribución que se parece un poco, llamada distribución t-Student, que es también parecida a una campana pero más dispersa, menos concentrada. Se le llama de cola pesada (heavy tailed) porque en la cola que son los extremos son más pesadas o gordas que la de la Normal. Y por último, es muy común emplear el teorema central del límite para los intervalos de confianza. Es decir, al principio hemos asumido que los datos originales se distribuían normalmente, pero esto no hace falta si tenemos una muestra grande, porque por el teorema central del limite la distribución de la media muestral, $\hat{\mu}$ seguirá siendo Normal porque es una suma de un numero grande de variables iid, que converge a una distribución Normal sin importar cuál sea la distribución de los datos. Por lo tanto, es frecuente que las personas utilicen el intervalo de confianza Z incluso cuando los datos no se distribuyen normalmente.

En resumen:

- Si tenemos datos Normales y conocemos la varianza:

$$IC_{1-\alpha\%} = \left[\hat{\mu} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

- *Si tenemos datos Normales y no conocemos la varianza:*

$$IC_{1-\alpha\%} = \left[\hat{\mu} - t_{n-1,\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}},\ \hat{\mu} + t_{n-1,\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}}\right]$$

- *Si además no tenemos datos Normales pero la muestral es grande (>30):*

$$IC_{1-\alpha\%} = \left[\hat{\mu} - z_{\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}},\ \hat{\mu} + z_{\alpha/2}\frac{\hat{\sigma}_1}{\sqrt{n}}\right]$$

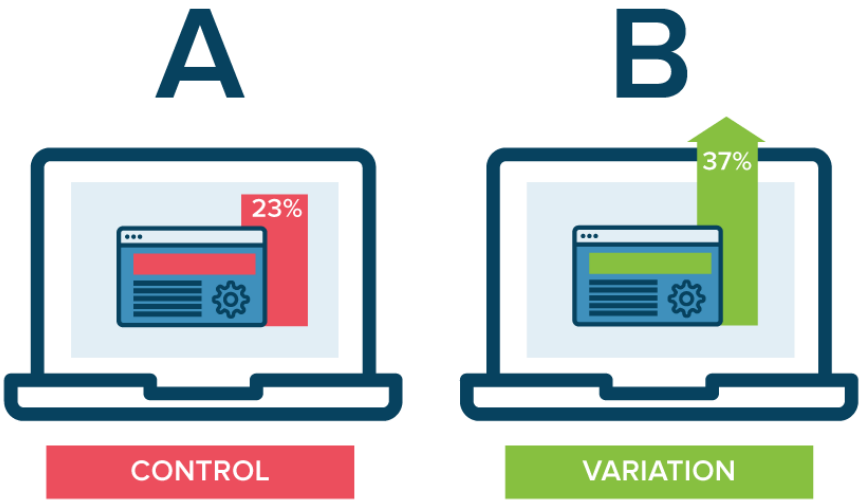
CONTRASTES DE HIPÓTESIS

Vamos a ver ahora otro concepto esencial de inferencia estadística, del enfoque frecuentista, que son las pruebas o contrastes de hipótesis. Las pruebas de hipótesis son más fáciles de entender cuando se consideran ejemplos prácticos del mundo real. Así que vamos a presentar algunos.

El primer ejemplo es el escenario clásico en el que somos unos fabricantes de medicamentos y queremos probar la eficacia de un medicamento.

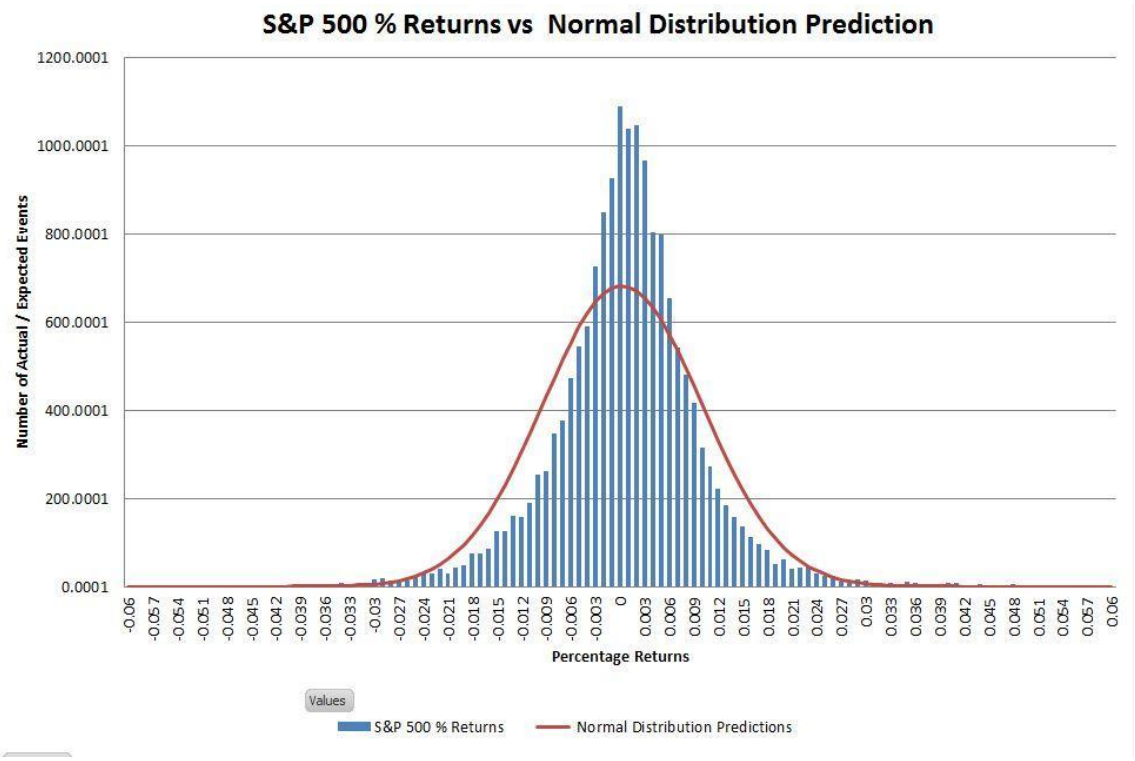
Queremos saber si este medicamento tiene alguna mejora con respecto al tratamiento estándar para alguna afección. Supongamos que estamos fabricando un medicamento para eliminar algunas cosas malas del cuerpo del paciente, sea lo que sea. El tratamiento estándar toma varios días para eliminar todas estas cosas malas. El nuevo fármaco que estamos desarrollando también tarda varios días en eliminar todas estas cosas malas. En tal experimento, dividiríamos a los participantes de nuestro estudio en dos grupos, el grupo de control y el grupo de tratamiento. El grupo de control es el grupo que recibe el tratamiento estándar actual o un placebo o ningún tratamiento. El grupo de tratamiento es el grupo que recibe el nuevo medicamento (que bien podría ser una pastilla, una vacuna, o un tratamiento de otro tipo). Una pregunta que podríamos querer considerar aquí es la siguiente, ¿el nuevo medicamento funciona más rápido? Es decir, ¿el nuevo fármaco tarda, en promedio, menos tiempo en actuar? ¿Y esta diferencia es estadísticamente significativa? Discutiremos lo que queremos decir con *estadísticamente significativo* más adelante. Si el medicamento nuevo tarda en promedio menos tiempo en actuar con una diferencia significativa (importante) esto es una buena noticia para nosotros porque este medicamento es mejor.

Vamos a ver otro ejemplo, supongamos que tenemos dos diseños de páginas web que queremos estudiar.



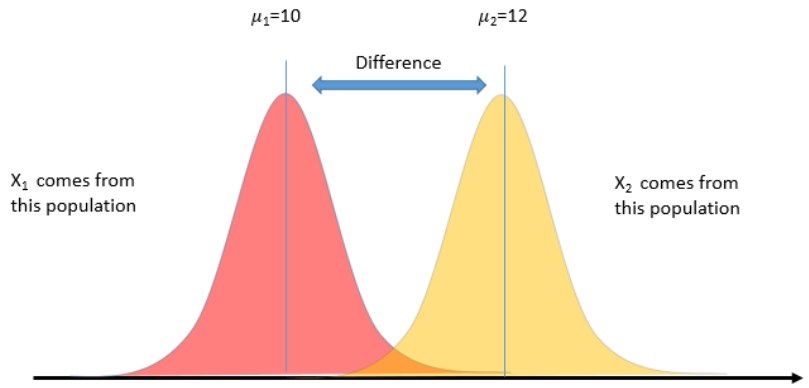
Queremos saber qué página web lleva al mayor tiempo de permanencia en esa página, por parte de los usuarios. En términos generales, queremos que los usuarios pasen más tiempo en nuestra página, porque eso significa que es más atractiva. En la práctica, lo que vamos a hacer es entregar cada página web a un grupo de usuarios diferentes y luego usar esos datos para determinar si el tiempo dedicado a cualquiera de las páginas es diferente o no. Por lo tanto, nuevamente, tenemos el mismo escenario en el que nos gustaría tomar medidas de dos grupos diferentes y queremos saber si la diferencia es estadísticamente significativa.

Otro ejemplo, supongamos que hemos estado observando algunas acciones financieras y queremos calcular los rendimientos diarios de esas acciones, si trazamos su distribución, se ven como curvas de campana centradas alrededor de cero.



Por supuesto, los rendimientos de las acciones no estarán exactamente centrados en cero. Y de hecho, esperamos que sean mayores que cero para que podamos ganar dinero. No basta con calcular el rendimiento medio diario y comprobar si es mayor que cero o no. A veces, el rendimiento diario individual será mayor que cero, pero a veces el rendimiento diario individual puede ser negativo y, de hecho, muy negativo. En general, quizás el rendimiento diario promedio sea mayor que cero, pero ser mayor que cero no es suficiente. Lo que queremos saber si el efecto es estadísticamente significativo.

Estos ejemplos nos dan algo de intuición sobre de qué se tratan las pruebas o contrastes de hipótesis y cómo se puede usar en el mundo real. Es importante reconocer que los ejemplos anteriores que acabamos de ver no son todos iguales. Cuando comparamos un medicamento nuevo y el tratamiento estándar o comparamos dos páginas web diferentes, tendremos dos grupos de datos. Tenemos un grupo correspondiente al fármaco nuevo y un grupo correspondiente al tratamiento estándar o al placebo. Estos contrastes serían de dos poblaciones o dos muestras.



Por otro lado, cuando comparamos el rendimiento diario de algunas acciones con un valor fijo como cero, solo tenemos un grupo de datos que provienen todos de la misma acción. En este caso, lo llamamos contraste de 1 población o de una muestra. Solo hay un grupo para comparar con un número fijo.

Hay otra distinción que podemos hacer pero para discutir esto, necesitamos introducir el concepto de hipótesis nula e hipótesis alternativa.

- Hipótesis nula H_0
- Hipótesis alternativa H_1

Supongamos que para probar nuestro nuevo fármaco, nuestra hipótesis nula es que el efecto del nuevo fármaco es el mismo que el efecto del tratamiento estándar. La hipótesis alternativa es que no son lo mismo. Las hipótesis quedarían:

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$

Usamos la media porque, por supuesto, cada medida será diferente. Lo que nos preocupa es el efecto promedio general de nuestro medicamento. Esto indica que bajo la hipótesis alternativa, el valor medio del grupo uno no es igual al valor medio del grupo 2. Nuestro objetivo cuando estamos haciendo pruebas de hipótesis es detectar si cada una de ellas podría ser cierta o no. Bien, entonces, ¿por qué estamos hablando de esto y cuál es el propósito de esta notación? En este caso, lo que acabo de describir se llama **prueba bilateral**, porque hay dos formas en las que el medicamento puede ser diferente del control. El efecto del fármaco puede ser menor o el efecto del fármaco puede ser mayor. Esos son dos escenarios diferentes bajo la misma hipótesis. Cuando es solamente una opción, lo llamamos **prueba unilateral**.

Tengamos en cuenta que esto también pasa con nuestro ejemplo de acciones financieras porque podemos verificar si el rendimiento diario medio no es igual a cero. O podemos verificar si el rendimiento diario medio es mayor que cero, específicamente, ya que ser menor que cero no es necesariamente algo que nos importe. De nuevo, se trata de una prueba bilateral en el primer caso y unilateral en el segundo caso.

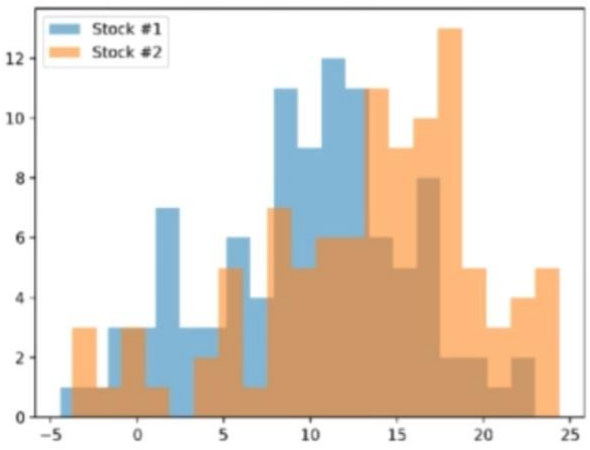
SIGNIFICACIÓN ESTADÍSTICA

¿Qué es y por qué nos preocupa esta idea de significación estadística? ¿Por qué no podemos simplemente elegir el que tenga el mejor valor promedio y luego afirmar que es el mejor? Bueno, pensemos en esta idea, estamos comparando dos grupos de pacientes, el grupo de tratamiento y el grupo de control.

Tomamos algunas medidas para ambos grupos y luego tomamos el promedio. Cuando tomamos el promedio del grupo de tratamiento y el grupo de control, uno de esos promedios será mayor que el otro. Pero ¿eso significa que son significativamente diferentes?

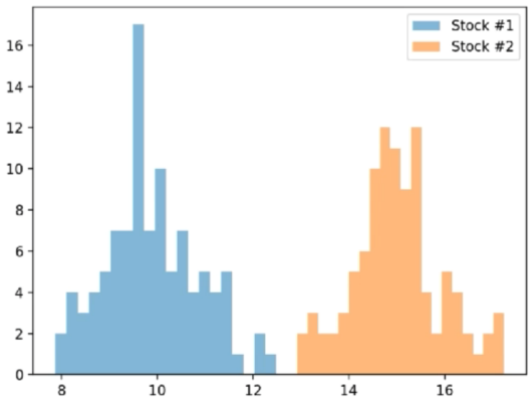
Supongamos que tomo una moneda y la lanzo 100 veces y calculo la tasa de ganancia empírica de sacar cara, luego tomo otra moneda, la lanzo 100 veces y calculo la tasa de ganancia empírica de sacar cara. Claramente, uno de estos tendrá una mayor probabilidad empírica de sacar cara. Digamos que en la moneda 1 salió cara 52 de cada 100 veces, pero en la moneda número dos salió cara 49 de cada 100 veces. ¿Significa esto que la moneda número uno realmente tiene una mayor probabilidad de sacar cara en comparación con la moneda número dos? La respuesta es no. Entonces, ¿cuál es la solución? Quizás esto pueda resultar sorprendente, pero esta idea de la prueba de hipótesis está estrechamente ligada a los intervalos de confianza. Para dar un poco de intuición sobre esto, consideremos un ejemplo más extremo del lanzamiento de la moneda. Lanzo solo una vez la moneda número uno, obtengo cara, pero para la moneda número dos, obtengo cruz. Claramente, este no es un buen experimento. Nadie tiene una mayor probabilidad de que salga cara simplemente porque yo obtuve cara una vez. Ahora consideremos el caso en el que lanzo cada moneda 10 veces. Esto es una mejora, pero intuitivamente todavía no es muy confiable. ¿Qué tal 100 veces, esto parece mejor? ¿Qué tal 1000 veces? Esto parece incluso mejor. Está claro que para el experimento perfecto, aunque no sería práctico, un número infinito de muestras nos daría una respuesta definitiva.

Ahora, consideremos otro escenario, supongamos que estamos comparando acciones y estamos midiendo los rendimientos de cada acción. Con la acción número uno, la rentabilidad diaria promedio es de 10 puntos básicos y la desviación estándar es de 5 puntos básicos. Con la acción número dos , el rendimiento diario promedio es de 15 puntos básicos y la desviación estándar es de 6 puntos básicos.



En este caso, ¿podríamos decir que el stock número dos es mejor que el stock número uno? La respuesta sería no. Hay tanta superposición en sus distribuciones que no está claro cuál es realmente es mejor.

Ahora, consideremos un resultado diferente, supongamos que el rendimiento diario promedio de la acción número uno sigue siendo de 10 puntos básicos, pero la desviación estándar ahora es 1. Suponga que el rendimiento diario promedio de la acción número dos sigue siendo de 15 puntos básicos, pero la desviación estándar también es ahora 1. Con base en esta imagen, ¿estaríamos más seguros al decir que la acción número dos es mejor que la acción número uno?



Intuitivamente, esta imagen debería dar más confianza al hacer tal declaración porque hay menos superposición entre las dos distribuciones.

Así que, ¿qué hemos aprendido en resumen? Hemos aprendido que comparar dos grupos no es simplemente una cuestión de comparar sus promedios, sus promedios siempre serán diferentes. En cada experimento, uno debe tener un promedio más alto que el otro. Así que esto no es útil para comparar. Además, hemos descubierto dos nuevos factores cruciales que parecen ser importantes. Son, número uno, el número de muestras recolectadas y número dos, la varianza en las muestras. Esto es exactamente lo que importaba con los intervalos de confianza. En general, todos estos factores son muy relevantes cuando evaluamos si los resultados de nuestra prueba de hipótesis son o no estadísticamente significativos. Si la diferencia entre el promedio de los dos grupos no es lo suficientemente grande, entonces el resultado no es estadísticamente significativo. Si no hay suficientes muestras, el resultado no es estadísticamente significativo. Si la varianza en los datos es demasiado grande, nuevamente, el resultado no es estadísticamente significativo.

P-VALOR

Ahora vamos a hablar de una noción importante a la hora de trabajar e interpretar los resultados de un contraste de hipótesis. Esta noción se conoce como el p-valor. Es importante tener en cuenta que estamos viendo los conceptos más básicos para explicar o para recordar el tema de contrastes de hipótesis, sin comprender las matemáticas subyacentes de una manera demasiado específica, sino mayormente intuitiva. Pero aún así se recomienda que siempre vayamos un poco más allá de los conceptos básicos para comprender cómo funcionan realmente las pruebas de hipótesis. Probablemente no sea una buena idea simplemente conectar y procesar los datos en una prueba de hipótesis sin pensar en el significado de cara al mundo real detrás del proceso o detrás del contraste que estamos planteando. Dicho esto, ahora tendremos una visión muy útil porque nos va a permitir ver qué hay detrás de este proceso, en términos de entradas o inputs, y cómo interpretar las salidas. Es decir ¿qué entradas tenemos que pasar a una prueba de hipótesis y qué salidas obtengo? ¿Cómo interpreto esos resultados? En primer lugar, recordemos que anteriormente vimos dos tipos de pruebas, hablamos sobre las pruebas unilaterales y bilaterales, así como las pruebas de una o dos muestras.

Un ejemplo sería comprobar la diferencia entre dos medias para ver la diferencia entre dos fármacos. ¿Es el efecto medio del fármaco nuevo mejor que el del tratamiento estándar? Esa sería una prueba unilateral de dos muestras.

Otro ejemplo sería, ¿el rendimiento medio diario de las acciones es igual a cero o no es igual a cero? Esa sería una prueba de una muestra y bilateral.

Entonces esto ya nos define cómo será la entrada o el input que vamos a necesitar en nuestra prueba de hipótesis. Para la prueba de una muestra, los datos serían una matriz unidimensional de números, siendo esos números las muestras del conjunto de datos. Para la prueba de dos muestras, los datos serían dos

unidimensionales. A continuación, pasaremos estos inputs a una función que me hará el contraste de hipótesis y todos los cálculos. Obviamente, es muy importante elegir bien la función según el tipo de prueba que queremos hacer. Básicamente, para los ejemplos que vimos anteriormente, generalmente la prueba Z o la prueba T son buenas elecciones. Tanto la prueba Z como la prueba T funcionan tanto para una muestra como para dos muestras. En Python, la función de ztest de statsmodels se usa para las pruebas de una muestra y para las dos pruebas de dos muestras, por lo que se puede usar la misma función en ambos casos. Esta función no aparece en el paquete scipy, pero aquí sí aparece la prueba T. Las funciones son ttest_1samp para el caso de 1 muestra y ttest_ind para el caso de 2 muestras.

Librería	**1 muestra**	**2 muestras**
Statsmodels	ztest(x)	ztest(x1,x2)
Scipy	ttest_1samp(x)	ttest_ind(x1,x2)

Después de usar estas funciones, obtendremos algunos resultados, esencialmente, esto generalmente consistirá en dos elementos, el estadístico de prueba y el p-valor. Estamos más interesados en el p-valor (p-value) o valor p. Que lo podremos interpretar mucho más fácilmente. Entonces, ¿qué significa este p-valor? Esencialmente, el valor p nos dice si la diferencia que se observó es estadísticamente significativa o no. Por supuesto, el valor p es solo un número. De hecho es una probabilidad, así que estará entre 0 y 1.

Entonces, **¿cómo podemos interpretar el p-valor?**

Básicamente, vamos a pensar en el p-valor como la probabilidad de fiabilidad de la hipótesis nula. Si esa probabilidad es muy pequeñita rechazaremos la hipótesis nula porque será poco fiable, y si no, no la rechazaremos. Entonces, ¿cómo saber cuán pequeño es el p-valor? ¿Cómo decidir si es demasiado pequeño, lo suficiente para rechazar la H0? Aquí es donde entra el nivel de significación alfa. El nivel de significación está estrechamente relacionado con el nivel de confianza, por eso también veremos más adelante que los contrastes de hipótesis están estrechamente relacionados con los intervalos de confianza.

El Nivel de Significación (NS) α, es igual a 1 – el Nivel de Confianza (NC) γ:

$$\alpha = 1 - \gamma$$

Entonces si nosotros fijamos el nivel de significación en su valor usual de 5% (0.05) eso significa que estamos asumiendo un nivel de confianza del 95%. Lo que resta para llegar al 100%. Otros valores usuales para el nivel de significación son el 1% (0.01) que está relacionado con un NC del 99% y el nivel de significación del 10% (0.1) que está relacionado con un NC del 90%.

Si $\alpha = 5\%: \gamma = 95\%$ Si $\alpha = 1\%: \gamma = 99\%$ Si $\alpha = 10\%: \gamma = 90\%$

Por lo tanto, si fijamos el $\alpha = 5\%$, y nos encontramos con un p-valor como resultado del contraste que es menor que ese α, eso significa que el p-valor es demasiado pequeño y rechazaríamos la H0. En este caso se dice que el resultado de nuestra prueba de hipótesis fue estadísticamente significativo. Entonces la salida o el resultado del contraste será un resultado binario, la prueba produjo un resultado estadísticamente significativo o no.

Por ejemplo, imagina que tenemos el siguiente contraste:

- H0: la media (verdadera) de los retornos de una acción es cero.

- H1: la media (verdadera) de los retornos de una acción es distinta de cero.

Observamos nuestros datos (muestrales) y la media muestral resulta ser de 100bp con varianza de 10. Si el resultado del contraste de que la media es cero versus que no lo es, da un p-valor muy pequeño, menor que 0.05, entonces se rechazará la H0. Y tenemos un resultado estadísticamente significativo. Cuando sucede lo contrario no rechazamos H0 porque no hay suficiente evidencia para rechazarla y por tanto no es estadísticamente significativo el resultado.

Veamos ahora otro ejemplo:

- H0: el tiempo promedio de efectividad del nuevo tratamiento B es mayor que el estándar A.

- H1: el tiempo promedio de efectividad del nuevo tratamiento B es menor que el estándar A.

Supongamos que estamos probando un grupo de fármacos, el Grupo A es el grupo de control y el Grupo B es el grupo de tratamiento, seleccionamos un nivel de significación del 5%. Nos interesa saber si el tratamiento nuevo es más rápido que el anterior (media del grupo tratamiento es menor que media del grupo de control). Esto que es lo que nos interesa saber lo ponemos en la Hipótesis alternativa. Y lo contrario sería lo que iría en la H0. Después de llamar a la función de prueba Z o prueba T, encontramos que el valor P es 0.01, y ya que 0.01<0.05, declaramos que el efecto del tratamiento es estadísticamente significativo, y como el p-valor es muy pequeño, menor que 0.05 ($\mathrm{p-valor} < \alpha$) entonces se rechazará la H0. Y podemos decir que el nuevo tratamiento es mejor.

Otro ejemplo, imagina que tenemos los retornos de una acción y queremos comprobar que son mayores que cero:

- H0: la media de los retornos de una acción es menor que cero.

- H1: la media de los retornos de una acción es mayor que cero.

Seleccionamos un nivel de significación del 1%. Obtenemos un p-valor igual a 0.2. Como el p-valor es mayor que 0.01 ($\mathrm{p-valor} > \alpha$) entonces NO se rechazará la H0. No tenemos un resultado estadísticamente significativo y no se rechaza la hipótesis nula. Aunque el promedio muestral sea mayor que cero.

Ahora, podríamos preguntarnos, ¿cómo se hace para elegir el nivel de significación, por qué el valor por defecto o usual es el 5%? En realidad no existe una regla general, sino que probablemente sea mejor consultar algunos artículos en nuestro campo particular según los datos que estemos estudiando (médicos, financieros, etc.) para determinar qué es lo convencional y qué es lo aceptable. También debe aplicarse el sentido común. Si el costo de una falsa alarma es muy alto, por ejemplo imaginemos que producir un medicamento cuesta miles de millones de dólares, entonces es posible que deseemos tener un umbral o α más estricto, más pequeño.

Finalmente, es importante distinguir entre significación estadística y significación práctica, en algunos casos, una diferencia estadísticamente significativa puede no importar de cara al sentido práctico. Por ejemplo, si descubrimos que algún medicamento es más rápido en ser efectivo en promedio en un segundo, pero su fabricación cuesta miles de millones de dólares, ¿es esta una inversión práctica? La respuesta sería no. Por otro lado, supongamos que encontramos que la tasa de clics de un anuncio es del 1%, pero la tasa de clics de otro anuncio es de 1.1%. Y obtenemos que este resultado es estadísticamente significativo. En este caso, este pequeño aumento en la tasa de clics podría generar una diferencia apreciable en los ingresos de nuestra empresa. Entonces aquí sí habría significación práctica. Así que tanto a la hora de plantear los contrastes como a la hora de interpretarlos siempre tenemos que pensar en el sentido práctico del asunto.

INTERPRETACIÓN DEL RESULTADO DEL CONTRASTE

Ahora bien, una cosa que vale la pena mencionar es que para interpretar y explicar la conclusión sobre el resultado del contraste, la terminología que se utiliza es bastante estricta. Sabemos que si encontramos que el p-valor es muy pequeño, por debajo de nuestro nivel de significación, entonces podemos rechazar la hipótesis nula. En este caso, decimos que el resultado es estadísticamente significativo.

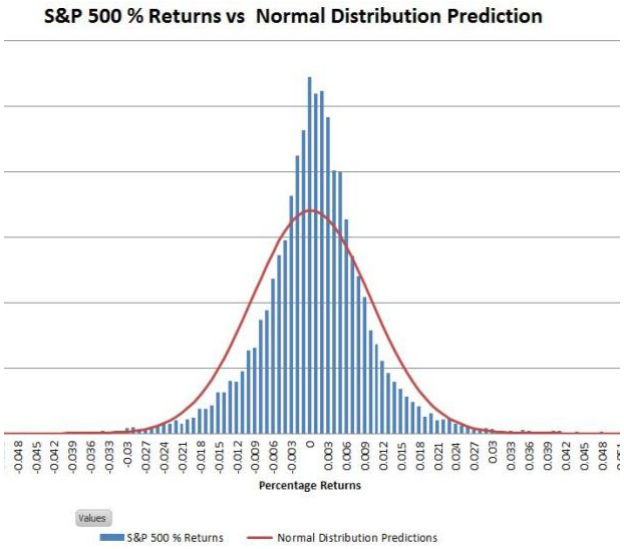
Pero, ¿qué pasa si el valor p no está por debajo de nuestro nivel de significación? Pues nunca diremos que aceptamos la hipótesis nula sino que NO la rechazamos. Es decir las únicas dos opciones son rechazar la hipótesis nula o no rechazar la hipótesis nula. Si obtenemos un p-valor por encima del nivel de significación, no diremos que aceptamos la hipótesis nula. Por ejemplo, supongamos que probamos el tratamiento en cinco personas. Imagina que el resultado es un p-valor grande. Entonces no se puede rechazar la hipótesis nula. El hecho de que no se haya encontrado suficiente evidencia para rechazar la hipótesis nula no implica que la hipótesis nula sea verdadera. Sino que no hay suficiente evidencia para rechazarla, porque quizás con mas muestras el p-valor se reduzca y se pueda rechazar. En términos generales, si quieres ser muy, muy cuidadoso, nunca aceptarás nada, solo rechazarás la hipótesis nula o dejarás de rechazar la hipótesis nula.

OTROS CONTRASTES

Hemos visto anteriormente que este enfoque o estos tipos de contrastes se pueden aplicar a muchos campos, por ejemplo para saber:

- ¿Qué tratamiento tuvo el mejor efecto?
- ¿Qué página web tuvo el mayor engagement?
- ¿Qué anuncio tiene la tasa de clics más alta?

Sin embargo, en realidad hay muchos tipos diferentes de pruebas que no están relacionadas con las que hemos visto hasta ahora, pero que básicamente funcionan más o menos de la misma forma. Por ejemplo, supongamos que tenemos estos datos, una pregunta común es preguntarnos si los datos se distribuyen normalmente.



Un ejemplo de esto son los rendimientos de las acciones cuando trazamos un histograma vemos que se asemeja a la normal porque se ven como una curva de campana. Y mucha gente asumiría que esta es una distribución Normal. Pero, de hecho, hay algunas formas de demostrar que esto es falso. Una forma es utilizar una prueba de hipótesis como por ejemplo la prueba de Kolmogorov Smirnov. Se llaman pruebas

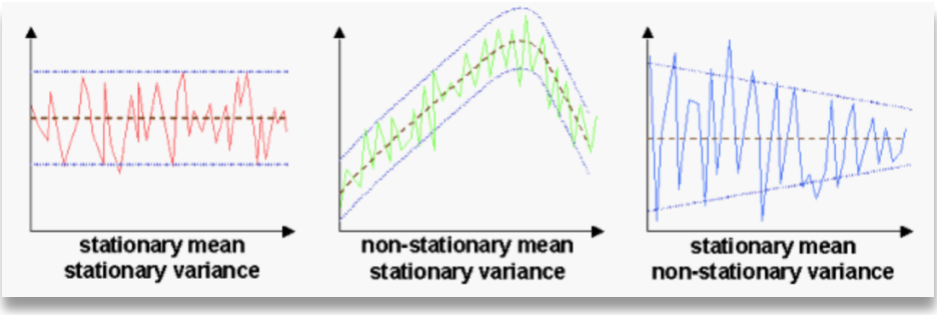
de bondad de ajuste (Goodness-of-fit tests), en general, porque verifican cuan bien se ajusta o no una distribución a la muestra de datos.

En la hipótesis nula asumimos que la distribución es Normal y en el alternativa que no lo es:

- H0: Los datos son Normales
- H1: Los datos NO son Normales

Podemos obtener una salida que será similar a la de los contrastes que hemos visto anteriormente, un valor P. Entonces, si los datos son tan diferentes a una distribución Normal que deberíamos rechazar su normalidad, el valor P resultante será muy pequeño. Pero tengamos en cuenta que la interpretación del valor P requiere que sepamos cuáles son en realidad las hipótesis nula y alternativa. Una vez que sepamos cuáles son, no es necesario que sepamos realmente cómo funciona la prueba.

Ahora vamos a ver otro ejemplo, cuando trabajamos con datos de series de tiempo, un requisito común antes de aplicar los métodos de series de tiempo es que los datos sean estacionarios. Esto significa que la distribución de los datos no cambia con el tiempo. Básicamente, debería verse algo como en el primer plot.



Que no hayan tendencias. O que no cambie la varianza con el tiempo. Aunque la prueba real es bastante complicada, el enfoque es simple. Nuevamente, es solo una cuestión de simplemente conectar los datos, plantear las hipótesis y verificar el valor P. En este caso, la hipótesis nula es que la serie de tiempo NO es estacionaria y la alternativa es que es estacionaria.

- H0: La serie NO es estacionaria
- H1: La serie ES estacionaria

Esto se llama prueba de Dickey Fuller de estacionariedad. Por lo tanto, si encontramos que el valor P es menor que nuestro nivel de significación, declararemos que nuestra serie de tiempo es estacionaria. Porque hemos rechazado la hipótesis nula de que no lo es.

Bien, para resumir, hasta ahora hemos hablado en profundidad sobre cómo comparar dos medias de dos grupos, o una media con un valor fijo, aunque hay otros tipos de contrastes de hipótesis como acabamos de ver. Aunque en la práctica no es necesario estar al tanto de cómo se plantea o funciona la prueba para llegar al resultado, ya que esto lo va a hacer por nosotros la función, pero sí es importante conocer los tipos de pruebas que podemos hacer, cómo se plantean las hipótesis y cómo se interpretan los resultados.

RESUMEN

Vamos a concluir y resumir lo que hemos aprendido en esta sección. En esta sección hemos hablado de inferencia estadística desde el punto de vista clásico frecuentista. El propósito no era aprender sobre esto ya que probablemente ya conocías estos métodos, sino recordarlo bien y tenerlo a mano para que nos

sirva de contraste con respecto al punto de vista bayesiano que veremos en la siguiente sección. Así que espero que al final podamos comprender las similitudes y diferencias entre el enfoque frecuentista y el bayesiano. Sin embargo, hay una buena razón para haber revisado esta sección, incluso si no nos importa particularmente el enfoque frecuentista. Y es que estos dos enfoques no son mutuamente excluyentes. Por ejemplo, podemos tener un sistema en vivo que utilice el enfoque bayesiano, del que pronto aprenderemos, pero aún así puedo hacer algunos análisis fuera de línea después de recopilar los datos en vivo, usando el enfoque frecuentista. En otras palabras, en general, podemos terminar usando ambos enfoques con los mismos datos.

Bien, entonces, en resumen ¿qué aprendimos en esta sección? Bueno pues en general vimos que la estimación de máxima verosimilitud no cuenta toda la historia. Si queremos comparar dos grupos de cosas, no podemos simplemente comparar sus promedios muestrales por ejemplo y ya está, o sus estimadores ML del parámetro que nos interese, porque eso es una estimación puntual del estimador ML. En cambio, necesitamos incluir otros factores, como la dispersión de los datos y la cantidad de muestras que hemos recolectado. Esto nos llevó al concepto de intervalo de confianza. Aprendimos a cuantificar el intervalo de confianza utilizando la distribución Normal. Esto se justifica por el hecho de que podemos emplear el Teorema Central del Límite cuando tenemos muchos datos, como por ejemplo, en un sistema en línea moderno. Sin embargo, tengamos en cuenta que esto sigue siendo solo una aproximación la mayor parte del tiempo. En la sección siguiente donde hablaremos del enfoque bayesiano aprenderemos a calcular la distribución del parámetro desconocido.

También aprendimos cómo realizar pruebas de hipótesis, en particular las pruebas A/B, que esta es quizás una de las técnicas más prácticas que podemos aprender en ciencia de datos. En la práctica, esta técnica se puede usar en muchas aplicaciones del mundo real. Cualquier cosa, desde comparar páginas web hasta la fiabilidad de las piezas mecánicas de una nave espacial. Las pruebas A/B que vimos son la prueba Z y la prueba T. La prueba Z supone que la media muestral se distribuye normalmente y que se conoce la varianza de los datos. Sin embargo, en los sistemas modernos, generalmente se recolectan tantos datos que se justifica el invocar al Teorema Central del Límite. Aprendimos que hay pruebas de 1 muestra y de 2 muestras. También aprendimos sobre las pruebas bilaterales y unilaterales.

A continuación, para entender la brecha entre el enfoque frecuentista y el bayesiano, hablemos de algunas extensiones y limitaciones de lo que hemos visto hasta ahora. Al igual que con los intervalos de confianza, podemos dejar de suponer que conocemos la desviación estándar. En este caso, la prueba Z se convierte en la prueba T y el intervalo de confianza Z se convierte en el intervalo de confianza T. Dado que la prueba T está incluida en la librería de Python scipy, esencialmente no se necesita mucho trabajo para usar esta prueba. Sin embargo, como hemos dicho antes, casi siempre tendremos tantos datos que ambos, la Z y la T, llegarán a la misma conclusión.

Hemos visto varios ejemplos, uno de ellos las tasas de cliqueo. Esto es muy útil para sistemas que rastrean donde hacen clic los usuarios, o qué anuncio tiene más engagement. En este caso, dado que los datos son Bernoulli, la media muestral no es Normal. En ese ejemplo es posible que para el ratio de cliqueo se use una prueba que se llama prueba de la Chi cuadrado $\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$. Esto puede ser de interés si en nuestros datos el tamaño muestral n es pequeño.

Otra situación que no consideramos es cuando tenemos muestras pareadas, lo que no suele ser relevante en los sistemas en línea con miles o millones de usuarios, donde solo nos interesa la tasa de clics de un enlace. Sin embargo, puede haber situaciones en las que deseemos realizar una prueba de pares. Por ejemplo, imaginemos que queremos probar un medicamento, pero en lugar de tener dos grupos separados de personas, solo tenemos un grupo de personas y toma el tratamiento estándar, y a ese mismo grupo de personas se le aplica el tratamiento nuevo. Entonces tenemos dos muestras o dos grupos de datos. Pero en cada uno de estos grupos de datos, una observación en un grupo corresponde a otra observación en el segundo grupo. Porque son las mismas personas. En esta situación, no podemos hacer una prueba de dos muestras independientes como hicimos antes con la prueba Z ni con la prueba T. En su lugar, tendríamos que hacer una prueba para datos pareados o dependientes, que es otro tipo de prueba. Ya que las que hemos visto son para muestras independientes. Además las pruebas que vimos son de tipo paramétrico, donde asumimos conocer la distribución de los datos, sin embargo para otro tipo

de datos como datos donde el orden importa o donde no se conoce la distribución, es mejor usar **pruebas no paramétricas**.

Por otro lado, hay otra situación que no consideramos y es en la que tenemos más de dos grupos de datos, por ejemplo, en lugar de solo probar un tratamiento, con respecto a otro, tenemos el tratamiento A, el B, el C, el D, etc. En este caso, el nivel de significación no es fiable. ¿Por qué podría ser esto? Bueno, recordemos que al utilizar un nivel de significación del cinco por ciento, estamos asumiendo un cinco por ciento de posibilidades de una detección falsa o falsa alarma, un 5% de error, de falso positivo. Sin embargo, cada vez que realizamos una nueva prueba, es más probable que encontremos un falso positivo. En la práctica, los investigadores modifican los niveles de significación alfa, con esta corrección conservadora que se llama corrección de Bonferroni donde se divide el alfa por el numero de test que se realizan:

$$\alpha_{Bonferroni} = \frac{\alpha}{\#\ tests}$$

Hay muchas otras preocupaciones en las pruebas A/B frecuentistas que no aparecen en el enfoque bayesiano. Los investigadores a menudo se preocupan por el poder o potencia de una prueba, también conocida como probabilidad de detección. Esto está influenciado por varias cosas, como por ejemplo el número de muestras, la varianza en los datos, el nivel de significación, etc.

Otra preocupación es cuando los investigadores quieren mirar los datos para buscar significancia o no ejecutar una prueba hasta su finalización por razones fuera del alcance de este curso, por ejemplo una razón práctica fácil de entender sería si la obtención de cada dato costase dinero. Básicamente, la idea está en que el valor p puede fluctuar en función del número de muestras. Imagina que estás trabajando como científico de datos y todos los días miras tus datos y verificas el valor p. Teniendo en cuenta que estamos en este punto y hemos recopilado todos los datos el p-valor se ha reducido. Después de recopilar 100 muestras, el valor p no es significativo. Pero después de recolectar 200 muestras, su valor p es significativo y podemos concluir que quizás el anuncio A es mejor que el anuncio B; sin embargo, originalmente teníamos la intención de ejecutar la prueba para 10.000 muestras. Entonces esto no estaría permitido.

En algunas situaciones incluso podría conducir a un dilema ético. Imaginemos que estamos realizando una prueba de un fármaco y descubrimos que el fármaco funciona. Imagina que el grupo de control y el grupo de tratamiento son diferentes. Es posible que como el medicamento funciona, se sienta obligado a ayudar a los pacientes del grupo de control proporcionándoles el medicamento para que se sientan mejor. Pero el paradigma frecuentista dice que no se le permite hacer esto si deseamos una respuesta válida en la prueba de hipótesis.

Entonces, ¿cuál es la conclusión? La conclusión es que ninguno de estos problemas aparece en el paradigma bayesiano. Además, veremos cómo el paradigma bayesiano puede conducir a un software que se adapte a los usuarios a medida que recopila más datos. No tenemos que preocuparnos porque no se nos permita echar un vistazo a los datos, obtener un resultado no válido porque finalizó la prueba antes de tiempo. No tenemos que preocuparnos por la potencia de la prueba o la cantidad de muestras que recolectamos. Tampoco tenemos que preocuparnos por las reglas muy estrictas que hay en el enfoque frecuentista sobre la terminología, por ejemplo, si rechaza o no la H0, o cuál es realmente el significado de un intervalo de confianza.

Como recordarás, el intervalo de confianza no es la probabilidad de que lo que está estimando se encuentre entre el punto A y el punto B. Sin embargo, en el paradigma bayesiano, obtendremos una distribución de nuestra estimación y podremos responder a la pregunta de ¿cuál es esta probabilidad? Y creo que todos podemos estar de acuerdo en que esta idea de rechazar o no rechazar es un concepto extraño. Una pregunta mucho más simple sería ¿cuál es la probabilidad de que el anuncio publicitario A sea mejor que el B? ¿o cuál es la probabilidad de que el medicamento A funcione mejor que el B? Los métodos frecuentistas no pueden responder a estas preguntas. Como veremos en la siguiente sección, el enfoque bayesiano conduce a una interpretación mucho más natural.

PRUEBAS A/B BAYESIANAS

DILEMA EXPLORACIÓN-EXPLOTACIÓN

Supongamos que estamos en un casino y tenemos la opción de elegir entre dos máquinas tragamonedas. Supongamos también que se trata de máquinas tragamonedas muy simples. Donde o ganas o pierdes. No hay nada intermedio. Y si ganas, el premio es siempre el mismo. Entonces, sin perder generalidad, podemos decir que por ejemplo, si ganas la máquina te da un euro y si pierdes te da cero euros.

Repasemos ahora una secuencia imaginaria de eventos. Vamos a intentar pensar cómo elegirías qué máquina jugar. Entonces, al comenzar, todo lo que vemos son dos máquinas, no sabemos absolutamente nada sobre ellas más que lo que acabamos de decir, pueden ganar o perder, y cuando se gana, se obtiene un euro. ¿Cómo elegimos en qué máquina tragamonedas jugar?

De momento, al principio, no importa porque no sabemos nada sobre ninguna de las dos máquinas. Cualquier elección sería equivalente. Así que escojamos una máquina tragamonedas, la número uno, y digamos que perdimos. Entonces, ¿cuál es nuestro próximo movimiento en este punto? Sabemos algo sobre la máquina tragamonedas número uno, sabemos que jugamos una vez y perdimos una vez. ¿Qué sabemos sobre la máquina tragamonedas número dos? Bueno, todavía no sabemos nada.

- Máquina 1: perdimos
- Máquina 2: nada

Nuevamente, vamos a pensar qué máquina tragamonedas elegiríamos para jugar. La mayoría de nosotros probablemente elegiría la número dos. Y es útil pensar por qué elegir la número dos en lugar de la número uno. Bueno, todos sabemos cómo calcular probabilidades, la probabilidad de éxito sería la cantidad de veces que ganamos, dividida por la cantidad de veces que jugamos. En el caso de la máquina tragamonedas número uno, eso sería cero sobre uno, que es cero. Pero para la máquina tragamonedas número dos, es cero sobre cero, lo cual no está definido.

- P(ganar en máquina 1) = 0 / 1 = 0
- P(ganar en máquina 2) = 0 / 0 = indefinido

Y hay algo en nuestra intuición que nos dice que indefinido es mejor que cero, aunque numéricamente no se puede comparar. Bien, digamos que jugamos a la máquina tragamonedas número 2 y ganamos. Entonces hemos jugado en la máquina tragamonedas número uno, una vez y perdimos. Y hemos jugado 1 vez en la máquina tragamonedas número 2 y ganamos.

- Máquina 1: perdimos
- Máquina 2: ganamos

Entonces, ¿a cuál jugamos a continuación?

Una vez más, la mayoría de nosotros por intuición, elegiría la máquina tragamonedas número dos, lo que haría parecer hasta ahora que la máquina tragamonedas número dos tiene más posibilidades de ganar al cien por ciento que la máquina tragamonedas número uno, que tiene una probabilidad del cero por ciento de ganar.

- P(ganar en máquina 1) = 0 / 1 = 0
- P(ganar en máquina 2) = 1 / 1 = 100%

Ahora vamos a pensar ¿estadísticamente, hay algún problema con mis cálculos? ¿Es incorrecto que asigne un cero por ciento a la máquina número uno y un cien por ciento a la máquina número dos? Estas estimaciones son exactamente lo que llamaríamos estimación de máxima verosimilitud. Pero, ¿hay algo incompleto en estas medidas? Luego lo veremos.

Vamos a jugar a la máquina tragamonedas número dos, nuevamente, observamos el resultado y lamentablemente perdemos.

- Máquina 1: perdimos
- Máquina 2: ganamos, perdemos

Pero si miramos estas máquinas en forma de probabilidad, la máquina número dos todavía se ve mejor que la 1. La máquina número uno tiene un cero por ciento de posibilidades de ganar y la máquina número dos tiene un 50 por ciento de posibilidades de ganar.

- P(ganar en máquina 1) = 0 / 1 = 0
- P(ganar en máquina 2) = 1 / 2 = 50%

Así que juguemos de nuevo en la máquina tragamonedas número dos. Y, desafortunadamente, perdemos una vez más en la máquina 2.

- Máquina 1: perdimos
- Máquina 2: ganamos, perdemos, perdemos

Nuestra probabilidad actualizada de ganar en la máquina 2 ahora es del 33 por ciento.

- P(ganar en máquina 1) = 0 / 1 = 0

- P(ganar en máquina 2) = 1 / 3 = 33%

Tengamos en cuenta que esto sigue siendo mejor que la máquina tragamonedas número uno. Entonces, si usamos nuestra estimación de probabilidad para elegir qué máquina jugar, elegiremos la máquina número dos.

Así que volvemos a jugar en la máquina número dos y, lamentablemente, volvemos a perder.

- Máquina 1: perdimos
- Máquina 2: ganamos, perdemos, perdemos, perdemos

Ahora que hemos perdido tres veces y ganado una, la probabilidad de ganar con la máquina número dos es del veinticinco por ciento.

- P(ganar en máquina 1) = 0 / 1 = 0
- P(ganar en máquina 2) = 1 / 4 = 25%

En este punto, si pregunto a qué máquina decides jugar a continuación, habría más desacuerdo sobre si usar la máquina número uno o la máquina número dos. Y con esto me refiero a la primera vez que elegimos la máquina número dos. Como recordarán, la primera vez que elegimos jugar a la máquina número dos, no sabíamos nada al respecto. Y por eso la elegimos, porque podría haber una posibilidad de ganar si no teníamos suficiente información para decidir. Ahora tenemos más datos sobre la máquina número 2. Hemos ganado una vez y hemos perdido tres veces. Así que calculamos nuestra tasa de ganancias como un 25 por ciento. Por otro lado, para la máquina número uno, nuestra tasa de ganancias calculada es cero por ciento, pero solo hemos jugado una vez. Sabemos que esto es de alguna manera menos preciso, por lo que es posible que deseemos jugar a la máquina número uno nuevamente para estar más seguros.

¿Qué tipo de herramientas matemáticas utilizamos en la mente para llegar a esa conclusión? ¿Por qué vamos en contra de hacer una elección codiciosa basada en la estimación de probabilidad de la tasa de ganancias? Bueno después de pensar en ello, si fuéramos un estadístico tradicional que realiza una prueba A/B frecuentista, estaríamos convencidos de que estamos abordando este problema de manera incorrecta. La forma correcta sería decidir antes incluso de entrar al casino cuántos datos debemos recopilar. Esto nos ayudará a determinar el poder estadístico del experimento, lo que llamamos potencia del test. También necesitamos determinar cuál será el tamaño del efecto (effect size). El tamaño del efecto está relacionado con la diferencia entre las tasas de ganancia de las dos máquinas tragamonedas.

Entonces lo que haremos es un cálculo de la potencia y esto me dirá cuántas muestras necesito recolectar en este punto. Pero algunas campanas de alarma deberían estar sonando en nuestra cabeza. ¿Cómo podemos saber el tamaño del efecto si para empezar al inicio ni siquiera hemos jugado en ninguna de las máquinas tragamonedas? Esa es solo una de las muchas razones por las que las pruebas estadísticas tradicionales desde el enfoque frecuentista son algo incómodas.

Bien, digamos que hacemos los cálculos y nos dicen que necesitamos jugar 10000 veces. Ahora, imaginemos el escenario. Hemos jugado en la máquina tragamonedas 1, 5000 veces. La máquina tragamonedas número uno ha ganado tres veces de 5000. Jugamos otras 5000 veces en la máquina 2, y en esa hemos ganado 4000 veces de 5000.

De 10000 veces:

- Máquina 1: Ganamos 3 veces de 5000.
- Máquina 2: Ganamos 4000 veces de 5000.

Pero para que esta sea una prueba estadística válida a los ojos de un estadístico, debemos completar el experimento hasta la 10000. No podemos simplemente detener el experimento antes y decir, creo que la máquina tragamonedas número dos es mejor que la máquina tragamonedas número uno. Eso invalidaría los resultados. Otra razón más por la que las pruebas estadísticas tradicionales son bastante incómodas.

También es posible que estemos ante un escenario más serio. ¿Qué pasa si estamos probando un fármaco que salva vidas y ha tenido éxito el 97% del tiempo en el grupo de prueba? ¿Continuaríamos el experimento dando el placebo al grupo de control, que no tiene ningún efecto? ¿Eso es ético?

Entonces, tal vez deberíamos descartar este método tradicional y usar un método que pueda adaptarse a medida que se recopilan nuevos datos. Esto sería similar a lo que hicimos originalmente usando nuestra intuición. En esta sección, aprenderemos cómo podemos hacer eso algorítmica y cuantitativamente en lugar de solo por simple intuición.

Lo que hemos visto con el ejemplo de las máquinas tragamonedas son dos fuerzas opuestas. Por un lado, deseamos ser un buen estadístico y recopilar una gran cantidad de datos para que podamos hacer una estimación más precisa. A eso lo llamamos **exploración**. Por otro lado, queríamos elegir la máquina tragamonedas que tenía la tasa de ganancias más alta para poder ganar más dinero. A eso lo llamamos **explotación**. Y debido a que estos dos objetivos se oponen, lo llamamos un **dilema**. De ahí el término **dilema de exploración / explotación (explore / exploit dilema).**

Esta sección está dedicada a los algoritmos que resolverán este problema:

1. Epsilon Greedy, que es algo que se usa mucho en aprendizaje por refuerzo.
2. Optimistic Initial Values (valores iniciales optimistas).
3. UCB1 (las siglas de Upper Confidence Bound).
4. Otros UCB: UCB2, UCB Normal.
5. Muestreo de Thompson (Thompson Sampling), también conocido como Bandido Bayesiano (Bayesian Bandit).
6. UCB-Bayes.

Y para no perder de vista el panorama general, todos estos son algoritmos que podemos utilizar en lugar de una prueba A/B tradicional. Son métodos que superan algunos de los problemas incómodos de la estadística tradicional frecuentista, como el tamaño del efecto, el deseo de detener el experimento antes de tiempo y el controvertido p-valor. Estos métodos son adaptables, lo que significa que aprenden sobre la marcha, lo que podría considerarse ventajoso, especialmente en entornos comerciales “en línea” donde estamos obteniendo cada vez más datos nuevos.

APLICACIONES DEL DILEMA EXPLORACIÓN-EXPLOTACIÓN

Vamos a discutir algunas de las aplicaciones de lo que vamos a aprender en esta sección. De cara al mundo real. Existen muchas aplicaciones de estos métodos en el mundo real, de hecho, los algoritmos que veremos son unos de los algoritmos más prácticos que aprenderemos. La cuestión está en que el concepto de comparar cosas para ver cuál es mejor, se puede aplicar a casi cualquier caso o cualquier negocio. Siempre se puede hacer uso de estos algoritmos en cualquier momento en que queramos comparar dos cosas, en Marketing e internet para elegir qué diseño, qué botón, qué título, qué publicación es mejor, en Medicina para comparar medicamentos, en Ingeniería e industria para comparar piezas, en todos estos casos y más, podemos aplicar estos métodos.

Por ejemplo, supongamos que somos una compañía que hace algún artículo que las personas van a usar. Imagina que somos Apple y acabamos de crear un iPhone nuevo y espectacular y queremos contárselo al mundo. Le pedimos a un diseñador que nos haga un par de anuncios. Como el diseñador es muy talentoso, ambos anuncios nos gustan. Entonces, ¿cuál deberíamos elegir? Por supuesto, debemos elegir el anuncio para el iPhone en el que creemos que es más probable que el usuario haga clic o compre. Una forma posible de medir esto es determinar la tasa de cliqueo de cada anuncio.

Ahora bien, ¿cómo determinamos la tasa de clics? Bueno, es simplemente la relación entre los clics y el número total de impresiones. Entonces, por ejemplo, si mostramos un anuncio 1000 veces, pero solo 10 de esas veces, el usuario hizo clic en él, entonces tenemos una tasa del 1%. Entonces, ¿cómo puedo medir el porcentaje de clics en la práctica? Bueno, podemos hacer un experimento. Digamos que mostramos el primer anuncio a un millón de usuarios y el segundo anuncio a otro millón de usuarios. Y luego comparo las tasas de clics que obtengo de cada anuncio. Y obtengo la respuesta. Pero hay un problema con eso. ¿Cómo sé que un millón es el número correcto? De hecho, ¿por qué no muestro el anuncio 1000 veces solamente? ¿o por qué no mostrar el anuncio solo 10 veces?

Bueno, aquí es donde debemos recordar lo que vimos cuando hablamos de probabilidades, la única forma de tener precisión absoluta en nuestra estimación estadística es recolectar un número infinito de muestras. A medida que recolectamos más y más muestras, el intervalo de confianza de la estimación disminuye. Entonces, la respuesta es recolectar tantas muestras como sea posible. Pero hay otro problema con eso: si un anuncio es mejor, eso necesariamente significa que el otro anuncio es peor. Eso significa que el mejor anuncio generará más ganancias y el otro anuncio generará menos ganancias. Entonces, si muestro el peor anuncio un millón de veces, eso significa que he desperdiciado un millón de impresiones por una tasa de clics sub-óptima. En otras palabras, mi deseo es el de mostrar siempre solo el mejor anuncio para aprovechar la tasa de clics óptima pero esto está en desacuerdo con mi otro deseo de tener una respuesta precisa sobre cuáles son las tasas de clics.

Ahora, obviamente, Apple no es la única compañía en el mundo que vende cosas, la industria de la publicidad en línea es una industria de miles de millones de dólares. Casi todas las empresas modernas del mundo utilizan publicidad en línea. En otras palabras, las técnicas que estamos aprendiendo aquí son aplicables a cualquiera de las empresas modernas que necesitan hacer este tipo de estudios. Sin embargo, ni siquiera tiene que ser un negocio exitoso para hacer uso de estas técnicas. Simplemente puede ser alguien que tenga un sitio web en Internet. Suponiendo que tiene tráfico y puede elegir entre dos o más opciones en su sitio web, puede aplicar estas técnicas. Por ejemplo, supongamos que contratamos a un diseñador para crear un nuevo diseño para nuestro sitio web y nos gustaría saber si el diseño anterior obtiene mejor tráfico que el nuevo diseño. O quizás nos gustaría saber si un botón grande de comprar funciona mejor en la parte superior de su página de ventas o en la parte inferior. Quizás nos gustaría saber qué tipo de precio utilizar en los productos del sitio web. Por ejemplo, podríamos vender el producto por veinte euros fijos, o podría venderlo en 19.99. ¿Cuál tiene más probabilidades de generar más conversiones? Por supuesto, todos estos son en la base el mismo problema, deseamos saber qué opción es la mejor y utilizar exclusivamente la mejor opción, pero necesita recopilar datos para hacer esa elección.

Como nota al margen, desde nuestra perspectiva, tanto las tasas de clics como las tasas de conversión se tratan de la misma manera, ya que ambas son una forma de calcular una tasa de éxito genérica. Obviamente, desde una perspectiva monetaria, un clic no es tan valioso como una conversión. Pero para nosotros, no necesariamente estamos midiendo los ingresos o las ganancias en sí, solo la tasa de éxito.

Otro gran ejemplo es el servicio de noticias de Facebook y, por supuesto, esto no es exclusivo de Facebook. Esto se aplica a cualquier empresa que tenga un servicio de noticias como The New York Times. Idealmente, queremos tener artículos de noticias en los que la gente haga clic o lea. ¿Y por qué es eso? Verás, cuando las personas hacen clic en estos artículos, ven más anuncios. Cuando te muestran un anuncio, ganan dinero. Así es el círculo de la vida. Los anunciantes ganan dinero cuando compramos cosas. Y los sitios web que muestran anuncios obtienen dinero de los anunciantes simplemente por usar ese espacio en su sitio web para publicar el anuncio. Así que The New York Times quiere que hagas clic en los artículos y Facebook quiere mantenerte involucrado desplazándote por las noticias durante el mayor tiempo posible. Cuanto más te desplazas, más anuncios ves. De nuevo, ¿qué estamos haciendo? Estamos midiendo una tasa de éxito, eso significa que hicimos clic en el artículo de Facebook. Eso significa que leemos, miramos o de alguna manera interactuamos con el elemento del servicio de noticias. Una vez que hayamos medido la tasa de éxito de cada elemento de la fuente de noticias, podemos mostrárselos a los usuarios en orden, donde ese orden depende directamente de la tasa de éxito, así que enseñaríamos primero el anuncio con tasa de éxito más alta, y así sucesivamente.

Si recordamos la distribución de Bernoulli se usa para medir el éxito o el fracaso. Son ideales para el escenario en el que solo tenemos dos resultados, haga clic o no haga clic, compre o no compre. Pero, ¿qué pasa si la recompensa es un valor continuo? En ese caso, una distribución Gaussiana puede ser más apropiada. Algunos ejemplos son la medición de ingresos, o ganancias, o el número de usuarios.

En resumen, todo esto lleva al dilema de exploración- explotación, que es uno de los conceptos más importantes en el aprendizaje automático, sobre todo en aplicaciones prácticas.

EPSILON GREEDY

Vamos a ver un algoritmo que probablemente será uno de los más importantes de esta sección, Epsilon Greedy. La razón es porque se usa mucho en aprendizaje automático, sobre todo aprendizaje por refuerzo, Q-learning y Deep QL. Por otro lado, veremos que los otros métodos, como UCB1 y el método del bandido bayesiano, a menudo funcionan mejor. Pero, en general, todos estos algoritmos tienen la misma idea detrás, desde la perspectiva de programación. Todos nos ofrecen una forma de equilibrar la exploración y la explotación.

Recordemos cuál es el dilema de explorar y explotar. Si solo tomamos una estimación puntual de máxima verosimilitud de la tasa de ganancia, esto es muy perjudicial para la exploración y la recopilación de datos. Podríamos tener suerte o mala suerte con esa estimación en particular y luego terminar explotando algo

sub-óptimo porque la tasa de ganancia no era la óptima. Esta estrategia tiene un nombre, a esto le llamamos el método codicioso (greedy en inglés). Este término básicamente significa hacer algo a visión corta o inmediata, en otras palabras, es como usar solo la información disponible de forma inmediata en ese momento, como una heurística para tomar una decisión. En nuestro caso, ser codicioso significa jugar en la máquina con la mayor probabilidad máxima de ganar sin tener en cuenta la cantidad de datos que hemos recopilado o la confianza que tenemos en las tasas de ganancias que calculamos.

Entonces, si esto se llama estrategia codiciosa o método Greedy, el método Epsilon Greedy es un tipo de modificación de esta estrategia codiciosa. Epsilon Greedy dice que en lugar de tomar siempre la acción codiciosa, tendré una pequeña probabilidad de hacer algo completamente al azar. Es decir, con una pequeña probabilidad, voy a elegir una máquina tragamonedas al azar sin tener en cuenta cuál es la tasa de ganancias. Y, por supuesto, esa pequeña probabilidad viene dada por el valor de epsilon. Por lo general, elegiríamos el valor de epsilon como 5% o 10%.

Si tuviéramos que escribir un pseudocódigo, así es como se compararían lo métodos:

Greedy:

```
while True:

    j = argmax(predicted bandit means)

    x = play bandit j and get reward

    bandits[j].update_mean(x)
```

Epsilon Greedy:

```
while True:

    p = random number in [0,1]

    if p < epsilon:

        j = choose a random bandit

    else:

        j = argmax(predicted bandit means)

    x = play bandit j and get reward

    bandits[j].update_mean(x)
```

Supongamos que para cada algoritmo nunca dejamos de jugar. Entonces cada algoritmo va dentro de un ciclo infinito. En el caso puramente codicioso o método Greedy a secas, seleccionamos la opción que tiene a mayor media actual. Una vez que hemos seleccionado, jugamos en esa opción, en esa máquina y recolectamos la recompensa llamada X. Recuerda que X puede ser cero o uno. Ganar 1 euro o perder que es no ganar nada. Una vez que tengamos X podemos actualizar la estimación de la media de la máquina que seleccionamos porque tenemos un resultado nuevo.

En el caso del método Epsilon Greedy, hay un paso adicional en el que seleccionamos la máquina: en lugar de simplemente tomar el máximo, comenzamos eligiendo un número aleatorio entre cero y uno. Si es menor que epsilon, elegimos al azar de una distribución uniforme. De lo contrario, elegimos el que tiene promedio o tasa máxima. Esto asegura que elegimos una apuesta aleatoria con probabilidad epsilon. Finalmente, el último paso sería jugar la opción seleccionada y luego actualizamos la estimación.

Vamos a analizar algunos puntos clave que son importantes para analizar la eficacia del método. Recordemos que la razón por la que queremos explorar, que se logra al tener un épsilon distinto de cero, es para que podamos recopilar datos sobre cada opción, o cada máquina en la que podemos jugar. Y la razón por la que queremos recopilar datos sobre cada máquina es porque queremos que nuestras estimaciones de la tasa de ganancias sean más precisas. Pero, ¿en qué momento decimos que tenemos suficientes datos como para dejar de explorar? ¿Qué sucede si dejamos que el algoritmo continúe ejecutándose para siempre? En ese escenario, nuestro algoritmo nunca dejará de explorar y, por lo tanto, nuestra recompensa colectiva total será sub-óptima. Porque por definición, si una de las opciones es óptima, la otra no lo es. Ahora, supongamos que tenemos dos opciones con tasas de ganancia del 90% y 80%. Claramente, si supiéramos esta información de antemano, solo jugaríamos en la máquina con una tasa de victorias del 90% y nunca jugaríamos en la que tiene una tasa de victorias del 80%. Pero por supuesto, no podemos estimar con precisión esas tasas y saber cuál es mejor sin pasar antes por el paso de recopilar los datos. Por otro lado, a largo plazo, nunca alcanzaremos una tasa de ganancias del 90% debido al hecho de que siempre tenemos una pequeña posibilidad en el Epsilon Greedy de explorar y jugar en la máquina sub-óptima. En resumen, podemos desglosar nuestro cálculo de la tasa de ganancias promedio de esta manera.

$$E(R) = (1 - \epsilon)\ 0.9 + \epsilon \left(\frac{0.8 + 0.9}{2}\right)$$

Primero, supongamos que después de mucho tiempo hemos identificado la máquina con el resultado óptimo, de modo que durante el tiempo 1-epsilon seleccionaremos la máquina óptima, la que nos da la recompensa esperada de 0.9 (90%). Y por otro lado, el épsilon % de las veces, elegiremos a cualquiera de las opciones con igual probabilidad, porque en este caso era una selección aleatoria, ambas elecciones tienen la misma probabilidad de ser elegidas. Entonces, cuando eso sucede, nuestra recompensa esperada es a medio camino entre 0.8 y 0.9. Por lo tanto, nuestra recompensa total esperada no es la máxima.

Una opción para mejorar este proceso es tener un épsilon que en vez de ser un valor fijo pequeño, sea una función decreciente como estas que se han utilizado en varias aplicaciones en la literatura:

$$\epsilon(t) \propto \frac{1}{t}$$

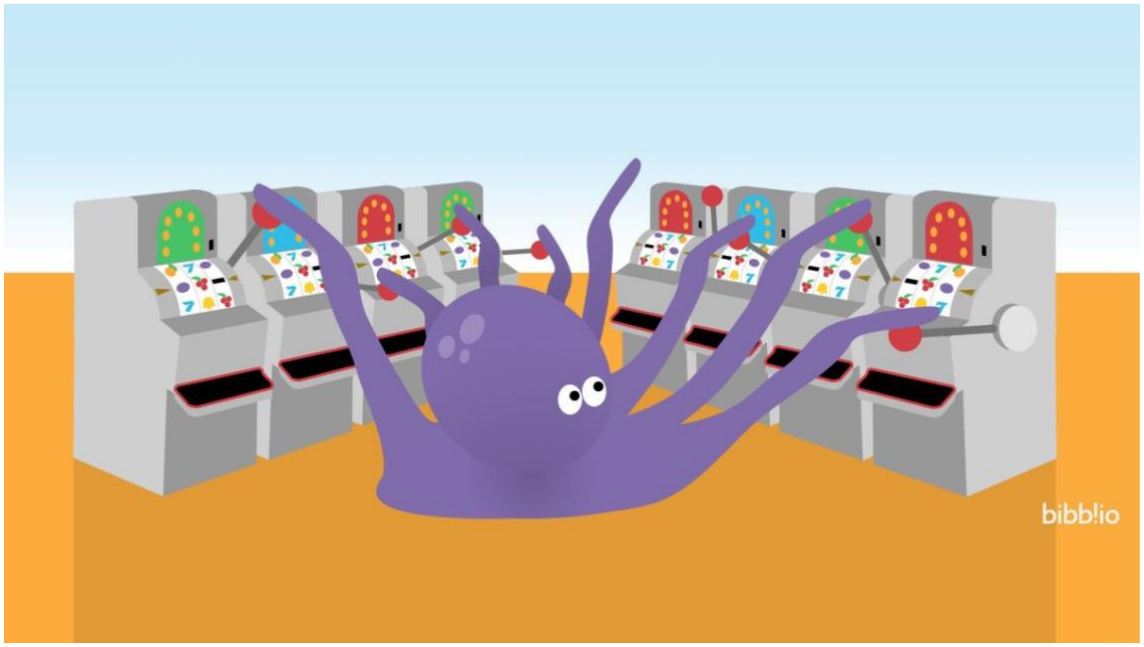
$$\epsilon(t) = \max(\epsilon_0 - kt\ ,\ \epsilon_{min})$$

$$\epsilon(t) = \epsilon_0 \alpha^t$$

$$\epsilon(t) = \frac{a}{\log(bt + c)}$$

La primera hace que el valor de epsilon decrezca proporcionalmente a uno sobre el número de pasos. La segunda deja que epsilon decaiga linealmente y luego se detenga en cero o en un valor pequeño, ligeramente mayor que cero. La tercera es una especie de enfriamiento exponencial y la cuarta con el logaritmo. Hay muchas opciones pero la idea es siempre la misma. Hacer que épsilon disminuya con el tiempo.

BANDIDO MULTIBRAZO: SOLUCIÓN FRECUENTISTA CON TEST A/B



Antes de ver los resultados de aplicar el algoritmo de Epsilon Greedy para ponernos en práctica, vamos a ver uno de los problemas clásicos del aprendizaje por refuerzo, el problema del bandido multibrazo (multi-armed bandit) y vamos a resolverlo a través del método clásico de la prueba A/B que vimos en la sección anterior. Y luego lo haremos con Epsilon Greedy para ver en qué se diferencian los dos algoritmos.

El problema del bandido multibrazo consiste en que a un agente se le ofrece la posibilidad de jugar con N máquinas tragaperras, a las que se les suele llamar "bandidos" o "brazos", que ofrecen diferentes recompensas. La recompensa que ofrece cada uno de los bandidos, cada una de las máquinas, viene dada por una distribución de probabilidad que es diferente. Distribución que el agente no conoce a priori. El objetivo es maximizar el beneficio obtenido después de jugar una cantidad dada fijada de antemano. Esto es, descubrir el bandido que ofrece la recompensa promedio mayor en el menor número de jugadas. Para, de este modo, maximizar el beneficio jugando la mayor cantidad de veces posible únicamente con este bandido.

El problema de identificar al mejor bandido es un problema más complejo de lo que parece al principio. Dado que la recompensa viene dada por una distribución de probabilidad, es posible que una de las peores máquinas nos devuelva grandes recompensas en las primeras jugadas. De forma análoga, con el mejor bandido puede que no obtengamos ningún premio en las primeras jugadas. Por lo que si el agente no realiza las suficientes tiradas de prueba es posible que se decante por un bandido que no sea el óptimo.

En el lado opuesto se encuentra el problema de jugar demasiadas veces con un bandido no óptimo en la fase de exploración. Siendo esto lo que se conoce como el dilema *exploración-explotación*. Esto es, ¿cuándo debemos parar de explorar para explotar el conocimiento adquirido? Que ya mencionamos anteriormente.

Una de las posibles soluciones que tenemos en nuestra mano para resolver este problema es utilizar un Test A/B. Esto es, evaluar durante un periodo de tiempo todos los bandidos por igual y decidir una vez finalizado este periodo de prueba cuál es el óptimo. O, si los datos no son concluyentes, volver a realizar otra prueba con más muestras.

Vamos a ver los resultados de simular la utilización de un Test A/B para el Bandido Multibrazo en Python. Si se ejecuta el código se puede ver que son necesarias 7 iteraciones, 3500 simulaciones (porque en cada iteración se hacen 500 simulaciones) para identificar el mejor bandido. Aunque solamente para los dos mejores, ya que los tres peores se han podido descartar después de una sola iteración. Resultados que

pueden cambiar ligeramente en caso de que se modifique el valor de la semilla. Con esto, después de 8500 simulaciones se han obtenido 692 casos positivos o favorables, lo que corresponde a un 8,1%. Por debajo del 10% máximo que se esperaría del mejor bandido. Aunque una vez obtenido este resultado ya se puede jugar solo con este bandido y mejorar la recompensa media. En resumen, este es un método que requiere muchas pruebas antes de obtener unos resultados concluyentes, por eso podemos estar mucho tiempo jugando con bandidos que no son óptimos. Por eso vamos a ver otros métodos alternativos.

BANDIDO MULTIBRAZO: EPISILON GREEDY

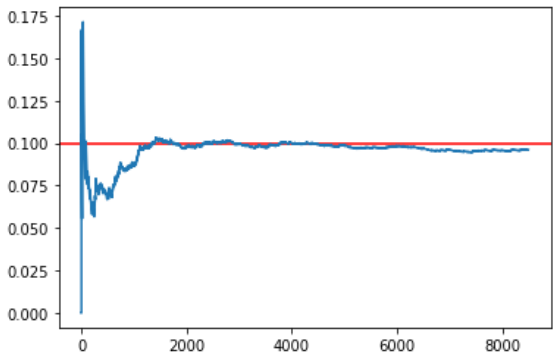
Vamos a seguir con el mismo ejemplo del Bandido Multibrazo. Ya hemos visto los resultados de resolver el problema mediante un test A/B frecuentista. Con ese algoritmo, la solución fue jugar con cada uno de los bandidos una cantidad de veces determinada hasta que estábamos seguros de cuál era el mejor de los bandidos. Esta aproximación no es eficiente, ya que en muchos casos se puede saber rápidamente cuáles son los peores, por lo que se puede plantear otra estrategia más eficaz. Y no perderíamos tiempo jugando con los peores. Tiempo o dinero.

La estrategia del algoritmo de Epsilon Greedy en la que se selecciona el mejor de los bandidos hasta ese momento, salvo un porcentaje de veces en las que se juega de forma aleatoria, nos ofrece una alternativa que proporciona un balance para el dilema de exploración y explotación. La estrategia Epsilon Greedy es realmente sencilla. En esta, en primer lugar, se decide si se juega con el mejor bandido, aquel que ha devuelto la mayor recompensa promedio hasta el momento, o de forma completamente aleatoria. El porcentaje de veces en las que la estrategia jugará de forma aleatoria se seleccionará mediante un valor epsilon. Así se obtendrá la mejor recompensa con la información disponible, al mismo tiempo que es posible explorar otras soluciones con las tiradas aleatorias. Esta simple estrategia permite maximizar la recompensa ya que se jugará preferentemente con el bandido que ha ofrecido la mayor recompensa hasta ese momento. Sin tener que esperar a probar con cada uno de los bandidos la cantidad de veces que ha definido al principio, como en el caso frecuentista.

Es importante tener en cuenta que si el valor de epsilon es muy pequeño, el algoritmo no podrá identificar rápidamente la mejor solución. Pero si el valor es muy alto, una vez identificada la mejor solución, seguirá jugando una cantidad de veces elevada con una solución que no es la óptima. Por lo que el valor de epsilon es un compromiso que tenemos que tener en cuenta para establecer un balance entre la exploración y la explotación.

Para ver los resultados, aplicados al mismo ejemplo que pueden encontrar en los códigos de Python del curso, primero se tiene que decidir el porcentaje de veces que se jugará de forma aleatoria, por ejemplo, un 5%. Una vez hecho esto solamente se tiene que seleccionar un número aleatorio y en base a este seleccionar el bandido. La selección va a ser espilon veces de forma aleatoria y el resto de las veces va a seleccionar el bandido que tiene la mejor recompensa hasta el momento.

De cara a comparar la solución de Epsilon Greedy con la solución obtenida con el test A/B, en primer lugar vamos a ver cómo funciona el algoritmo con 8500 jugadas. En este caso se obtiene una recompensa media de 9.6%, bastante superior al 8,1% que se observó con el test A/B. Además, se puede comprobar la evolución de la recompensa media, para lo que se puede imprimir la recompensa media en cada jugada.

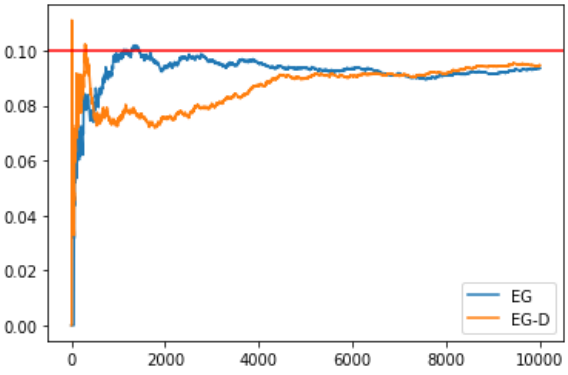


Lo que muestra que, en torno a las 1000 jugadas, la recompensa media obtenida ya se acerca a la final. Lo que indica que en este punto el algoritmo ha decidido jugar mayoritariamente con el bandido que ofrece una recompensa del 10%. Una conclusión a la que se ha llegado bastante más rápido con este método que con el uso del test A/B. En las primeras jugadas se puede ver una recompensa promedio por encima del máximo, pero es algo que puede suceder debido a la aleatoriedad de las recompensas. Aunque esto se corrige rápidamente a medida que aumentan el número de jugadas. Posiblemente en torno a las 1000 jugadas ya no sea necesario explorar otros resultados. Pero el algoritmo seguirá jugando un 5% de las veces aleatoriamente, algo que veremos después cómo se puede mejorar.

BANDIDO MULTIBRAZO: EPSILON GREEDY CON DECAIMIENTO

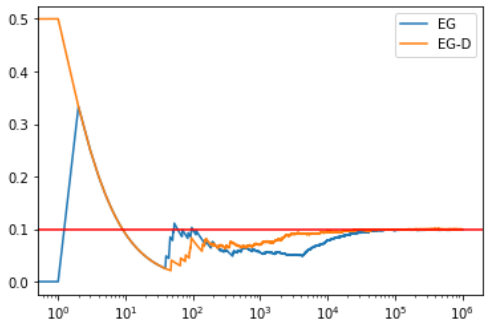
Anteriormente vimos cómo se podía usar la estrategia Epsilon Greedy para resolver un problema tipo bandido multibrazo. Una estrategia que nos había dado mejores resultados que un test A/B. Pero esta estrategia tiene un problema, una vez que se sabe cuál es el mejor bandido se continuará jugando una cierta cantidad de veces con bandidos que no son el óptimo. Lo cual se puede resolver utilizando una estrategia **Epsilon Greedy con decaimiento**, es decir, cambiando la probabilidad con la que se juega con bandidos aleatorios, es decir, en vez de usar un epsilon fijo, hacer que a medida que aumenta el conocimiento del agente del entorno, el epsilon decrezca. De este modo, en los primeros episodios o iteraciones, cuando se desconoce la recompensa esperada, se juega aleatoriamente y, a medida que se tiene más información, se juega de manera óptima. Se pueden usar diferentes estrategias para obtener un valor de epsilon que decaiga con el tiempo. Ya que solamente es necesario que este decaiga a medida que aumente el número de episodios. Otra opción es comenzar con un valor de épsilon dado y reducir el valor de este multiplicándose en cada episodio por un factor que sea inferior a la unidad. Así, el valor en cada tirada sería epsilon * decay^N donde N es el número de episodios.

Ahora vamos a comparar los resultados que se obtienen para 10.000 episodios con Epsilon Greedy y Epsilon Greedy con decaimiento.



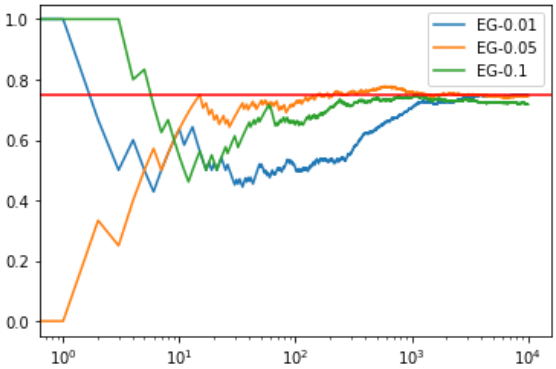
Lo primero que se puede ver es que el método sin decaimiento tiene una recompensa de 0,093, mientras el método con decaimiento de 0,094. Una ligera mejora para el segundo, pero que todavía no es significativa. El valor que se debería esperar para el método sin decaimiento es de 0,098. Esto es 95% por la recompensa del mejor bandido más 5% por la recompensa promedio de todos los bandidos. Esto es: 0,95 * 0,1 + 0,05 * (0,18/3) lo que da 0,098. Por otro lado, a medida que aumenta el número de episodios el valor de Epsilon Greedy con decaimiento debe acercarse al del mejor bandido. Ya que, el valor de epsilon tenderá a cero, por lo que solamente jugará con este. Esto es algo que se puede comprobar aumentando el número de episodios. Por ejemplo, con un millón de episodios más los valores pasarán a ser de 0,0985 y 0.0998 para Epsilon Greedy y Epsilon Greedy con decaimiento respectivamente. Valores ya muy cercanos a los que se esperaría para ambos agentes.

La evolución de los 1,1 millones de episodios se puede ver en la siguiente gráfica. En donde se puede ver que ambas estrategias se acercan al valor óptimo a partir de los 10.000 episodios.



La estrategia Epsilon Greedy con decaimiento ofrece ciertas ventajas respecto a Epsilon Greedy, especialmente cuando el número de episodios aumenta. Al seleccionar el bandido inicialmente al azar y, a medida que avanzan los episodios, cada vez de una forma más avariciosa permite obtener mayores recompensas. Ya que una vez que se conoce cual es el mejor bandido ya no es necesario explorar en busca del mejor.

Es importante saber que el valor de epsilon, como hemos dicho anteriormente, puede afectar al resultado final. Por ejemplo, si tomamos tres epsilon diferentes: 0.01, 0.05 y 0.1 y hacemos primero 10.000 iteraciones y se ve que el epsilon más pequeño 0.01 es más lento en converger hacia la recompensa óptima. Pero si incrementamos las iteraciones vemos que a largo plazo es el que mejor se acerca.



Entonces con menos repeticiones, y con epsilon más pequeños se obtienen recompensas acumuladas altas, más rápidamente que con epsilon más pequeños como el del 1%. Con epsilon 1% también se converge pero más lentamente. Sin embargo si miramos a largo plazo, la recompensa acumulada es menor para epsilon más grandes, y es mejor para el epsilon más pequeño. Entonces hay un trade-off entre

si queremos una convergencia rápida o si queremos una recompensa más alta a largo plazo. En algunas aplicaciones tendremos que tener en cuenta el tiempo que tenemos para el experimento, por ejemplo, si no nos podemos permitir demasiadas iteraciones y no podemos esperar a que converja a muy largo plazo, entonces un epsilon más alto estará bien. La decisión tiene que reflejar bien nuestros requerimientos.

VALORES INICIALES OPTIMISTAS

Hasta ahora hemos visto que con la estrategia llamada Epsilon-Greedy se obtienen mejores resultados que con un test A/B. Aunque Epsilon-Greedy tiene un problema: cuando el número de episodios a jugar es elevado, continúa explorando los peores bandidos con una probabilidad fija epsilon. Para solucionar esto, vimos la opción de que este parámetro se reduzca con el tiempo. Comprobando que así se obtienen mejores resultados a largo plazo. Ya que una vez se ha identificado al bandido que ofrece la mayor recompensa promedio el agente estará más tiempo explotando este, en lugar de explorar el resto de los bandidos. Pero aún así es posible reducir la fase de exploración si se cuenta con una estimación inicial optimista de la recompensa de cada bandido. Una estrategia a la que se conoce con el nombre de Valores Iniciales Optimistas. El método de Valores Iniciales Optimistas es una modificación extremadamente simple del método Epsilon Greedy. En otras palabras, ni siquiera necesitamos ya el concepto de epsilon que propone una exploración aleatoria. La forma en que funciona es la siguiente, la estimación de la media muestral se puede actualizar en función de la estimación en el paso anterior anterior. Nuestra estimación inicial (de lo que partimos) es cero para asegurar que podemos obtener exactamente la media muestral. El método de Valores Iniciales Optimistas dice esto: ¿qué pasa si en lugar de inicializar la estimación inicial a cero, elegimos un valor realmente grande y lo incluimos en la estimación de la media muestral? De esta manera, no estamos realmente estimando la media. De hecho, estamos sobre-estimando la media.

Entonces se podría usar este punto de partida para seleccionar el bandido mediante una estrategia puramente avariciosa, esto es, con epsilon igual a cero, un Greedy puro sin epsilon. Así, si la estimación es correcta siempre se explotará al mejor bandido, obteniendo de este modo la mayor recompensa posible porque solamente se ha jugado con el mejor bandido. Por otro lado, si no es así, al actualizar la recompensa con los datos reales se podrá comparar esta estimación con las estimaciones iniciales. En caso de que la estimación para otro bandido sea mejor, se pasará a jugar con este. Así hasta que se pase a jugar siempre con el bandido que ofrece las mayores recompensas.

¿Por qué tienen que ser ***optimistas*** los valores iniciales?

La estrategia se llama Valores Iniciales Optimistas y esto es así porque los valores tienen que ser realmente optimistas para que este método pueda funcionar de forma correcta. Para esto nos podemos fijar en un ejemplo.

Supongamos que tenemos dos bandidos con una recompensa promedio de 1 y 2 respectivamente. Así, si fijamos unos valores iniciales de 3 y 2 el agente jugará inicialmente con el primero de los bandidos, al que se le ha asignado una recompensa promedio de 3. Siendo esto así hasta que se compruebe que la recompensa real del primero es inferior a 2, el valor que se le ha asignado al segundo bandido. Por lo que el agente pasa a explotar únicamente el segundo de los bandidos, ya que su recompensa esperada es 2.

Por otro lado, si los valores iniciales no son optimistas el agente puede explotar un bandido que no sea óptimo. Algo que se puede comprobar si se asigna valores iniciales de 0,5 y 0,2. En este caso el agente jugará solamente con el primero ya que la recompensa esperada es 1 y esta es mayor que 0,2, la asignada al segundo. Por lo que la solución obtenida no será la óptima.

A la hora de aplicar esta estrategia hay que tener en cuenta que es posible que no se obtenga un resultado óptimo. Algo que se produce porque el agente no explora en ningún momento otras opciones. Como se ha visto una posible causa sería no seleccionar correctamente los valores iniciales. Lo que llevaría al agente a seleccionar una solución que no es óptima. Por otro lado, también se puede dar este problema si la recompensa promedio de dos bandidos es muy parecida. En este caso, si en algún momento, debido al carácter aleatorio de las recompensas, la recompensa estimada para un bandido es inferior a la real y

existe otro bandido con una mejor, es posible que se seleccione el segundo y no se vuelva al primero en ningún momento. Por lo que la solución final no será la óptima. Un problema similar también se puede dar con Epsilon Greedy, aunque en este caso, al explorar siempre las recompensas de todos los bandidos, el problema suele ser puntual. Esto es, solo hasta que se comprueba que la solución no es la óptima.

Entonces la idea del método es la siguiente. Primero, inicializamos a nuestros bandidos para que sus medias iniciales sean valores muy grandes. Como nota al margen, deben ser finitos porque si eliges infinito, no puedes hacer más cálculos con él. El infinito más cualquier cosa sigue siendo infinito. Luego, dentro de nuestro experimento, en lugar de hacer cualquier tipo de exploración aleatoria, simplemente elegimos el bandido con la media estimada más alta ya que inicializamos este valor para que sea muy grande. Esto no está necesariamente cerca de la verdadera media, sino que es una sobreestimación de la verdadera media. El siguiente paso es considerar la parte del código donde elegimos con quien jugar a continuación. Como hemos dicho antes, ya no necesitamos elegir nada al azar. Podemos simplemente actuar de manera codiciosa y elegir al bandido con la mayor media estimada. La pregunta es, ¿por qué funciona esto? Vamos a pensarlo.

Nuestro objetivo es equilibrar la exploración y la explotación, la exploración es lo mismo que decir que deseamos recopilar una gran cantidad de datos. Entonces, ¿qué sucede con nuestra media estimada? Si es al principio de nuestro experimento y aún no hemos recopilado muchos datos, entonces nuestra media estimada será muy grande, no porque la media verdadera sea grande, sino porque establecimos el valor inicial muy grande y aún no puede converger al verdadero valor. Por otro lado, una vez que hayamos recopilado muchos datos, la media estimada se hará cada vez más pequeña hasta que dejemos de elegir a ese bandido. Recordemos que la media muestral es el promedio aritmético de todas las muestras que se recopilaron. Entonces, incluso si tenemos un valor extremadamente grande, su efecto desaparecerá cuando tengamos 1.000 o 10.000 muestras.

Una pregunta interesante que podemos hacer es esta: en Epsilon Greedy, vimos que nuestra estimación de la media de cada bandido converge hacia la media verdadera. ¿Crees que esto sucederá con el método de valores iniciales optimistas? La respuesta es: no tiene por qué. Recordemos que, dado que utilizamos el método codicioso, no hay garantía de que recopilemos una gran cantidad de muestras para ninguno de los bandidos. Lo que sucederá es que si las medias estimadas para los bandidos sub-óptimos descienden por debajo de la media estimada para el bandido óptimo, dejaremos de explorar esos bandidos sub-óptimos por completo. Y eso es porque estamos siendo codiciosos. Entonces, el único resultado que podemos esperar es que esas medias estimadas estén por debajo de la media estimada de los bandidos óptimos. No que hayan convergido hacia sus verdaderas medias. De hecho, ni siquiera podemos esperar que el bandido óptimo tenga una buena estimación de la media verdadera porque el valor inicial podría haber sido tan alto y el número de ensayos lo suficientemente bajo, como para que la media estimada siga siendo una sobre-estimación al final del experimento.

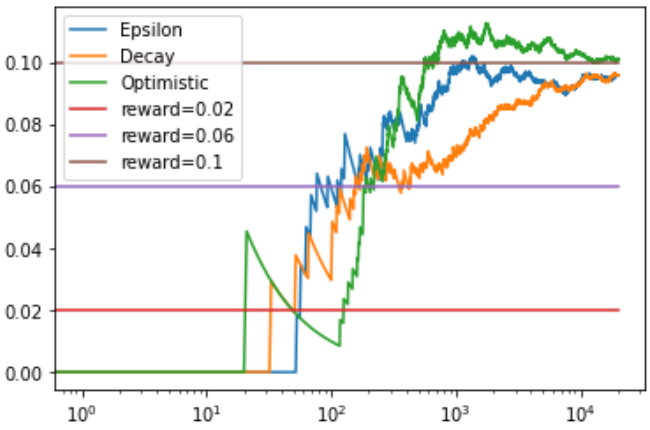
La siguiente pregunta que quiero abordar es ¿cuál es el papel del valor inicial entonces? Bueno, si establecemos este valor muy alto, esto hará que el algoritmo explore más, porque el método codicioso creerá que la apuesta tiene una alta recompensa incluso cuando no la tiene.

¿Qué tan altos deben ser los valores iniciales? Si establecemos el valor inicial extremadamente alto, entonces estamos diciendo que queremos más exploración porque tomará más tiempo para que el valor estimado de la media descienda hacia su verdadero valor. Si establecemos el valor inicial solo un poco alto, entonces estamos diciendo que solo queremos un poco de exploración, porque tomará mucho menos tiempo para que la media estimada descienda hacia su verdadero valor.

Entonces, el valor inicial es un hiper-parámetro que controla la cantidad de exploración. Los valores iniciales altos significan más exploración y los valores iniciales más pequeños significan menos exploración. Pero cuando fijemos un valor inicial, debería ser mayor que el típico valor de recompensa que se esperaría obtener, para que el algoritmo funcione.

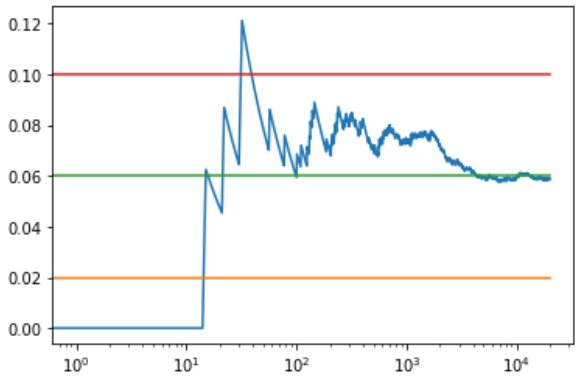
BANDIDO MULTIBRAZO: VALORES INICIALES OPTIMISTAS

Veamos el siguiente ejemplo donde simulamos datos para tres bandidos, con tasas de ganancia de 2%, 6% y 10% respectivamente. Esto quiere decir que el bandido óptimo es el último. Los resultados se muestran en la siguiente gráfica, para los métodos Epsilon Greedy, Epsilon Greedy con decaimiento y Valores Iniciales Optimistas.



Aquí se puede observar que si se configuran bien los Valores Iniciales Optimistas, es el método que detecta antes cuál es el mejor bandido. Para comparar los algoritmos también nos podemos fijar en el número de veces que ha jugado cada bandido con las soluciones que no son óptimas. En el caso de Epsilon Greedy han sido 381 y 355 con el primero y segundo bandido, respectivamente. Valores que mejoran ligeramente con en el caso de Epsilon Greedy con decaimiento, 333 y 352 para el primero y segundo, respectivamente. Finalmente, en el caso de Valores Iniciales Optimistas solamente se han explorado los dos primeros bandidos en 90 y 16 casos, respectivamente. Esto es una ventaja de cara a no seleccionar soluciones sub-optimas.

Veamos también que, en caso de que se configure mal el método, no dará buenos resultados. En caso de no fijar los valores iniciales correctamente, es posible que el agente nunca seleccione el tercer bandido que es el mejor, como se puede ver en el siguiente gráfico.



En este caso lo que se puede observar es que el agente ha jugado principalmente con el segundo bandido (franja horizontal verde), ya que los valores iniciales han provocado que juegue poco con el tercero (franja horizontal roja). Provocando que la estimación de la recompensa esperada no sea la correcta. En concreto solamente ha jugado 6 veces con el último bandido, con lo que ha estimado que su recompensa promedio es 0,015, muy por debajo de la real e incluso la del segundo bandido.

En resumen, la estrategia de Valores Iniciales Optimistas para el problema Bandido Multibrazo, si se configura correctamente, puede ofrecer mejores resultados que Epsilon Greedy al reducir la fase de exploración. Aunque en su contra requiere disponer de un conocimiento previo de los bandidos para poder configurarlo correctamente.

UCB1

Ahora vamos a hablar sobre otro método para resolver el dilema de explotación-exploración. El método se llama UCB1 y pertenece a una familia de métodos que se denominan UCB que son las siglas en ingles de Upper Confidence Bound (Límite de confianza superior). Si pensamos un momento en la secuencia de algoritmos que hemos estado viendo, tiene sentido que los veamos en este orden porque esencialmente, cada una de las ideas toma la idea anterior y la hace un poco más compleja.

Epsilon Greedy fue el primer algoritmo que vimos en la sección, donde tenemos una pequeña probabilidad de exploración aleatoria para que nunca nos quedemos atascados con una estimación inexacta de la media. Valores Iniciales Optimistas lo lleva un paso más allá al usar un tipo de exploración más natural en lugar de solo tener una probabilidad aleatoria de exploración. Simplemente hacemos que la media sea artificialmente alta para que cada bandido sea elegido con más frecuencia hasta que aprendamos que el verdadero promedio no es realmente tan alto. Ahora, daremos un paso más y preguntaremos, ¿hay alguna manera de pensar en ese límite superior de una manera más probabilística? En otras palabras, en lugar de intentar adivinar un buen límite superior, que darle como valor inicial al algoritmo, ¿podemos usar las reglas de probabilidad para inferir ese límite superior?

Aquí vamos a usar la idea de confianza. Nosotros queremos tener confianza en nuestras predicciones y sabemos que para tener confianza en nuestras predicciones, debemos recopilar muchos datos. Cuando recopilemos más y más datos, nuestra confianza en la predicción aumenta cada vez más. Ahora, podemos suponer que esto nos lleva naturalmente a intervalos de confianza, pero vamos a ir en una dirección diferente. Hablaremos de desigualdades. Y más tarde usaremos los intervalos de confianza para desarrollar algo de intuición. Pero por ahora, vamos a trabajar un poco de teoría. La forma básica de una desigualdad se ve así:

$$P(sample\ mean\ -true\ mean \geq error) \leq f(error)$$

Nos dice que la probabilidad de que la diferencia entre la media muestral (estimación) y el verdadero valor de la media sea mayor que algún error, esa probabilidad es menor o igual a alguna función de ese error. Esta función suele ser una función que disminuye su valor a medida que aumenta el valor del error. Pensemos ¿por qué el lado izquierdo tiene sentido? Y la razón es que necesitaríamos recolectar un número infinito de muestras para que nuestra media muestral sea igual a la verdadera media. Si esto no es así, habrá algún error en nuestra medición. Lo que queremos preguntar es qué tan grande es este error. Bueno, podemos expresar esto como una pregunta de probabilidad. Puedo preguntar cuál es la probabilidad de que mi error sea mayor que algún valor y ¿puedo limitar esa probabilidad? Sabemos que el lado derecho es un límite superior porque los dos lados están relacionados por un signo menor o igual. En realidad, en este punto, probablemente sea útil expresar esa probabilidad con una variable real y un ejemplo de alguna función f que pudiéramos utilizar, solo para ejemplificar y poder entender bien lo que estamos diciendo.

Por ejemplo, podemos preguntarnos cuál es la probabilidad de que mi error de medición sea mayor que algún valor t. Y para que quede claro, t debe ser positivo para que podamos decir que la probabilidad de que nuestro error de medición sea mayor que t es menor o igual que uno sobre t. Donde 1/t es solo un ejemplo de una función decreciente. No necesariamente tiene que ser esa.

$$P(sample\ mean\ -true\ mean \geq t) \leq \frac{1}{t}$$

Pero pensemos por qué esto tendría sentido. Digamos que t es un valor muy pequeño, en ese caso, el lado derecho se vuelve más grande. Eso tiene sentido porque la probabilidad de que mi error sea mayor que un valor muy pequeño debería aumentar a medida que ese valor pequeño se hace más pequeño. Por el contrario, digamos que t es un valor muy grande. En este caso también tiene sentido porque la probabilidad de que mi error sea mayor que un valor muy grande debería disminuir a medida que t aumenta. En otras palabras, la probabilidad de ser mayor que un error grande, es menor y la probabilidad de ser mayor que un error pequeño es mayor. Entonces, para resumir, queremos preguntar ¿cuál es la probabilidad de que mi error sea mayor que un valor t? Y resulta que podemos hallar un límite superior para esta probabilidad como una función decreciente de t.

De hecho, cuando tenemos una función en el lado derecho, y esa disminución es proporcional a uno sobre t, esa es la desigualdad de Markov. Hay otra desigualdad llamada Chebyshev, que disminuye proporcionalmente a uno sobre T-cuadrado. Esto se consideraría mejor ya que la función del lado derecho disminuye más rápido, por lo que podemos garantizar que el error está por debajo de un número aún menor. Finalmente, llegamos a la desigualdad de Hoeffding's, que es un límite aún más estrecho, porque disminuye exponencialmente en T-cuadrado. Eso es como una curva gaussiana, que disminuye más rápido que cualquier polinomio. Y esta última desigualdad es la que realmente vamos a usar en el método UCB1.

- Desigualdad de Markov: decrece proporcional a $1/t$.
- Desigualdad de Chebyshev: decrece proporcional a $1/t^2$.
- Desigualdad de Hoeffding's: decrece exponencialmente en t^2.

$$P(\bar{x}_n - \mu \geq t) \leq e^{-2nt^2}$$

El algoritmo en sí es bastante fácil de implementar. Pero antes vamos a tratar de comprender todos los símbolos en esta desigualdad, aunque esto es opcional porque realmente no vamos a hacer uso de eso. Esto es solo para poder tener una mejor comprensión. Entonces, $\bar{x}_n$ es la media muestral de X después de haber recolectado n muestras. μ es la verdadera media, entonces $\bar{x}_n - \mu$ es el error en nuestra medición de la media de X. Y finalmente t es un valor de error arbitrario y n es el numero de muestras. Lo que podemos decir de esto es que a medida que recolectamos más y más muestras, el error se vuelve cada vez más pequeño ya que la función exponencial se hace más pequeña a medida que n aumenta. Ahora bien, hay otros tipos de UCB que son menos populares y, por lo tanto, están fuera del alcance de este curso. La idea es la misma pero usan otra función para calcular el límite superior del error.

Ahora vamos a ver cómo funciona el algoritmo con el pseudocódigo y luego podemos hablar sobre por qué tiene sentido.

- Ciclo:

$$j = \arg\max\left(\bar{x}_{n_j} + \sqrt{2\frac{\log n}{n_j}}\right)$$

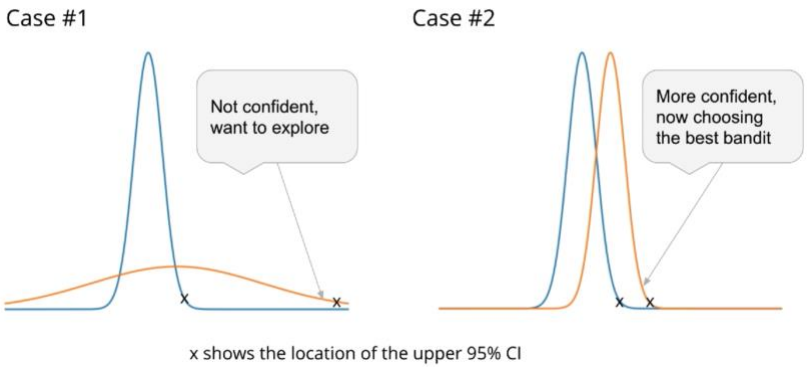
jugar al bandido j, actualizar las recompensas de cada bandido, etc.

- Seguimos siendo codiciosos con respecto al límite superior.

Básicamente, la idea es que este algoritmo va a seguir el mismo patrón básico de los anteriores, pero dentro del ciclo vamos a usar un algoritmo codicioso basado en la suma de estos dos términos. El primer término es la media muestral del bandido j. El segundo término es un límite superior del error. Discutiremos de dónde viene esta fórmula en detalle un poco más adelante. Pero ahora fijémonos en que realmente esto es como el método de valores iniciales optimistas, excepto que en lugar de usar una estimación optimista para la media, usamos la media muestral real, más algún límite superior del error. Como podemos ver, esto sigue siendo un algoritmo codicioso con respecto al límite superior. Entonces, primero, vamos a pensar un momento en la intuición detrás de esto, pero en términos de intervalos de

confianza, aunque sabemos que eso no es exactamente lo que estamos haciendo. Pero vamos a pensarlo desde esa perspectiva.

En primer lugar, ¿funcionaría el uso del límite superior del intervalo de confianza? Como vimos en la descripción del algoritmo, se actúa de manera codiciosa con respecto a este límite superior. Entonces esto sigue siendo similar a cuando usamos valores iniciales optimistas. Entonces, hay dos casos a considerar.



El primer caso es en el que todavía no hemos recogido muchas muestras. Entonces, nuestra estimación es muy mala de momento, y por tanto nuestro intervalo de confianza es grande. En ese caso, lo que quisiéramos es explorar a este bandido para poder recolectar más muestras. Usar el límite superior del intervalo de confianza sería útil ya que como el intervalo es grande, este sería un valor muy alto, y si somos codiciosos escogeremos el bandido con el límite superior más alto, y entonces podremos explorarlo bien. El segundo caso es en el que hemos recolectado muchas muestras, ya tenemos bastante información sobre un bandido en particular. Entonces, en ese bandido, mi estimación es muy precisa y ya no necesito explorarlo más, a no ser que sea el mejor. En ese caso, el algoritmo asegura que la única forma en que puedo seguir explorando a este bandido es si es cierto que es realmente alto y es más alto que los límites superiores de los otros bandidos. Así que usar una especie de límite superior de confianza es razonable.

Ahora, veamos cómo se interpreta exactamente esto en UCB1.

$$j = \arg\max\left(\bar{x}_{n_j} + \sqrt{2\frac{\log n}{n_j}}\right)$$

El primer término es la media muestral, las recompensas promedio. Por supuesto aquí queremos elegir el bandido que tenga la media más alta (la mayor recompensa). Si la media muestral de un bandido en particular es mayor, es más probable que se elija a este bandido, que es lo que queremos. Este sería el aspecto de la explotación. ¿Qué pasa con el segundo término en UCB1? El límite superior del bandido j viene dado por la media muestral del bandido j más la raíz cuadrada de dos veces el logaritmo de n dividido por n_j. Veamos cada uno de los términos dentro de la raíz cuadrada. Es importante diferenciar entre n y n_j. La n a secas representa el número total de jugadas que hemos realizado hasta el momento. Entonces, si jugáramos con tres bandidos una vez que comenzara cada uno, serían tres. Por otro lado, el n_j representa la cantidad de veces que jugamos solo con el bandido j. Entonces, si jugáramos con los tres bandidos una vez cada uno, entonces el n_j sería uno para cada uno de los bandidos. Finalmente, quiero señalar que esto no proviene de ninguna derivación de probabilidad. Sino que es solo un hiper-parámetro, que es lo que define a este tipo de algoritmos como heurística. Si hiciéramos esta constante más grande, entonces le daríamos un límite superior más grande. Y si lo hiciéramos más pequeño, entonces le daríamos un límite superior más pequeño.

La siguiente pregunta a considerar es, ¿por qué funciona esto?

Consideremos lo que le sucede a la n a secas y a la n_j. Digamos que por ejemplo hemos estado ignorando al bandido j=1 durante mucho tiempo, de modo que la n_1 es muy pequeña. Entonces n se hará más grande y eventualmente el numerador será más grande que el denominador. Y esto hará que el límite superior para el bandido 1 sea haga más grande. Esto hace que elijamos en cierto punto al bandido j=1 porque su límite superior es tan grande, que ya es más grande que todas las otras opciones. Por otro lado, supongamos una situación diferente. Imagina que ya hemos explorado el bandido j=1 y ha pasado tiempo, entonces n_1 es grande y n también es grande. Como efecto del logaritmo que se le aplica a n, el numerador crece más lento que el denominador, porque la función logarítmica crece mucho más lentamente que la lineal. Entonces, digamos que jugamos al bandido 1 mil veces y solo jugamos con ese, entonces $n = n_1 = 1000$ y $2 * \log(1000)/1000 = 0.01$ aproximadamente.

Entonces, como podemos ver, a medida que estos valores del tamaño muestral (el numero de jugadas) crecen, el denominador supera al numerador y el límite superior se reduce a cero, y todo lo que nos queda son las medias muestrales, que es lo que deseamos. Eso tiene sentido, porque después de recopilar una gran cantidad de datos, las estimaciones ya son suficientemente precisas y no es necesario explorar más. En general, el límite de log(n)/n cuando n tiene a infinito es cero.

Ahora, aunque la matemática detrás de todo esto realmente es opcional sí quiero mencionar ¿de dónde viene este límite superior? Intuitivamente, podemos tomar el lado derecho de la desigualdad de Hoeffding's, y darnos cuenta de que si tomamos el logaritmo en ambos lados y despejamos t, se obtiene que t es igual a:

$$P(\bar{x}_n - \mu \geq t) \leq e^{-2nt^2}$$

$$p = e^{-2nt^2}$$

$$t = \sqrt{-\frac{\log p}{2n_j}}$$

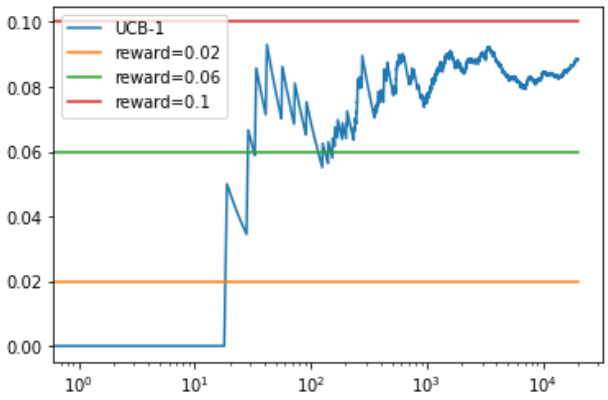
UCB1 simplemente dice, como heurística, que tomemos p igual a $p = n^{-4}$. En otras palabras, queremos que la probabilidad del límite superior disminuya como uno dividido por n a la 4. Si sustituimos esto ya vemos de dónde ha salido el límite que hemos puesto en el algoritmo:

$$t = \sqrt{-\frac{\log n^{-4}}{2n_j}} = \sqrt{\frac{2\log n}{n_j}}$$

En resumen, la idea básica del método UCB1, es estimar un límite de confianza superior para determinar cuál es la solución óptima. Para ello se usa la desigualdad de Hoeffding. Una desigualdad que se puede aplicar a cualquier distribución. La desigualdad de Hoeffding indica la probabilidad de que la media muestral se desvíe de la verdadera media. Con lo que es posible estimar el intervalo de confianza que podíamos esperar. Un método heurístico para la determinación de la probabilidad es que esta se reduzca con el tiempo. Que al principio sea menos estricto, pudiendo explorar más, pero a medida que aumenta el número de datos ser más estrictos. En el caso de UCB1 se fija el criterio como n^{-4}. Ahora, el algoritmo UCB1 seleccionará en cada momento el bandido cuya recompensa esperada junto con su intervalo de confianza devuelva el valor más elevado.

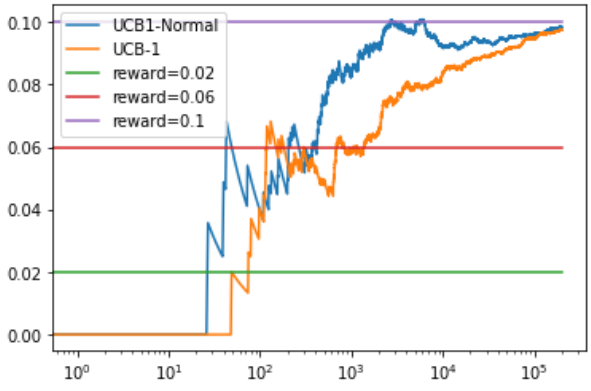
BANDIDO MULTIBRAZO: UCB1

Ahora que conocemos la fórmula con la que se selecciona el bandido mediante UCB1 se puede implementar en Python. Es importante destacar que cuando se observa un empate entre bandidos, al igual que en métodos anteriores se elige el bandido con el que se va a jugar de forma aleatoria.



En los resultados vemos que UCB1 se decanta bastante rápido por el bandido óptimo. Aunque en este caso, debido a que los valores de las recompensas están bastante cerca, realiza muchas jugadas con el bandido que no es óptimo. Lo que produce que la recompensa promedio que obtiene no esté tan cercana al valor de recompensa real del bandido óptimo. En resumen, UCB1 es un algoritmo que se basa en la desigualdad de Hoeffding para obtener un intervalo de confianza para la recompensa de cada uno de los bandidos, es decir, selecciona el bandido cuya recompensa observada más el limite superior del intervalo de confianza sea máximo.

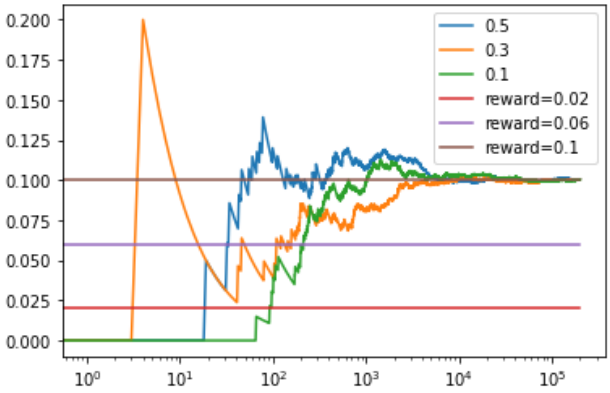
Una modificación del UCB1 es el UCB1-Normal. El método UCB1-Normal es una mejora de UCB1 en el que se asume que las recompensas de cada uno de los bandidos son generadas mediante una distribución Normal. Lo que permite seleccionar al mejor bandido en base al nivel de confianza que se deduce de esta distribución. Minimizando de este modo el número de tiradas en fase de exploración. Una de las características de este UCB1-Normal es el hecho de garantizar un mínimo de jugadas para cada uno de los bandidos. Evitando de este modo que algunos de los bandidos no se exploren debido a una mala racha inicial. Aunque esto puede conllevar una menor explotación de los resultados.



En esta ocasión se puede ver que UCB1-Normal selecciona más rápidamente el mejor bandido en comparación con UCB1. Aunque, como en otros casos, a largo plazo los resultados de la recompensa media obtenida es similar a la de UCB1. La mejora también se puede apreciar comparando el número de veces que el UCB1-Normal ha jugado con cada bandido respecto a UCB1. En el caso de UCB1-Normal solamente ha seleccionado 494 veces el 1ro frente a las 2.883 de UCB1, por lo que ha jugado con este bandido no óptimo solamente un quinto de las veces. Por otro lado, para el segundo bandido sería 5.854 en el caso de UCB1-Normal frente a las 8.807 de UCB1. Lo que indica que UCB1-Normal ha seleccionado el bandido óptimo en más ocasiones que UCB1.

BANDIDO MULTIBRAZO: UCB2

Hemos visto cómo aplicar el UCB1 y el UCB1-Normal al problema del bandido multibrazo. En ambos la idea es estimar un límite de confianza superior para la recompensa de cada uno de los bandidos, seleccionando en cada momento el que tenga la recompensa media más el límite de confianza mayor. En esta ocasión vamos a ver cómo se puede implementar el método UCB2 para un problema Bandido Multibrazo. El método UCB2 es una mejora de UCB1 en la que se reduce el número de veces en las que se selecciona un bandido que no sea el óptimo. Reduciendo de este modo el número de tiradas en el modo exploración. Aunque esto se hace a costa de un algoritmo algo más complejo. El límite va a ser diferente, porque en UCB2 la fase de exploración se divide en trozos de tamaño variable en las que se juega con cada uno de los bandidos una cierta cantidad de veces. Además vamos a contar con un hiper-parámetro nuevo llamado alfa: *α,* que influye en el ratio de aprendizaje del algoritmo. Consideremos tres valores distintos para *α=0.1, 0.3 y 0.5.*



En esta ocasión lo primero que se puede apreciar es cómo la recompensa promedio después de 10000 tiradas es prácticamente la óptima en todos los casos. Aunque, a medida que se aumenta el valor del parámetro *α*, vemos que el agente se decanta más rápidamente por la solución óptima. Lo que se puede ver en el número de veces totales que el agente ha jugado con soluciones sub-óptimas.

En el caso de *α* igual a 0.5, el agente ha seleccionado 16 veces el primer bandido y 24 el segundo, por lo que ha jugado 199960 veces de 200000 con el bandido óptimo. Por otro lado, cuando *α* es 0.1 el agente ha jugado 66 veces con el primer y segundo bandido, jugando 199868 con el óptimo. Números que explican claramente los resultados vistos en la figura.

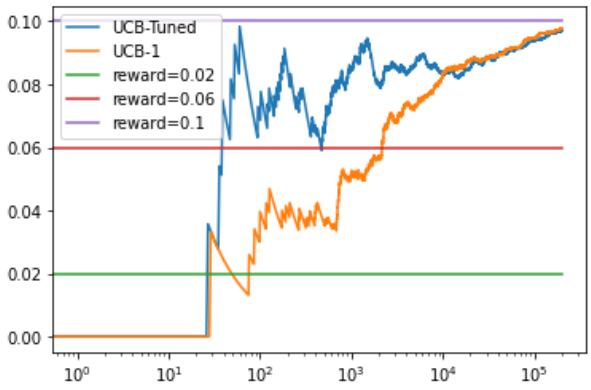
UCB2 es un método que demuestra ser el más eficiente de los vistos hasta el momento, aunque esto sea a costa de una mayor complejidad que UCB1.

BANDIDO MULTIBRAZO: UCB1-TUNED

Anteriormente hemos visto UCB2, un algoritmo que ha ofrecido mejores rendimientos que UCB1 para nuestros bandidos basados en una distribución binomial. En esta ocasión vamos a ver UCB1-Tuned (también conocido como UCB-Tuned), una mejora de UCB1 en el que se modifica la fórmula con la que se calcula el límite de confianza superior.

El método UCB1-Tuned propone una mejora de la expresión usada en UCB1 para determinar el límite de confianza superior para cada uno de los bandidos, modificando la regla de selección. En concreto, incorpora información sobre el promedio de los cuadrados de las recompensas. El algoritmo UCB1-Tuned

suele superar empíricamente a UCB1 en cuanto a la veces que selecciona el mejor bandido. Además de ser un método menos sensible a la variabilidad de las recompensas.



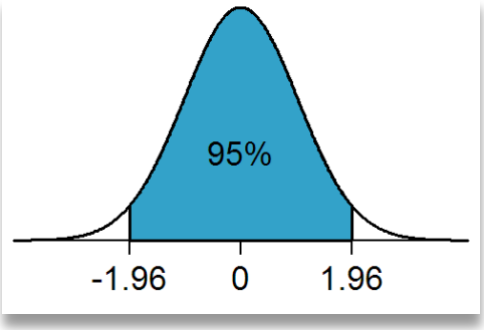
Lo que se puede ver en los resultados del ejemplo es que el rendimiento de ambos métodos es similar a largo plazo, convergiendo ambos hacia valores muy próximos al valor teórico del mejor bandido. Aunque el método UCB1-Tuned ha detectado el bandido óptimo más rápido que UCB1. Por lo que demuestra ser un método más adecuado cuando el número de tiradas disponibles es reducido.

PRIOR CONJUGADO

Ahora antes de ver un nuevo algoritmo que hace uso del enfoque bayesiano, llamado Bayesian Bandit (Bandido bayesiano) o Thompson Sampling (Muestreo de Thompson), vamos a hablar de un concepto clave que usaremos en este enfoque. Primero vamos a ver cómo los intervalo de confianza son una herramienta muy intuitiva para poder entender la idea detrás de todo esto, porque aunque no los vamos a utilizar realmente, nos proporcionan una buena imagen de lo que sucede por detrás. Y de alguna manera ejemplifica el dilema de exploración / explotación.

Cuando tenemos solo una pequeña cantidad de datos, el intervalo de confianza es grande, perdemos en precisión y hay más posibilidades. Cuando tenemos una gran cantidad de datos, ganamos en precisión y entonces el intervalo de confianza se hace más pequeño. Cuando el intervalo de confianza es grande, eso significa que deberíamos explorar más (recopilar más datos, más información) y cuando el intervalo de confianza es estrecho, eso significa que deberíamos explorar menos (porque tenemos suficiente ya). Y la explotación vendría cuando el intervalo de confianza sea estrecho y tenga un valor elevado. Lo que equivale a decir que tenemos una buena precisión y confianza en que la recompensa de ese bandido es alta y por eso queremos explotarlo. Entonces, lo bueno del intervalo de confianza es que nos dice de una manera indirecta dónde podría estar el verdadero valor de la recompensa.

Pero hagámonos una pregunta, ¿por qué el intervalo de confianza es simétrico, por qué siempre tiene esta curva en forma de campana?



De hecho, eso se debe a que los intervalos de confianza se basan en el Teorema Central del Límite, que dice que las sumas de variables aleatorias tienden a una distribución normal, dando igual la distribución original de los datos, mientras se tenga una muestra suficientemente grande. Dado que vamos a hablar de un método del enfoque bayesiano, no vamos a utilizar intervalos de confianza porque pertenecen al enfoque frecuentista clásico. Sin embargo, el concepto de intervalos de confianza nos da bastante intuición. Lo que realmente queremos cuando miramos un gráfico como este es la distribución del parámetro desconocido, en nuestro ejemplo, el parámetro que desconocemos es la media de las recompensas. El enfoque bayesiano nos da las herramientas que necesitamos para calcular esto, porque en el paradigma bayesiano, todo es una variable aleatoria, incluso los parámetros de las distribuciones. Eso significa que si antes dijimos que la recompensa venía de alguna distribución con alguna media, ahora vemos que la media también tiene una distribución, ya que todo es una variable aleatoria, todo tiene una distribución.

Para determinar la distribución de la media, vamos a llamarla parámetro θ, y vamos a estructurar el problema en términos de un problema de estadística bayesiana. La cantidad que estamos buscando es $P(\theta|X)$, donde θ es la media de la variable aleatoria X original con distribución Bernoulli y la X dentro de esta expresión se refiere a los datos que hemos recopilado. Por conveniencia, combinaremos los dos, pero nos daremos cuenta de que son símbolos sobrecargados. Cuando utilizamos X mayúscula, podríamos estar refiriéndonos a la variable aleatoria o podríamos estar refiriéndonos a los datos que hemos recopilado. Así que hay que prestar mucha atención al contexto y así no tendremos nunca ningún problema. Por supuesto, el lugar desde el que partimos en el enfoque bayesiano es el Teorema de Bayes.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

En el lado izquierdo, tenemos $P(\theta|X)$, que es una distribución que llamamos posterior (o a posteriori). A la derecha, tenemos tres cosas. La primera, $P(X|\theta)$, es la verosimilitud (likelihood), la misma probabilidad que usaría en una estimación de máxima verosimilitud. Podríamos interpretar eso como la probabilidad de los datos dado el parámetro θ. Luego, $P(\theta)$ en sí mismo se llama prior (o distribución a priori), es la distribución de θ cuando no sabemos nada sobre X o, en otras palabras, cuando no hemos recopilado ningún dato. Y $P(X)$ en el denominador se llama evidencia. Como estamos tratando de encontrar $P(\theta|X)$, que es una distribución sobre la variable aleatoria θ, resulta que $P(X)$ es constante con respecto a θ y, por lo tanto, podemos escribir esto como una proporcionalidad en lugar de una ecuación.

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

Ahora, ¿por qué querríamos deshacernos de la evidencia y convertir esto en una proporcionalidad? Bueno, en general, no conocemos este denominador y es intratable o a veces incluso imposible de calcular.

Como recordarán, podemos expresar el denominador como una integral sobre el numerador, ya que ambos expresan la probabilidad conjunta.

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int P(X|\theta)P(\theta)d\theta}$$

Dado que θ es una variable aleatoria continua, integramos en lugar de sumar. El problema es que, si recuerdas de tus estudios de cálculo, la integración es muy difícil. No se trata solo de aplicar mecánicamente un conjunto de reglas como la diferenciación. Entonces lo que se suele hacer es usar métodos de aproximación como el de Monte Carlo del que hablaremos más en la siguiente sección, que calculen el valor aproximado de la integral. Sin embargo, como sabemos, si esto está diseñado para ejecutarse en tiempo real, tampoco es factible ejecutar una simulación de Monte Carlo para obtener una respuesta.

El truco de la Estadística Bayesiana es que hay pares especiales de distribuciones en las que podemos aprovechar la proporcionalidad e ignorar la evidencia. Estas se denominan **Conjugate pairs o Pares conjugados**. Esto significa que para un par de distribuciones, una distribución de muestreo y una

distribución a priori, la distribución a posteriori resultante pertenece a la misma familia paramétrica de distribuciones que la distribución a priori. Y en este contexto, el prior se llamaría **Prior conjugado (Conjugate prior).**

La historia es la siguiente. En probabilidad, usamos casi siempre un conjunto fijo de distribuciones comunes. Tenemos la Gaussiana, la Bernoulli, la Binomial, la Poisson y así sucesivamente. En general, la distribución a posteriori no encaja bien en ninguna de estas distribuciones comunes. Por ejemplo, no puedo decir que si mi verosimilitud es Normal, y mi prior es uniforme, entonces la posterior es Normal. Eso no funcionará. Sin embargo, los pares conjugados son especiales. Son especiales porque cuando se combinan la verosimilitud y la prior, y se obtiene la posterior, esta sí proviene del mismo tipo de distribución que tenía la prior.

Entonces, si por ejemplo mi prior es Gaussiano, entonces la posterior será también será Gaussiana. Pero esto depende del hecho de que la verosimilitud coincida con la prior de manera que sean conjugadas.

$$Normal \propto Normal \times Normal$$

Por lo tanto, no podemos elegir cualquier verosimilitud, esto solo funciona para ciertas distribuciones. Otro ejemplo que veremos es que si los datos son Binomiales y elegimos una prior con distribución Beta, la posteriori tendrá distribución Beta también.

$$Beta \propto Binomial \times Beta$$

Entonces, ¿en qué otras distribuciones estamos interesados? Como sabemos, en nuestro ejemplo, nuestra verosimilitud es una Bernoulli. Por ejemplo en un lanzamiento de una moneda, donde tenemos estos resultados de los lanzamientos $X = \{x_1, x_2, \dots, x_n\}$, imagina que nos interesa saber la probabilidad de sacar cara, entonces la Bernoulli tomaría valor 1 si sale cara y 0 si sale cruz. Sabemos que podemos expresar la probabilidad como el producto de las probabilidades de cada lanzamiento:

$$P(X|\theta) = \prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}, \text{ donde } x_i \sim Bernoulli(\theta)$$

En este caso, estamos usando θ para representar el parámetro de la Bernoulli que es la probabilidad de éxito, sea "éxito" lo que nos interese, por ejemplo que salga cara.

Bueno, resulta que si elegimos nuestro prior para que sea la distribución Beta, esto es un par conjugado para la verosimilitud de la Bernoulli.

$$P(\theta|X) \propto P(X|\theta)P(\theta) = \left(\prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

Vamos a demostrar que ese es el caso, para que no parezca una elección completamente arbitraria. Pero veamos primero de manera intuitiva por qué tiene sentido. Sabemos que en una situación de ganar o perder (1 y 0) donde la recompensa es binaria, la media de esa distribución debe ser un número entre cero y uno. Resulta que el soporte de la distribución Beta está en el rango de cero a uno. Así que por esa parte cumple con el requisito. En la expresión de la distribución Beta, tenemos dos parámetros $\alpha \; y \; \beta$, que son los parámetros de esta distribución. La función $B(\alpha, \beta)$ que aparece en la fórmula, es igual a esta expresión:

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha,\beta)}$$

Gamma (Γ) en sí es una función especial que resulta ser una integral elegante. Pero básicamente puedes pensar en ello como una generalización de la función factorial para números no enteros. En realidad, la función Γ se define tanto para números reales como complejos, aunque eso estaría fuera del alcance de

este curso. En cualquier caso, en realidad nunca tenemos que preocuparnos por calcular la función $B(\alpha, \beta)$.

Entonces, ¿por qué nos gustan las expresiones como esta en el aprendizaje automático bayesiano? Bueno, si miramos de cerca, vemos que tanto θ como $1 - \theta$ aparecen en la base tanto de la distribución a priori como de la función de verosimilitud. Además, dado que la función $B(\alpha, \beta)$ es constante con respecto a θ, se puede eliminar este factor de la proporcionalidad.

Como estos son productos sobre la misma base, con exponente diferente, es lo mismo que poner que es igual a la base elevada a una suma de los exponentes. Y ahora podríamos combinar todos los términos θ y todos los términos $1 - \theta$, para obtener un solo término θ y un solo término $1 - \theta$. Pero aquí está el truco, como podemos ver, la posteriori final tiene exactamente la misma forma que una distribución Beta, que, como recordaremos, ¡es el mismo tipo de distribución que elegimos para la prior!

$$P(\theta|X) \propto \left(\prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}\right)\left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \left(\prod_{i=1}^{n} \theta_i^x (1-\theta)^{1-x_i}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$= \left(\theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{\sum_{i=1}^{n}(1-x_i)}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$= \left(\theta^{\alpha-1+\sum_{i=1}^{n} x_i}(1-\theta)^{\beta-1+\sum_{i=1}^{n}(1-x_i)}\right)$$

$$= \left(\theta^{(\alpha+\sum_{i=1}^{n} x_i)-1}(1-\theta)^{(\beta+n-\sum_{i=1}^{n} x_i)-1}\right) = Beta\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$$

Esto se debe a que la distribución beta tiene esta forma general como distribución. En conclusión, podemos decir que la distribución posteriori para θ, dado un conjunto de datos X es una distribución Beta con los siguientes parámetros:

$$Beta\left(\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i\right)$$

Esto depende del hecho de que la prior es una distribución $Beta(\alpha, \beta)$, y que la probabilidad es Bernoulli.

Una parte del proceso que no podemos olvidar en un aprendizaje automático bayesiano es cómo elegir el valor de α, β en primer lugar. Esto define la forma de nuestra distribución prior. Tenemos la suerte de que la distribución Beta es lo suficientemente expresiva como para darnos algo que tenga mucho sentido, como por ejemplo la distribución uniforme. Y es que la distribución $Beta(1,1)$ resulta ser una distribución $Uniforme[0,1]$ y esto suele ser una buena elección para la prior.

Lo que viene a decir esta elección es que no tenemos ni idea de cuál puede ser el resultado, por lo que asignamos la misma probabilidad a cada valor posible. Sin embargo, si nuestro problema en concreto proviene de una industria en la que realmente sí se tienen conocimientos previos sobre cuál podría ser esta probabilidad a priori, entonces debería hacer uso de ese conocimiento. Por ejemplo, sabemos que las tasas de clics en los anuncios suelen ser muy pequeñas, alrededor del 1% o 2%, entonces, si tenemos ese conocimiento debido a la experiencia en el dominio, esto debería estar codificado en la prior.

Un punto de confusión es el hecho de cómo actualizar la distribución a posteriori cuando se trata del problema real del bandido multibrazo. Recordemos que en el problema del bandido, actualizamos nuestro modelo después de recolectar cada muestra de X. Entonces, a diferencia de nuestro cálculo teórico, nunca tenemos n muestras, sino que tendremos una sola muestra. Por supuesto, en el transcurso de un experimento, sí podemos llegar a recolectar muchas muestras.

El truco en el aprendizaje automático bayesiano, es que vamos a considerar que la distribución a posteriori en un paso se convierte en la distribución a priori en el siguiente paso. Lo cual tiene sentido porque es una manera de actualizar nuestra información y tomarla como evidencia para seguir avanzando.

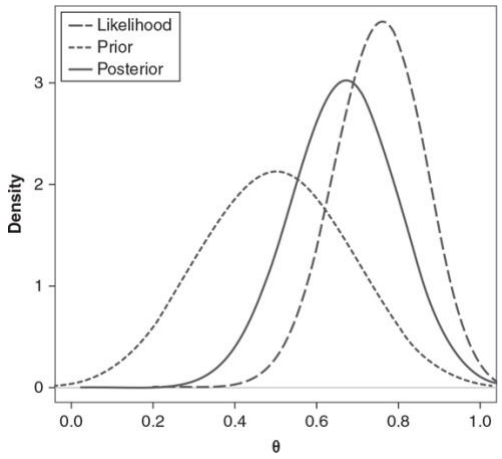
Vamos a ver un ejemplo de cómo podemos hacer esta actualización en varios pasos de tiempo. Empezamos con un prior Beta(1,1), la distribución Uniforme. A continuación, tiramos del brazo del bandido y obtenemos por ejemplo una ganancia, x=1. La posteriori por lo tanto y por la fórmula que vimos antes sería Beta(1+1,1+1-1) que es una Beta (2,1). En el próximo paso, la Beta(2,1) se convierte en prior. Ahora digamos que sacamos un uno nuevamente, entonces x=1 y nuestro nuevo posterior es Beta(2+1,1+1-1) que es una Beta(3,1). Así que este es nuestro nuevo posterior. Pero en el siguiente paso, la Beta(3,1) se convierte en la prior. Entonces, digamos que en el siguiente paso perdemos, obtenemos x=0, entonces nuestro posterior ahora es Beta(3+0,1+1-0) = Beta(3,2).

En resumen, tengamos en cuenta que cuando hablamos de teoría, en estadísticas bayesianas y aprendizaje automático bayesiano, generalmente definimos la posteriori en términos de varias muestras pero en aplicaciones con actualización en línea (online applications) por lo general, calculamos la posteriori en términos de una sola muestra.

BANDIDO BAYESIANO O MUESTREO DE THOMPSON

Anteriormente, cubrimos todos los fundamentos, como los prior conjugados y cómo actualizar el prior para obtener el posterior después de recopilar nuevos datos. Ahora vamos a discutir cómo usar esto realmente en un algoritmo. Volveremos nuevamente a nuestra imagen intuitiva de los intervalos de confianza, excepto que ahora ya no vamos a usar los intervalos de confianza.

En cambio, ahora que sabemos sobre las posteriors en el paradigma bayesiano, podemos dibujar las distribuciones a posteriori exactas y llegar a las mismas conclusiones.

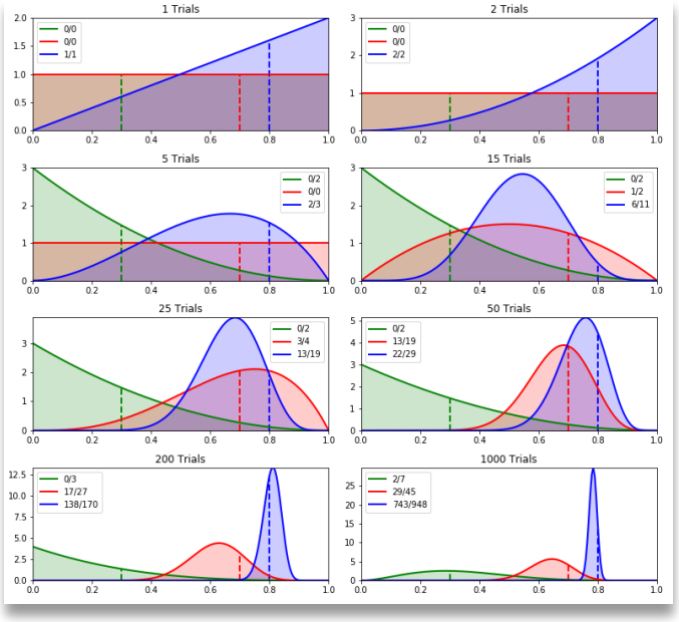


Como sabemos, la posteriori representa nuestra creencia sobre la distribución del parámetro θ, que es la recompensa media de un bandido. Teóricamente, θ podría estar en cualquier lugar de este rango (entre 0 y 1 si lo expresamos en porcentajes), aunque obviamente las áreas con más masa de probabilidad son más probables. El método Bandido Bayesiano o Bayesian Bandit implica una estrategia diferente a la que hemos usamos anteriormente, que consistía básicamente en utilizar un límite superior.

En cambio, lo que vamos a hacer ahora es simplemente extraer una muestra de esta distribución. Este método se conoce también con el nombre de Muestreo de Thompson o Thompson Sampling. La parte de la palabra “muestreo”, en realidad tiene mucho sentido si lo pensamos detenidamente. ¿Qué es una distribución en primer lugar? Una distribución nos indica qué valores son los valores más probables que una variable aleatoria puede tomar. Al extraer muestras de esta distribución, estamos diciendo dame un

valor que caracterice a esta distribución y dejemos que eso determine qué bandido elegimos. Y en lugar de usar solo un valor, que es el límite superior, podemos usar todos sus valores generando esas muestras.

A medida que extraemos más y más muestras, esta distribución se volverá más y más delgada a medida que tengamos más y más confianza en nuestra creencia de dónde se encuentra la verdadera media.



Cuando extraemos una muestra de una distribución muy delgada, obtendremos un valor muy cercano al centro y será como si estuviéramos comparando a cada bandido por sus medias reales. Si no comprendes esto de inmediato, veremos la misma idea unas cuantas veces más desde una variedad de perspectivas. Solo recuerda el hecho principal, que es que en lugar de usar algún tipo de límite superior como antes, elegimos aleatoriamente cualquier valor que la media pueda tomar de acuerdo con la distribución a posteriori.

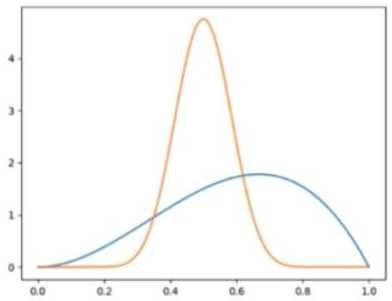
Ahora veamos el pseudocódigo que es otra perspectiva que creo que puede aclarar estas ideas.

```
class Bandit:

        def sample():

                return beta(a,b) . sample()

        def update(x):

                a = … , b = …

for n in range (NUM_TRIALS):

        j = argmax(b.sample() for b in bandits)

        x = bandit[j].pull

        bandit[j].update(x)
```

Dentro de la clase bandido, definimos una nueva función llamada sample, que nos devuelve una muestra de la distribución Beta dada por sus valores actuales de a y b. Aquí usamos a y b en representación de α, β.

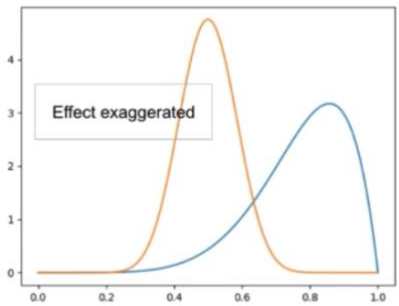
Como podemos ver, cuando elegimos nuestro bandido, tomamos el argmax (argumento máximo) con respecto a la muestra. Esto se opone a nuestros algoritmos anteriores cuando tomamos el argmax con respecto a un límite superior. Sería como decir que queremos ser codiciosos, pero con respecto a las muestras de las distribuciones a posteriori directamente, en lugar de alguna estadística estimada de las muestras.

Lo siguiente que creo que será útil es mirar los gráficos de las posterioris y ver cómo cambian a medida que recopilemos nuestras muestras. Comencemos observando un escenario común en el que tenemos una posteriori delgada y una posteriori más gordita.



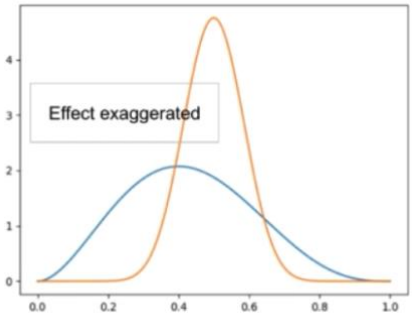
- Con la posteriori delgada, cuando tomamos muestras de esta distribución, el rango de valores que podemos obtener es bastante estrecho.
- Con la distribución más gordita, cuando tomamos muestras, potencialmente podemos obtener muchos valores diferentes porque el rango es más amplio.

Como podemos ver, existe alguna posibilidad de que la muestra de la distribución más gordita sea mayor que la muestra de la distribución más delgada. Pero también hay una posibilidad de que la muestra de la distribución más gordita sea menor que la muestra de la distribución delgada. Supongamos que extraemos una muestra de cada distribución y que los valores de la muestra de la distribución más gordita resultan ser más grandes. Eso significa que vamos a tirar de ese bandido. Digamos que jugamos con ese bandido y ganamos, entonces obtenemos un nuevo dato x=1. Luego, cuando actualizamos nuestra distribución más gordita (la azul), suceden dos cosas:



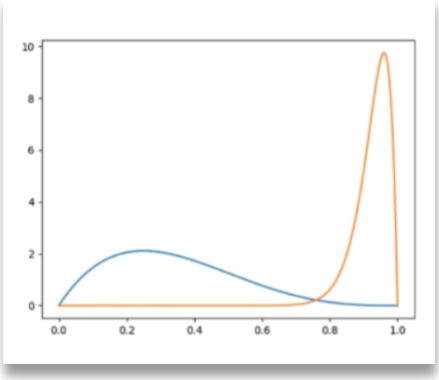
Como vemos en la figura anterior, la distribución de color azul se vuelve un poco más delgada porque tenemos más información y por tanto más precisión en nuestra estimación. Además, su pico se mueve un poco hacia arriba ya que obtener un dato nuevo igual a 1 hace que aumente la media.

A continuación, consideremos la situación en la que tomamos una muestra de la distribución más gordita y resulta ser más grande, pero cuando tiramos del brazo, obtenemos x=0. En este caso, ocurren dos cosas:



Primero, todavía se vuelve un poco más delgada porque ganemos o perdamos, aún obtenemos una muestra y, por lo tanto, tenemos más confianza en la estimación. Pero ahora el pico no se mueve hacia la derecha en los valores horizontales, sino que se reduce y se mueve hacia la izquierda, ya que un cero disminuiría la recompensa media.

Ahora por ejemplo consideremos otro escenario en el que nuevamente tenemos un bandido con una distribución gordita y un bandido flaco, o con distribución delgada. Pero esta vez la distribución más ancha (la azul) es tal que la mayor parte de la masa de probabilidad es menor que donde está la mayor parte de la distribución más estrecha.



¿Qué sucede cuando sacamos una muestra? Bueno, para el bandido gordito, en realidad, la mayoría de esas muestras serán menores que la muestra del bandido flaco y, por lo tanto, no terminaremos tocando ese bandido muy a menudo. Y eso es bueno. ¿Por qué eso es bueno? Hay dos cosas a las que prestar atención. Primero, su media, y segundo, su gordura. Como podemos ver, la media es baja (el pico de la azul está hacia los valores de la izquierda que son bajos, en comparación con los valores del pico de la naranja). Por lo tanto, no queremos jugar con este bandido porque es sub-óptimo y no da una recompensa tan a menudo como el bandido más delgado (naranja). Además, como el azul es gordo, o amplio, eso significa que tenemos poca confianza en nuestra estimación, lo que también significa que no hemos recopilado muchos datos sobre este bandido. Y eso es bueno. ¿Por qué eso es bueno? Nuevamente, porque es un bandido sub-óptimo, por lo que en realidad no queremos jugar con él. Entonces, en esta situación, el Bandido Bayesiano o Muestreo de Thompson nos protege de jugar con ese tipo de bandidos que en realidad queremos evitar porque no son óptimos. Si comparamos esto con Epsilon Greedy, seguramente seguiríamos jugando con el bandido malo de todos modos, ya que nuestra elección sería uniforme, y aleatoria.

Entonces, vamos a resumir lo que hemos visto. El Muestreo de Thomson, si bien creo que es muy intuitivo, también tiene la mayor parte de teoría de todos los algoritmos de bandido que hemos visto. Comenzamos con la idea intuitiva de que estamos interesados en algo como el intervalo de confianza de nuestra estimación de la tasa de ganancia. Esto encapsula perfectamente el dilema de exploración y explotación. Cuando el intervalo de confianza es amplio, queremos explorar para mejorar nuestro intervalo de confianza, y cuando los intervalos de confianza son estrechos, queremos aprovechar este hecho eligiendo

el bandido con la media más alta, para explotarlo. Pero los intervalos de confianza se basan en el Teorema Central del Límite y no describen con precisión la verdadera distribución de las tasas de recompensas de los bandidos. En cambio, el método bayesiano consiste en tratar las tasas de recompensas como una variable aleatoria, es decir, darle su propia distribución. En general, calcular una posteriori a partir de la regla de Bayes no es fácil porque usualmente involucran sumas intratables o integrales insolubles. Por ello usamos los priors conjugados para en lugar de calcular la posterior a mano, usar la proporcionalidad para demostrar que la forma de la posteriori se ajusta a una distribución particular y a partir de ahí se puede encontrar la constante de normalización, ya que la integral de cualquier distribución debe ser igual a uno. Luego, presentamos el algoritmo de Muestreo de Thompson, que nos dice que para elegir un bandido, simplemente ordenamos los bandidos basándonos en las muestras extraídas de sus posterioris. Vimos qué pasaba con estas distribuciones en cada paso. Cuando las distribuciones son anchas, exploramos más. Y cuando las distribuciones son delgadas, exploramos menos. Pero lo más importante, cuando la distribución óptima se vuelve delgada, podemos dejar las distribuciones sub-óptimas aunque sean anchas sin necesidad de explorarlas y perder optimalidad, ya que en este escenario el bandido óptimo suele tener los valores muestra más altos. Esto nos lleva a elegir mayormente el bandido óptimo, que es realmente como queremos que sea la explotación.

Una última pregunta que quiero responder es: de todos estos métodos que discutimos, ¿cuál deberíamos usar en la práctica? Todos parecen igualmente fáciles de implementar y solo los algoritmos y el razonamiento detrás de ellos son diferentes. Todo se reduce a esta regla. El aprendizaje automático es experimentación, no filosofía. Para elegir el mejor algoritmo, no hay una regla, sino que se debe experimentar y recopilar datos para tener una respuesta definitiva para nuestro conjunto de datos en particular. Porque un algoritmo funcione muy bien en un escenario en particular no quiere decir que también funcionará mejor para nuestro caso. Y tengamos en cuenta que hemos estado haciendo una simulación muy simple, pero los problemas del mundo real pueden tener cientos o miles de opciones. Y cuando aumentamos la dimensionalidad del problema, eso posiblemente podría cambiar las características de estos algoritmos. Por otro lado, debido a que los algoritmos de bandido son tan simples de implementar y tan populares, podemos aprovechar las experiencias de muchas otras empresas de la misma industria. Por tanto, es bueno investigar las experiencias de los demás antes de crearnos una idea sobre el algoritmo que deberíamos utilizar. En la industria de los ejemplos que solemos poner, de marketing online, el Muestreo de Thompson parece ser el ganador en muchos experimentos realizados en esta industria. Entonces, es por eso que hemos estado construyendo los algoritmos desde el más básico para eventualmente llegar a este punto. Sin embargo en aprendizaje reforzado (reinforcement learning), los algoritmos usualmente se usan en el contexto de Epsilon Greedy para equilibrar la exploración y explotación y compensar las dos a la vez.

UCB-BAYES

El algoritmo de Bandido Bayesiano o Muestreo de Thompson se basa en una idea del enfoque bayesiano con el que es posible obtener buenos resultados en problemas de tipo Bandido Multibrazo. Es uno de los algoritmos más antiguos que se utilizan para seleccionar los bandidos en este tipo de problemas. Se le llama “muestreo” porque para seleccionar el bandido toma una muestra aleatoria de distribución de probabilidad asociadas a cada uno de los bandidos. Su distribución a posteriori para ser más exactos. En el caso de que la recompensa se distribuya en base a una distribución Binomial se puede utilizar la distribución Beta que es su conjugada. Así, a cada uno de los bandidos se le asocia una distribución Beta donde los parámetros se obtienen mediante:

$$Prior \sim Beta(\alpha, \beta)$$

$$Posteriori \sim Beta(\alpha', \beta')$$

$$\alpha' = \alpha + \sum_i x_i$$

$$\beta' = \beta + N - \sum_i x_i$$

donde x_i son las recompensas obtenidas por el bandido en cada una de las tiradas y N es el número de sucesos de la distribución Binomial, o la cantidad de jugadas que tiene ese bandido. De modo que para seleccionar el bandido en cada una de las tiradas se obtendría una muestra de cada una de las distribuciones y se selecciona aquella con mayor probabilidad.

El Muestreo de Thompson no crea un mecanismo para resolver el problema, sino que asocia una distribución de probabilidad a cada uno de los bandidos. A medida que aumenta el número de veces que se ha jugado con cada uno de los bandidos, la forma de las distribuciones se irá haciendo más estrechas en torno a la media. Por lo que, al obtener una muestra de cada una de las distribuciones es más probable que esta se encuentre cerca del valor esperado para ese bandido.

Anteriormente se ha asumido que las recompensas se obtienen mediante una distribución Binomial, lo que nos permite justificar el uso de la distribución Beta. En el caso de que no se conozca la distribución de las recompensas se puede emplear la distribución $Normal(\mu_j, \sigma_j)$ para realizar el muestreo y seleccionar el bandido. En este caso los parámetros se pueden obtener a través de las siguientes expresiones:

$$\mu_j = \widehat{x_j} \qquad \sigma_j = \frac{1}{n_j + 1}$$

donde $\widehat{x_j}$ es la media observada en el bandido j y n_j son las veces que se ha jugado con el bandido j.

En la implementación del UCB1 no se ha asumido que las recompensa provengan de ninguna distribución. Por lo que se usó la desigualdad de Hoeffding para obtener una estimación generalizada del intervalo de confianza. En el caso de conocer cuál es la distribución se podría obtener el intervalo de confianza de una manera más precisa. Así, basándonos en lo que hemos visto para el Muestreo de Thompson, se puede asumir que una distribución Beta para las recompensas Binomiales y emplear como criterio de confianza una cantidad de desviaciones estándar para seleccionar el bandido. Es decir se puede usar esta expresión para obtener UCB un Bayesiano:

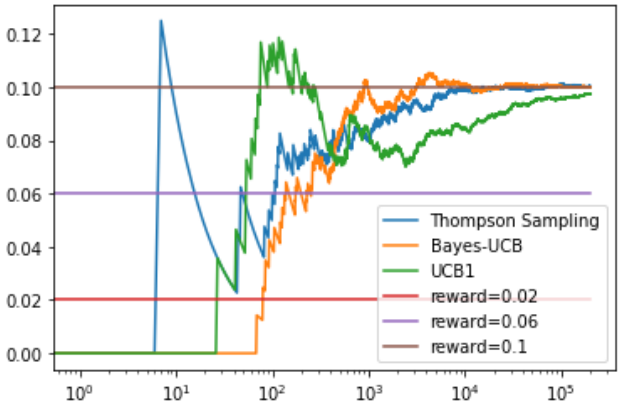
$$X_{Bayes-UCB} = \overline{x_j} + \gamma B_{std}(\alpha,\ \beta)$$

- α y β se calcula tal como se ha explicado anteriormente.
- $\overline{x_j} = \frac{\alpha_j}{\alpha_j + \beta_j}$ es la media de la distribución Beta.
- γ es un hiper-parámetro con el que se indica cuántas desviaciones estándar queremos para el nivel de confianza.
- B_{std} es la desviación estándar de la distribución Beta.

Y esto sería un UCB Bayesiano o también llamado UCB-Bayes.

BANDIDO MULTIBRAZO: BAYESIAN BANDIT Y UCB-BAYES

Veamos los resultados de comparar UCB1, Bayesian Bandit (Muestreo de Thompson) y UCB-Bayes en un ejemplo con tres bandidos con tasas de ganancia del 2%, 6% y 10%, cada uno:



Como se puede ver, tanto el Muestreo de Thompson (azul) como el UCB Bayesiano (naranja) convergen antes al bandido óptimo, en comparación con el UCB1. Lo que se puede ver también en el número de veces que ha jugado cada uno de ellos con los bandidos que no son óptimos. En el caso del Muestreo de Thompson se ha seleccionado un bandido no óptimo 406 veces (117 + 289), para Bayes-UCB han sido 593 veces (68 + 525) y para UCB1 son 11.339 (2680+ 8719) las veces que se selecciona uno de los bandidos no óptimos. Lo que nos indica que usar la distribución Beta para estos bandidos es una buena solución. Quizás el mayor problema que tienen tanto el Muestreo de Thompson como Bayes-UCB es que necesitan mucho más tiempo que los otros algoritmos para seleccionar al mejor bandido. Esto es así por la necesidad de evaluar la función Beta y esto pasa para ambos métodos.

Entonces, en resumen hemos visto dos métodos bayesianos para resolver un problema tipo bandido multibrazo: el Muestreo de Thompson y Bayes-UCB. Ambos métodos probabilísticos basados en ideas bayesianas. Estos métodos han demostrado seleccionar una mayor cantidad de veces el bandido óptimo al trabajar con bandidos basados en la distribución Binomial.

ESTIMACIÓN E INFERENCIA BAYESIANA

TIPOS DE PRIORS

En la sección anterior vimos el problema del bandido multibrazo que es un tipo de problema A/B que se puede extender a muchas áreas, no solo se interpreta como máquinas tragamonedas donde jugamos y ganamos o perdemos, sino que puede interpretarse y aplicarse a otras áreas como marketing online, donde nos interesa comparar dos paginas web, dos anuncios diferentes, dos botones diferentes, etc. O en medicina donde nos interesa comparar dos o más tratamientos diferentes. Y en general en muchos otros ejemplos de la vida real. Vimos tanto el enfoque clásico, como el heurístico, como el bayesiano. Una ventaja adicional del enfoque bayesiano (y por supuesto de los heurísticos también) es su facilidad para procesar información secuencialmente. Lo cual es muy útil para problemas que se llaman “en línea” porque se van actualizando los datos cada cierto tiempo.

Desde el enfoque bayesiano vimos que esto equivale a calcular la distribución a posteriori en el momento presente, utilizando la regla de Bayes, que me dice que esta es proporcional a la verosimilitud por la distribución a priori, y en el momento siguiente esa posteriori que ya tenemos se convierte en la priori, que es la información actualizada, para hallar una nueva posteriori que nos permita hacer la inferencia que nosotros deseamos. Entonces el enfoque bayesiano proporciona un procedimiento automático para expresar el aumento de nuestro conocimiento respecto al parámetro a medida que se recibe información adicional. Este es uno de sus aspectos más atractivos.

La mayor dificultad práctica del enfoque bayesiano es cómo especificar la distribución a priori: normalmente la información de la que disponemos es cualitativa y el enfoque bayesiano requiere que

establezcamos una distribución de probabilidad sobre sus valores. Entonces podemos considerar 4 casos distintos:

1. **La distribución a priori se conoce porque proviene de estudios anteriores**. Por ejemplo, supongamos que tratamos de determinar el porcentaje de elementos defectuosos en un proceso. Antes de tomar la muestra conocemos que se hizo un estudio similar hace unos meses y suponiendo que las condiciones no han cambiado, tomaremos la distribución a posteriori del estudio anterior como distribución a priori del estudio actual. Como segundo ejemplo, supongamos que una empresa está interesada en conocer el tiempo medio que las personas de una zona determinada dedican a navegar por Internet. Supongamos que conocemos un estudio de esta variable en otra zona de características similares, podemos tomar la distribución a posteriori del estudio realizado como distribución a priori para nuestra zona.

2. **La distribución a priori puede ser importante respecto a la muestral, pero la información existente es subjetiva y no formalizada.** Podemos comenzar por decidir el valor más probable del parámetro, que será la moda de la distribución, su rango de valores posibles (o que cubre el 99.9% de la distribución) y si la distribución es o no es simétrica con relación a la moda. Por ejemplo, en el caso de lanzar una moneda, probablemente nuestra opinión a priori sobre la proporción de las caras es 0.5 y esta será la moda de la distribución. Si pensamos que las desviaciones sobre este valor serían debidas a posibles desperfectos por el uso, es razonable suponer una distribución simétrica respecto a 0.5 y con una pequeña variabilidad respecto a este valor central. En un segundo sobre la edad de los estudiantes de universidad, nuestra estimación a priori dependerá mucho de las características de la universidad: si tiene o no programas para adultos, la importancia del 3er ciclo, etc. Para establecer nuestra opinión podemos fijar el valor más probable, el intervalo central donde debe estar el 50% de la densidad y la forma general de la distribución. Por ejemplo, supongamos que en una universidad sin programas para adultos pero con un amplio programa de tercer ciclo pensamos que el valor más probable (moda) es alrededor de 52 y que estamos seguros de que el estudiante más veterano debe tener más de 35 años y menos de 67.

3. **La información a priori es pequeña con relación a la muestral.** Podemos elegir una distribución a priori que refleje globalmente nuestra opinión, en particular la moda a priori y el rango de valores posibles, pero sin preocuparnos mucho del resto de los detalles. En estos casos elegiremos una distribución conjugada para el problema, que son las distribuciones que facilitan el cálculo de la posteriori. Ya hablamos sobre estas distribuciones en la sección anterior, vimos que para el caso de una variable Binomial, se podía usar la distribución Beta como la priori, y con ello obteníamos que la posteriori también iba a ser una Beta. Más adelante en esta sección veremos más detalles y otras posibilidades.

4. **La información a priori es despreciable frente a la muestral, o no queremos tenerla en cuenta en el proceso de inferencia.** En este caso podemos utilizar los métodos clásicos de los capítulos anteriores o utilizar el enfoque bayesiano con una distribución a priori no informativa o de referencia, que también veremos más adelante más detalles sobre ello.

DISTRIBUCIONES CONJUGADAS

El cálculo de la distribución a posteriori, como ya hemos mencionado anteriormente, puede ser complicado y requerir métodos numéricos. El problema se simplifica si podemos expresar aproximadamente nuestra información a priori con una distribución que facilite el análisis. Una familia de distribuciones a priori adecuada para este objetivo es aquella que tiene la misma forma que la verosimilitud, de manera que la posteriori pueda calcularse fácilmente al pertenecer a la misma familia que la priori. A estas familias se las denomina **conjugadas**.

Una clase $\mathcal{C}$ de distribuciones a priori para un parámetro vectorial θ es conjugada si cuando la prior pertenece a esa clase $p(\theta) \in \mathcal{C}$, entonces también lo hace la posteriori $p(\theta|X) \in \mathcal{C}$.

La distribución conjugada a priori se elige tomando como distribución la verosimilitud y modificando los valores de las constantes para que la función resultante sea una función de densidad y tenga características coincidentes con nuestra información a priori. Para el modelo Binomial esta es la distribución Beta que se presentó en la sección anterior.

Distribución	**Conjugada**
Bernoulli	Beta
Binomial	Beta
Binomial Negativa	Beta
Poisson	Gamma
Uniforme	Pareto
Exponencial	Gamma
Normal	Normal

DISTRIBUCIONES NO INFORMATIVAS O DE REFERENCIA: JEFFREYS PRIOR

Una distribución no informativa o de referencia pretende no modificar la información obtenida en la muestra. Intuitivamente, una distribución a priori no informativa para un parámetro de localización es aquella que es localmente uniforme sobre la zona relevante del espacio paramétrico, y escribiremos $p(\theta) = c$.

Sin embargo, esta elección tiene el problema de que si el vector de parámetros puede tomar cualquier valor real, la integral será igual a infinito y eso no puede ser, tiene que ser igual a 1:

$$\int_{-\infty}^{\infty} p(\theta) d\theta = 1$$

Pero en efecto, si podemos suponer que a priori un parámetro escalar debe estar en el intervalo $\theta \in [-h, h]$, donde h puede ser muy grande, pero es un valor fijo, la distribución a priori $p(\theta) = \frac{1}{2}h$ sí vale porque integra a uno. En cuyo caso se llama distribución propia, y si sucede lo contrario y no integra se llamaría impropia.

Hay otra distribución no informativa que se puede utilizar y tiene el nombre de **Jeffreys prior**. En general es bastante utilizada cuando no se tiene ninguna información sobre la distribución a priori. La formula es la raíz cuadrada del determinante de la información de Fisher que es una matriz que nos dice cuanta información puede obtenerse de una muestra sobre un parámetro desconocido.

$$p(\theta) \propto \sqrt{\det\big(I(\theta)\big)}$$

Donde θ puede ser un parámetro o un vector de parámetros, y $\det I$ es el determinante de la información de Fisher $I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right)$.

Vamos a ver un ejemplo con datos Normales con varianza conocida, y vamos a obtener la prior de Jeffrey no informativa para la media que sería el parámetro que desconocemos en este ejemplo. La información de Fisher es menos el valor esperado de la segunda derivada de la log-verosimilitud, con respecto al parámetro desconocido de la Normal (θ es su media):

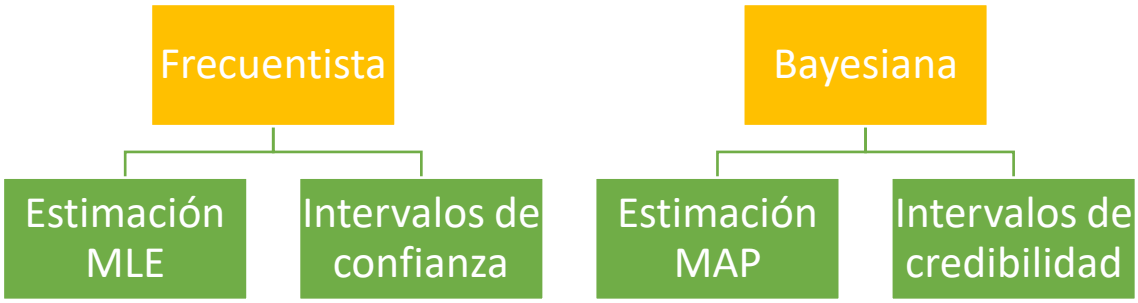
$$I(\theta) = -E\left(\frac{\partial^2 \log L}{\partial \theta^2}\right) = \frac{n}{\sigma^2}$$

Entonces la prior de Jeffrey sería proporcional a esta constante (n y σ son conocidos):

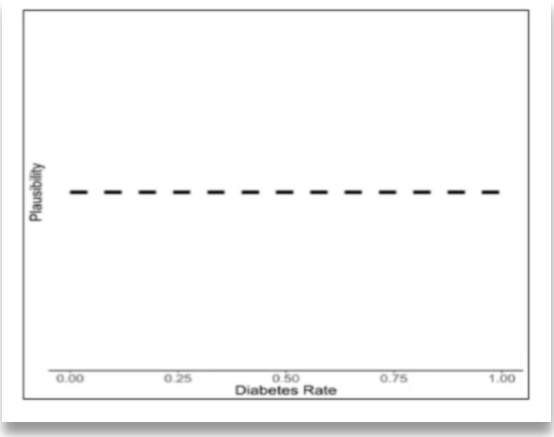
$$p(\theta) \propto \frac{\sqrt{n}}{\sigma}$$

ESTIMACIONES E INTERVALOS DE CREDIBILIDAD

La inferencia estadística podría dividirse de la siguiente manera. Desde el enfoque frecuentista sabemos que podemos obtener una estimación puntual de un parámetro (con el método de máxima verosimilitud) y un intervalo de confianza para esa estimación, con cierto nivel de confianza. Desde el enfoque bayesiano sabemos que nosotros obtenemos directamente una distribución para ese parámetro desconocido. Con ella también podríamos hallar una estimación puntual, escogiendo la moda de la distribución, lo que se llama *Maximum a posteriori estimation* o *MAP*. Y lo equivalente a intervalos de confianza se llamarían ahora intervalos de credibilidad.

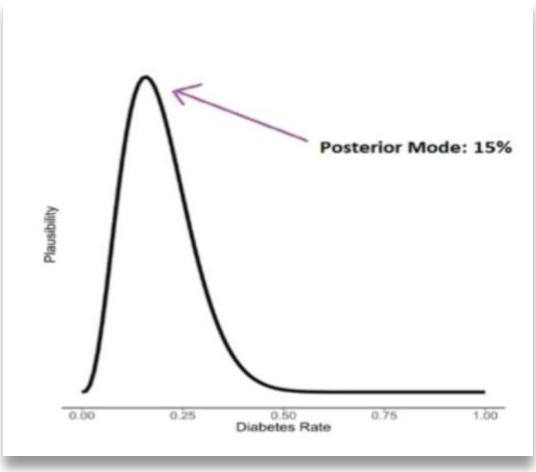


Veamos esto a través de un ejemplo. Se pide determinar la proporción de población que tiene diabetes. Sería impráctico e imposible preguntar a todas las personas en el mundo si tienen diabetes o no. Entonces lo que haríamos es estimar esta proporción basándonos en la información que sí podemos tener. De primeras, antes de recopilar ninguna información, no conocemos nada sobre cuál podría ser el valor de esta proporción. Así que podría ser cualquier valor entre 0 y 100%.



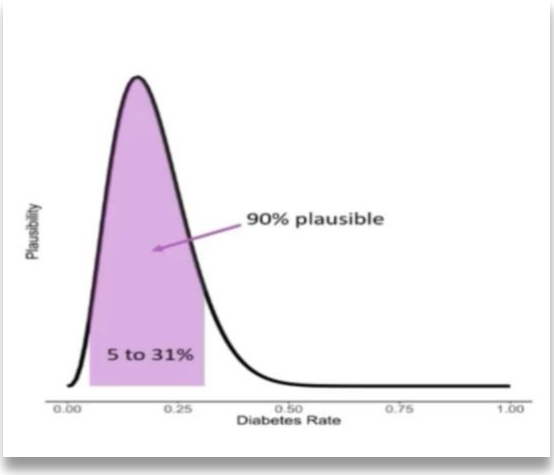
Para ganar información supongamos que entrevistamos a 20 personas y de ellas 3 resultan tener diabetes. Basándonos en esta nueva información, ¿cuál es el valor más probable para esa proporción? Sería 3/20=0,15.

La distribución a posteriori nos dice cuán probable es cada valor para ese parámetro que es la proporción de diabetes, dada la muestra que tenemos.

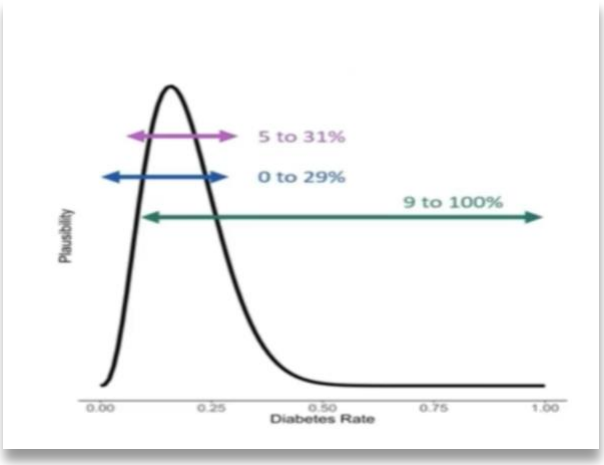


En esta distribución a posteriori ¿cuál es el valor más probable? (la moda) Tenemos que la moda es 15%. Esto sería una estimación puntual, que nos dice cuál es el valor que creemos más probable para el parámetro desconocido que nos interesa pero que no nos comunica cuánta confianza tenemos en esa estimación. Esa proporción también podría ser 10% o 20% porque son bastante probables también. De hecho hay un rango de valores posibles que podrían ser una buena elección. Ese rango de valores se denomina Intervalo de credibilidad. Es un rango de valores que representa un nivel dado de posibilidad. Y está basado en la distribución a posteriori.

Si ese rango está entre el 5% y el 31% en este ejemplo, eso equivale al 90% de credibilidad. La anchura de ese intervalo nos da un indicador de cuanta confianza tenemos en esa estimación puntual.



En este caso, si por ejemplo tomamos el rango de 0 a 29% también corresponde con un 90% de posibilidad, y si tomamos el de 9 a 100% también.

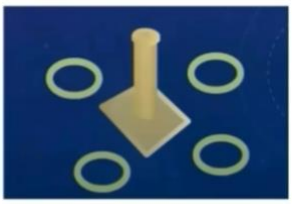


Los tres tienen un 90% de posibilidad. ¿Cual de los tres elegiríamos para comunicarlo como resultado? El último es muy amplio, tiene una gran cantidad de valores, prácticamente todo el intervalo de valores posibles que teníamos antes de recopilar ninguna información, y entre los otros dos el más estrecho y por tanto el que más confianza nos da es el primero.

Entonces siempre vamos a elegir el intervalo que contiene a la estimación, con mayor densidad y que sea el más estrecho para un nivel dado de posibilidad que en este caso fue del 90%.

Ahora bien, vamos a ver en qué se diferencian exactamente los intervalos de confianza clásicos y frecuentistas de los intervalos de credibilidad bayesianos. Los intervalos de confianza frecuentistas se pueden interpretar como un juego de lanzar anillas a un poste. Si nos dan un conjuntos de anillas de un tamaño determinado, podemos lanzarlas hacia el poste y tratar de averiguar cuán frecuente es que acertemos la anilla en el poste. Si nos dan anillas más grandes, es más fácil acertar. Y la proporción de éxito será mayor. Y si nos dan anillas más pequeñas es más difícil acertar. Y la proporción de éxito será menor. En el enfoque frecuentista, el poste representa el parámetro poblacional. La posición donde cae cada anilla representa un estadístico muestral (diferentes anillas, diferentes muestras). Y la anilla como tal representa el intervalo de confianza.





Cada vez que se lanza una anilla, tenemos una muestra y se calcula un intervalo de confianza. Nosotros podemos elegir el nivel de confianza (95% por ejemplo) antes de tirar las anillas, lo que equivale a tener que fijar el tamaño de las anillas antes de tirarlas pero una vez que las hemos tirado, NO hay un 95% de probabilidad de que la anilla esté alrededor del poste. Los intervalos de confianza están basados en una idea de muestras repetidas (las diferentes anillas). Entonces la idea es que se escoge el tamaño del IC (la anilla) tal que de muchas tiradas el 95% de las veces el IC va a incluir al parámetro poblacional. O lo que en el ejemplo sería de todas las veces que tiramos anillas, el 95% acertamos en el poste. Después de calcular un intervalo de confianza específico NO se puede decir que eso equivale a que haya un 95% de probabilidad de que el parámetro poblacional esté en ese intervalo de confianza, porque o está o no está. Entonces la interpretación correcta sería: que el 95% de los IC confianza construidos de manera similar contendrán al verdadero valor del parámetro. No que hay un 95% de probabilidad de que el parámetro esté entre esos valores del IC.

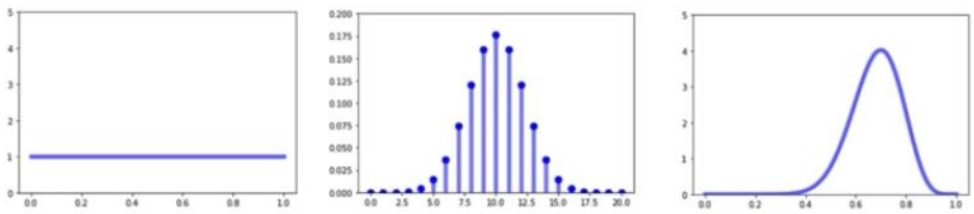
En el caso de intervalos de credibilidad sí se puede decir que hay un 95% de probabilidad de que el parámetro esté entre esos valores del intervalo de credibilidad porque está hallado a partir de la distribución de ese parámetro como variable aleatoria.

CONTRASTES DE HIPÓTESIS DESDE EL ENFOQUE BAYESIANO

Habiendo aprendido sobre el concepto de distribuciones conjugadas, estimaciones puntuales, intervalos de credibilidad, podríamos preguntarnos ahora, ¿cómo realizar contrastes de hipótesis en un entorno bayesiano?

Vamos a suponer que tenemos este ejemplo sencillo, del lanzamiento de una moneda y lo que nos interesa es la verdadera probabilidad de obtener cara por ejemplo. Vamos a suponer que nuestro prior es una uniforme, que es lo mismo que una $Beta(1,1)$ y que nuestra verosimilitud es una Binomial. En nuestro caso específico, digamos que lanzamos la moneda 20 veces y en la muestra obtuvimos 14 caras de 20 tiradas. Entonces la verosimilitud es una $Binomial(n,p)$ con parámetros $n = 20$ y $p = \frac{14}{20} = 0.7$.

Haciendo uso de las propiedades que vimos para las distribuciones conjugadas llegamos a que la distribución a posteriori es también una Beta como la priori pero con estos parámetros: $Beta(\alpha = 1 + 14\,,\ \beta = 1 + 20 - 14) = Beta(15,7)$.



Como conocemos la distribución exacta de la posteriori, podemos usar la función de densidad de probabilidad a posteriori, que simplemente describe cuánta probabilidad se asocia a cada posible evento.

¿Cómo lo hacemos? Digamos que estamos interesados en probar la hipótesis de que la proporción de caras verdadera es mayor que 0.5, es decir, $P(p > 0.5)$. Lo que haríamos es mirar la distribución Beta

posteriori actualizada $Beta(15,7)$ y buscamos la posición 0.5 y miramos el área que hay en el lado derecho del 0.5. Y esta sería la probabilidad de que la proporción de aciertos verdaderos (de sacar cara) sea mayor que 0.5. Entonces, la probabilidad de que la proporción de aciertos verdaderos sea mayor que 0.5 es del 96% aproximadamente:

$$P(p > 0.5) = 96\%$$

Casi toda la masa de probabilidad se encuentra a la derecha del 0.5.

Así que podemos estar bastante seguros de que la proporción de caras verdadera es mayor que 0.5.

¿Qué pasaría si la hipótesis fuera que es mayor que 0.6?

Nuevamente, solo marcamos el punto relevante aquí 0.6 y miramos la masa de probabilidad que se encuentra a la derecha de ese valor. Que sería alrededor de un 80% de probabilidad.

$$P(p > 0.6) = 80\%$$

Y podemos hacer lo mismo con la probabilidad de que la proporción de caras verdaderas sea mayor que 0.7. Que es aproximadamente el 45 por ciento.

$$P(p > 0.7) = 45\%$$

Es decir que el hecho de que sepamos cuál es la distribución exacta a posterior nos permite comprobar fácilmente esas hipótesis y calcular la probabilidad exacta de que la proporción verdadera se encuentre por encima de un cierto umbral que nos interese.

Pero, ¿qué hacemos si la función de verosimilitud no tiene una distribución conjugada simple como en este caso que es una beta y podemos hallar esas probabilidades de manera sencilla, sino que esa verosimilitud sigue otra distribución diferente para la que no hay definida ninguna distribución conjugada? Una posibilidad es utilizar métodos de muestreo que no proporcionan el resultado exacto, pero en la mayoría de los casos proporcionan una muy buena estimación de la distribución a posteriori. Los veremos a continuación.

MUESTREO POR RECHAZO O REJECTION SAMPLING

Una de las formas más comunes de estimar la distribución a posteriori es el Muestreo por Rechazo o en inglés Rejection Sampling. Vamos a formalizar el procedimiento de muestreo por rechazo.

```
Sample_list = empty list

For i=1,…,n:

        Obtenemos una muestra a partir de la probabilidad a priori y la condición.

        Si la muestra es correcta:

                Añadir a Sample_list

        Sino:

                Continuar iterando

End

Return Sample_list
```

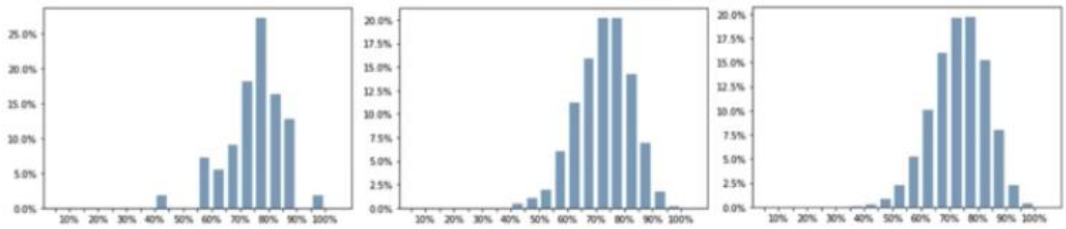
Comenzamos inicializando una lista vacía, que llamamos sample_list que es donde iremos guardando nuestras muestras resultantes del muestreo. Luego vamos a hacer n iteraciones. Y quizás nos preguntemos, ¿cómo puedo calcular el número de iteraciones? La respuesta es que depende de qué tan buena queremos que sea la estimación al final. Lo que normalmente se hace es definir un criterio de convergencia y decir que dejamos de hacer las iteraciones cuando la estimación de la probabilidad deseada no cambia mucho en, digamos, por ejemplo un 1% dentro de 100 iteraciones seguidas. Entonces, ahora, para cada una de las iteraciones, obtenemos una muestra a partir de la probabilidad a priori y la condición. Por ejemplo supongamos que estamos con el ejemplo anterior del lanzamiento de la moneda y aunque en ese caso claramente no necesitamos hacer un muestreo vamos a cogerlo porque es un ejemplo bastante sencillo para entender la idea del muestreo por rechazo. Y además nos permitirá comparar directamente los dos resultados ya que en realidad nosotros sabemos cuál es la distribución exacta a posteriori, era una Beta(15,7).

En nuestro caso la priori era una Beta(1,1) que es lo mismo que una uniforme. Entonces solamente hacemos una elección aleatoria en el intervalo 0 a 1. Digamos que por ejemplo obtuvimos 0.61. En el siguiente paso miramos la probabilidad condicional, la verosimilitud para determinar el número de caras.

En nuestro ejemplo sabemos que el numero de caras sigue una Binomial(n,p) donde n=20 y p va a ser igual a 0.61. Entonces muestreamos el numero de caras y digamos que resulta ser 13. Entonces, lo que tenemos es una combinación de dos cosas: la probabilidad muestreada de 0.61 en correspondencia con un numero obtenido de caras 13. Y ahora comprobamos si el número de aciertos coincide con lo que observamos en el la muestra. Recordemos que en la muestra original teníamos 14 caras. Entonces, lo que haríamos es desechar esa muestra nueva que hemos obtenido y no usarla. Y pasamos a la siguiente iteración.

En este caso, imagina que obtenemos una probabilidad de 0.45 y un numero de caras de 14 por casualidad. Como esto sí coincide con lo que tenemos en la muestra original, guardamos ese 0.45 como una muestra de nuestras proporciones de caras, que es lo que estamos muestreando, es decir, la guardamos en *sample_list*. Y lo que haremos será iterar este procedimiento muchas veces hasta que veamos algún tipo de convergencia en la distribución a posteriori.

Veamos los resultados del muestreo con estos histogramas de las frecuencias del valor de esa proporción que hemos muestreado y que hemos agregado a la *sample_list*.



En el primer histograma, se han realizado 1000 iteraciones. Posteriormente, realizamos 10.000 iteraciones. Y finalmente, 100.000 iteraciones. A medida que aumentamos el número de iteraciones, la distribución a posteriori se aproxima cada vez más y más a la distribución exacta, la Beta. Pero aquí vamos a ver algo muy importante. Una de las principales debilidades de este muestreo es que podríamos terminar rechazando muchos muestras, dependiendo de cuán improbable sea el evento, o dependiendo de a lo que condicionamos. En nuestro caso, condicionamos a que el número de caras sea exactamente 14. Como lo primero que hacemos es muestrear de la prior que es una distribución Uniforme y luego muestreamos de la distribución Binomial con n=20 y p=lo que se ha obtenido antes, se puede demostrar matemáticamente que solo en el 5% de los casos realmente vamos a observar un valor de 14. Lo que significa que el 95% de las muestras son rechazadas. Entonces, cuando realizo 1000 iteraciones, realmente significa que tenemos alrededor del 5% muestras, alrededor de 50. Entonces, el primer histograma solo consta de 50 puntos de datos. De manera similar, el segundo gráfico consta de 500 puntos de datos y en el último de 5000 puntos de datos, aproximadamente.

METROPOLIS-HASTINGS

Creo que todos podemos estar de acuerdo en que el procedimiento de muestreo de rechazo es bastante poco útil en el sentido de que solo aceptamos un 5% de la las muestras y rechazamos el otro 95%. Afortunadamente, hay un segundo método de muestreo que funciona mucho mejor en este caso, el algoritmo de Metrópolis-Hastings.

Es un método de Monte Carlo en cadena de Markov para obtener una secuencia de muestras aleatorias a partir de una distribución de probabilidad a partir de la cual es difícil el muestreo directo. El método de Montecarlo es un método no determinista o estadístico numérico, usado para aproximar expresiones matemáticas complejas y costosas de evaluar con exactitud. En la teoría de la probabilidad, se conoce como cadena de Márkov o modelo de Márkov a un tipo especial de proceso estocástico discreto en el que la probabilidad de que ocurra un evento depende solamente del evento inmediatamente anterior.

Comenzamos como antes con una lista vacía.

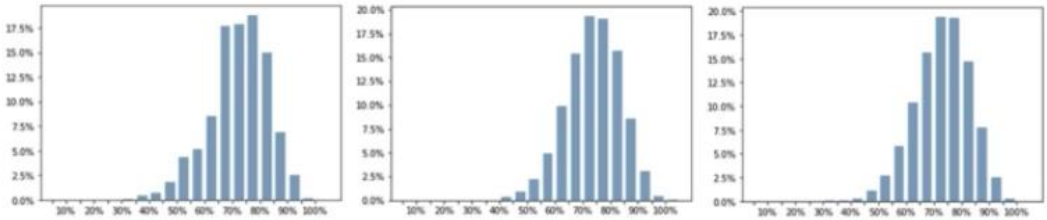
```
Sample_list = empty list

Current_sample = [0.5,14]

For i=1,…,n:

        Obtenemos una muestra "Proposal_sample" a partir de una uniforme para los parámetros.

        Calcular la probabilidad p de Current_sample

        Calcular la probabilidad q de Proposal_sample

        Definir un ratio de aceptación a=q/p

        Obtener una muestra uniforme u del intervalo [0,1]

        Si a>u:

                Aceptar Proposal_sample y usar como nueva Current_sample

        Sino:

                Rechazar Proposal_sample y continuar con Current_simple

        Guardar Current_simple en Sample_list

End

Return Sample_list
```

Y además, comenzamos con un valor muestral actual "current sample", que es cualquier valor para P (que en realidad lo desconocemos así que podemos ponerle que sea 0.5) y el valor de interés del numero de caras de nuestra muestra que sabemos que es 14. Y ahora, de nuevo, realizamos varias iteraciones. Primero, obtenemos de una uniforme una "propuesta de muestra"="proposal sample" para todas las variables excepto la que tenemos ya una muestra, es decir, para p porque en la muestra que tenemos ya sabemos que el numero de caras es 14. Entonces por ejemplo sigamos que obtenemos 0.61, 14. Lo que hacemos a continuación es calcular la probabilidad de la current sample. Y calculamos la probabilidad de la proposal sample. Y definimos un índice de aceptación, A, que es solo la probabilidad de la proposal

sample entre la de la current sample. Luego obtenemos un valor uniforme entre 0 y 1 y comparamos ese valor con A. Y, si la tasa de aceptación A es mayor que el valor u, obtenido de manera aleatoria con la uniforme 0-1, aceptamos la proposal sample y la usamos como nuestra nueva current sample en la siguiente iteración. Y si sucede lo contrario, rechazamos la muestra propuesta y nos quedamos con la actual. Y luego vamos guardando la current sample haya cambiado o no, en la sample list.

Vamos a notar un punto importante antes de que pasemos a los resultados de las simulaciones. Dijimos que dibujamos uniformemente la proposal simple, es decir, p porque el numero de caras está fijado en 14. Esta es una posible distribución de la que podemos muestrear en Metropolis Hastings. Pero también podemos usar una distribución diferente, por ejemplo, también podríamos tomar muestras de la distribución Normal, centrada alrededor de p. Entonces, en lugar de simplemente obtener cualquier valor entre cero y uno. Lo que la distribución normal hará es computar una distribución alrededor del valor de cero punto cinco, que es el valor del p en la current sample en el primer paso, y extraer una muestra de esta distribución. Por ejemplo, digamos que obtenemos 0.7. El único punto a considerar cuando utilizamos una distribución para muestrear aquí diferente de estas dos es que para que funcione, la distribución propuesta debe ser simétrica. Es decir que la probabilidad de obtener un 0.7 muestreando de una normal alrededor de 0.5 seria igual a la probabilidad de obtener un 0.5 si muestreamos de una normal centrada alrededor de 0.7.

Veamos los resultados:



Al contrario que en Rejection Sampling, ya con menos iteraciones nos acercamos bastante a la distribución beta que sabemos que es la exacta. Y esto se debe a que con Metropolis Hastings no estamos rechazando un 95% de muestras. Aquí aunque tenemos histogramas esto es porque hemos discretizado las distribuciones.

MÉTODOS DE MACHINE LEARNING BAYESIANOS

El aprendizaje automático (machine learning) se ha convertido en un tema muy importante en los últimos años. De hecho, hoy en día hay numerosas aplicaciones que dependen de estos métodos. Por eso seguramente, en este punto, nos estaremos preguntando cómo se relaciona la estadística bayesiana con el aprendizaje automático. Este será el tema central de esta última sección.

El primer punto clave a destacar es que prácticamente todo el aprendizaje automático se basa en la noción bayesiana de probabilidad. O dicho de otra manera, el aprendizaje automático siempre asume implícitamente que puede asignar probabilidades a eventos que no son repetibles solo para expresar su grado de creencia de que este evento está sucediendo.

Hay tres paradigmas básicos en el aprendizaje automático: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado. En esta sección, nos centraremos en el aprendizaje supervisado y el aprendizaje no supervisado.

Recordemos de qué se trata el aprendizaje supervisado. Por ejemplo, supongamos que tenemos un conjunto de datos de cancelaciones de clientes en el pasado, y para cada cliente conocemos las siguientes variables: el tipo de cliente, la región, la cantidad de productos, el uso, si ha habido un aumento de precio asignado a este cliente y si el cliente ha cancelado o no. Entonces el objetivo del aprendizaje supervisado

es desarrollar un modelo que razone si el cliente cancela o no en dependencia de las variables de entrada: el tipo de cliente, la región, la cantidad de productos, el uso, y si ha habido un aumento de precio asignado a este cliente. Tal modelo puede ser extremadamente útil porque podríamos aplicarlo a una base de datos de clientes actuales, para los que tenemos toda la información de entrada pero que no sabemos si este cliente cancelará en un futuro o no. Entonces este modelo podría ayudarnos a determinar qué tan probable es que alguno de mis clientes actuales cancele el contrato.

Entonces, el punto clave sobre el aprendizaje supervisado es que tenemos una variable de interés con etiquetas, que en nuestro caso, nos dice si un cliente cancela o no. Y nos gustaría predecir esas etiquetas a partir de un grupo de variables de entrada.

En contraste, en el aprendizaje automático no supervisado, no tenemos los datos etiquetados, sino que solo nos proporcionan los datos de las variables de entrada. Digamos que por ejemplo nos proporcionan dos variables de entrada, una que nos dice la cantidad de horas de uso de cada cliente, y otra que nos dice la cantidad de llamadas a soporte de cada cliente. Digamos que obtenemos los siguientes datos.

Entonces aquí el objetivo del aprendizaje no supervisado consiste en encontrar alguna estructura en los datos sin que se nos proporcione explícitamente esto que estamos tratando de predecir. Por lo tanto, el aprendizaje no supervisado está estrechamente relacionado con clusterización o clustering. En este caso, podríamos encontrar que hay dos grupos de clientes distintos, por ejemplo, aquellos con poco uso, pero una gran cantidad de llamadas a soporte y aquellos clientes con un uso más alto, pero un número bajo de llamadas a soporte.

Entonces, para resumir, la diferencia crucial entre el aprendizaje supervisado y el aprendizaje no supervisado es que en el aprendizaje supervisado se nos proporciona una etiqueta de interés con la que trataremos de construir un modelo. Y en el no supervisado no disponemos de esas etiquetas.

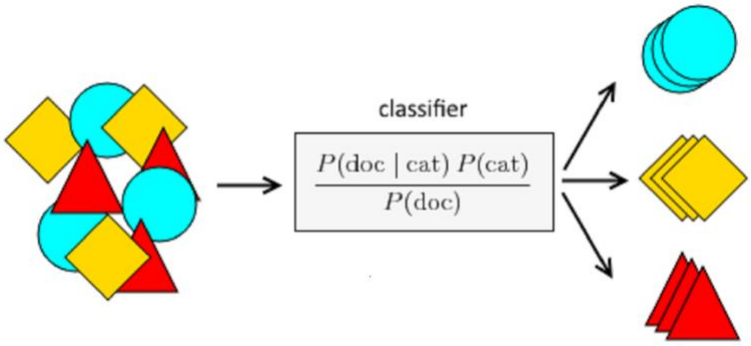
De esta manera, también podemos recalcar que es más fácil medir el rendimiento de los algoritmos de aprendizaje supervisado porque simplemente podemos dividir toda la base de datos que tenemos en una base de datos de entrenamiento y una base de datos de prueba, entrenar el algoritmo con la base de datos de entrenamiento, y luego aplicar este modelo para predecir las etiquetas de los datos de prueba y comparar esas predicciones que hace nuestro modelo con las etiquetas reales que tienen esos datos de prueba, porque disponemos de ellas. Y la diferencia entre lo real y lo predicho nos da alguna medida del desempeño de ese método.

Por el contrario, para el aprendizaje no supervisado, estamos interesados en aprender la estructura de los datos sin el objetivo explícito de predecir, aunque también una vez determinada la estructura, por ejemplo la agrupación, se podría predecir a qué grupo pertenecen observaciones futuras.

Como iremos viendo, los métodos bayesianos ayudan a los algoritmos de aprendizaje automático a extraer información crucial de pequeños conjuntos de datos y manejar los datos faltantes. Desempeñan un papel importante en una amplia gama de áreas, desde el desarrollo de juegos hasta el descubrimiento de fármacos. Y nos permiten estimar la incertidumbre en las predicciones, lo que resulta vital para campos como la medicina, la bioinformática, etc. En esta sección veremos algunos algoritmos de aprendizaje automático que se basan en el enfoque bayesiano.

NAIVE BAYES

Ahora vamos a ver uno de los principales algoritmos de aprendizaje supervisado que está basado en el enfoque bayesiano: Naive Bayes (NB). La palabra “naive” del inglés, significa en español "ingenuo" y esto es básicamente porque el algoritmo asume que las características de una medición, es decir, nuestras variables, son independientes entre sí. Esto es ingenuo porque esto en la práctica (casi) nunca es cierto. Sin embargo resulta que NB en muchos casos funciona bastante bien. NB es un algoritmo de clasificación muy intuitivo que hace la siguiente pregunta: "Dadas estas características, ¿esta medición pertenece a la clase A o B?"



Y la responde tomando la proporción de todas las mediciones anteriores con las mismas características pertenecientes a la clase A multiplicada por la proporción de todas las mediciones de la clase A. Si este número es mayor que el cálculo correspondiente para la clase B, entonces decimos que la medida pertenece a la clase A. Simple, ¿verdad? Sin embargo, fíjate que antes hemos dicho que miramos mediciones anteriores con las mismas características y, en la práctica, rara vez veremos muchas mediciones con conjuntos de características idénticos. De hecho, si tuviéramos mediciones idénticas en nuestros datos medidos previamente, podríamos clasificar duplicados exactos, lo que hace que la regla de Bayes sea prácticamente inútil para la clasificación.

Ahora bien, si, en cambio, asumimos ingenuamente que todas las variables son independientes entre sí, no tenemos que depender de duplicados exactos en nuestro conjunto de datos de entrenamiento para hacer una clasificación. Simplemente podemos tomar cada característica (variable) por separado y determinar la proporción de datos anteriores que pertenecen a la clase A que tienen el mismo valor para esta característica (o esa variable) solamente. Luego hacemos lo mismo con todas las demás variables y tomamos el producto. Nuevamente multiplicamos esto con la proporción de la clase A en el conjunto de datos y vemos si este número es mayor que si hiciéramos el cálculo correspondiente para la clase B.

Lo bueno de NB es que la suposición ingenua en realidad tiende a ayudar a la clasificación. Pensémoslo de esta manera: si dos características son dependientes, por ejemplo, la longitud del cabello y el género, entonces asumir que son independientes significa que tiene doble evidencia. Si tanto el género como el cabello largo están más asociados con ser fan de Justin Bieber, entonces asumir independencia nos hizo estar aún más seguro de que la chica que queremos clasificar es una “Belieber”.

Entonces, en resumen. La idea detrás del clasificador Naive Bayes es solo una cuestión de contar cuántas veces cada atributo coexiste con cada clase. Veamos la idea del algoritmo con un ejemplo muy intuitivo. Supongamos que estamos caminando y nos encontramos con un objeto rojo frente a nosotros. Este objeto rojo puede ser clasificado en estas tres posibles cosas: un bate, un gato o una pelota. ¿Qué dirías que es? Definitivamente asumiríamos que es una pelota. ¿Pero por qué es así?

Supongamos que estamos haciendo un modelo de ML y le hemos dado la tarea anterior de clasificar un objeto entre bate, pelota y gato. Al principio, querremos seguramente crear un modelo que identificará las características del objeto y luego lo mapeará con la clasificación que hemos definido, de modo que si un objeto es redondo, entonces será una pelota, o si el objeto es un ser vivo, entonces será un gato. En nuestro caso, si el objeto es rojo, diremos que lo más probable es que sea una pelota. Por que? Porque desde nuestra niñez y durante toda la vida, hemos visto pelotas rojas, pero no hemos visto nunca un gato rojo y muy probablemente tampoco un bate rojo.

Entonces, en nuestro caso, podemos clasificar un objeto mapeando sus características con nuestro clasificador, individualmente. Como en nuestro caso, este color rojo fue mapeado con un bate, un gato y una pelota, pero finalmente, obtenemos la mayor probabilidad de que el objeto rojo sea una pelota, el resultado final es que clasificamos ese objeto como una pelota.

Si miramos el Teorema de Bayes

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Aquí c representa el conjunto de clases: bate, gato o pelota. Y x representa cada variable individual.

$P(c|x)$: es la probabilidad a posteriori de la clase c dado el predictor (variable) x.

$P(c)$: es la probabilidad de la clase c.

$P(x|c)$: es la verosimilitud.

$P(x)$: probabilidad a priori.

Veamos otro ejemplo. Digamos que tenemos datos de 1000 frutas. Las frutas pueden ser: Plátanos, Naranjas, o de Otro tipo. Esas serían nuestras tres clases. Imagina que sobre esas frutas medimos las siguientes características (variables): si es larga o no, si es dulce o no y si es de color amarillo o no. La siguiente tabla me dice los resultados, sobre cuántas frutas cumplen las características: larga, dulce, color amarillo.

Fruta	Larga	Dulce	Amarilla	Total
Plátano	400	350	450	500
Naranja	0	150	300	300
Otras	100	150	50	200
Total	500	650	800	1000

De la tabla de arriba qué información tenemos?

- 50% de las frutas son plátanos
- 30% de las frutas son naranjas
- 20% son de otro tipo

Y también:

- De 500 plátanos, 400 (0.8) son largos, 350 (0.7) son dulces, y 450 (0.9) son amarillos.
- De 300 naranjas, 0 son largas, 150 (0.5) son dulces y 300 (1) son amarillas.
- Del resto de frutas (200), 100 (0.5) son largas, 150 (0.75) son dulces y 50 (0.25) son amarillas.

Toda esta información debería proporcionarlos la evidencia suficiente para poder predecir la clase de una fruta nueva que sea introducida. Obviamente imagina que nosotros no vemos la fruta nueva sino que solamente tenemos información sobre sus características, es decir, si es larga, dulce y amarilla, o no.

Entonces, digamos que se nos dan las características de esa nueva fruta y necesitamos predecir la clase. Si nos dicen que la fruta adicional es Larga, Dulce y Amarilla, podemos clasificarla usando la siguiente fórmula para cada clase, y la que tenga mayor probabilidad (puntuación) será la clase ganadora y será la opción en la que clasificaremos a esa nueva fruta.

Supongamos que las variables Larga, Dulce y Amarilla toman valores 1 o 0 según si la fruta cumple esa característica, o no la cumple, respectivamente. Entonces nuestro nuevo dato x corresponde a que las tres variables tomen valor 1:

Por ejemplo, para la clase Plátano:

$$P(Plátano|x = \{Larga, Dulce, Amarilla\}) = \frac{P(\{Larga, Dulce, Amarilla\}|Plátano)P(Plátano)}{P(\{Larga, Dulce, Amarilla\})}$$

Lo primero que vamos a notar es lo siguiente, en el denominador tenemos la evidencia: $P(x) = P(x = \{Larga, Dulce, Amarilla\})$ y como el método asume que las variables son independientes podemos transformar esa probabilidad en un producto de tres probabilidades:

$$P(x = \{Larga, Dulce, Amarilla\} = P(Larga)P(Dulce)P(Amarilla)$$

La independencia hace que la probabilidad condicional del numerador también pueda convertirse en un producto de probabilidades:

$$P(\{Larga, Dulce, Amarilla\}|Plátano) = P(Larga|Plátano)P(Dulce|Plátano)P(Amarilla|Plátano)$$

Entonces para hallar la posteriori de Plátano:

$$P(Plátano|\{Larga, Dulce, Amarilla\})$$
$$= \frac{P(Larga|Plátano)P(Dulce|Plátano)P(Amarilla|Plátano)P(Plátano)}{P(Larga)P(Dulce)P(Amarilla)}$$
$$= \frac{0.8 \times 0.7 \times 0.9 \times 0.5}{0.25 \times 0.33 \times 0.41} = 0.252$$

Para la clase Naranja:

$$P(Naranja|\{Larga, Dulce, Amarilla\})$$
$$= \frac{P(Larga|Naranja)P(Dulce|Naranja)P(Amarilla|Naranja)P(Naranja)}{P(Larga)P(Dulce)P(Amarilla)} = 0$$

Y para la clase Otras frutas:

$$P(Otra|\{Larga, Dulce, Amarilla\}) = \frac{P(Larga|Otra)P(Dulce|Otra)P(Amarilla|Otra)P(Otra)}{P(Larga)P(Dulce)P(Amarilla)}$$
$$= \frac{0.5 \times 0.75 \times 0.25 \times 0.2}{0.25 \times 0.33 \times 0.41} = 0.01875$$

De las tres opciones, la que nos devuelve **mayor probabilidad** es la primera (0.252) por lo que podríamos asumir que **si la fruta es larga, dulce y amarilla será un Plátano**.

Si nos fijamos, en los tres casos el cálculo de las probabilidades tienen las tres el mismo denominador, entonces realmente la cuestión sólo depende de lo que dé el numerador, que es la verosimilitud por la prior.

Ventajas y desventajas de los clasificadores Naive Bayes.

Pros:

- Computacionalmente rápido.
- Sencillo de implementar.
- Funciona bien con pequeños conjuntos de datos.
- Funciona bien con grandes dimensiones.
- Se desempeña bien incluso si la suposición ingenua no se cumple perfectamente. En muchos casos, la aproximación es suficiente para construir un buen clasificador.

Contras:

- Es necesario eliminar las características correlacionadas porque se votan dos veces en el modelo y puede dar lugar a una importancia exagerada.
- Si una variable categórica tiene una categoría en el conjunto de datos de prueba que no se observó en el conjunto de datos de entrenamiento, entonces el modelo asignará una probabilidad cero. No podrá hacer una predicción. Esto a menudo se conoce como "Frecuencia cero". Para solucionar esto, podemos utilizar técnicas de suavizado. Una de las técnicas de

suavizado más simples se llama estimación de Laplace. Sklearn aplica el suavizado de Laplace de forma predeterminada cuando entrena un clasificador Naive Bayes.

ANÁLISIS DISCRIMINANTE BAYESIANO

Anteriormente vimos un método muy sencillo que se fundamenta en el enfoque bayesiano para dar una solución al problema de clasificación supervisada: el método Naive Bayes. Y vimos también que su nombre proviene de la suposición "naive" (ingenua) de que las variables son independientes. En el método que veremos ahora se va asumir una hipótesis diferente, es decir, no vamos a asumir que las variables son independientes, pero sí vamos a hacer una suposición bastante fuerte que es que la distribución del vector multivariante de variables explicativas, dada la variable Y que contiene la información sobre las categorías, sigue una distribución Normal multivariante.

Es decir, sean $X = \{X_1, \dots, X_p\}$ las p variables explicativas y sea Y la variable dependiente, se asume que:

$$X|Y = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

Donde $k = 0, \dots, K-1$ son las K categorías o valores que puede tomar la variable dependiente Y, y si nos fijamos para cada categoría tenemos una distribución Normal con vector de medias $\boldsymbol{\mu}_k$ diferente pero con la misma matriz de covarianza $\boldsymbol{\Sigma}$.

En este caso estamos ante el análisis discriminante lineal (LDA), donde se asume que todos los grupos tienen la misma estructura de variación. Si los grupos tienen variaciones desiguales, para cada grupo tendríamos una matriz de covarianza diferente $\boldsymbol{\Sigma}_k$, y el enfoque se conoce como análisis discriminante cuadrático (QDA).

El objetivo del análisis discriminante en general, para cualquiera de los dos casos particulares, es construir una regla de clasificación basada en un conjunto de datos de entrenamiento, que contiene instancias etiquetadas, para que esta regla pueda clasificar nuevas observaciones futuras en una de las clases conocidas (o grupos). En ambos casos, el método tiene varias propiedades atractivas, por ejemplo su simplicidad conceptual y computacional. Sin embargo, la definición clásica del método, que es la que veremos ahora para entenderlo, se basa en la media empírica y el estimador muestral de la o las matrices de covarianza, que son altamente susceptibles a la presencia de atípicos en los datos.

Esto hace que la regla sea inapropiada en conjuntos de datos contaminados aunque haya un solo valor atípico ya puede darnos un resultado completamente diferente e incorrecto. Y si la regla está mal calculada esto hace que sea más probable la clasificación errónea de las observaciones futuras. Una solución propuesta por varios autores en la literatura de esta área, yo incluida, es considerar estimadores robustos para la media y la matriz de covarianza en la definición de la regla. Pero esto ya sería entrar en el tema de análisis robusto de datos, detección de atípicos y outliers para lo cual tienen un curso completo sobre todo eso.

Vamos a ver el enfoque tradicional. En ambos casos LDA y QDA asumimos que las probabilidades a priori, las medias y la matriz o matrices de covarianza son desconocidas. Y lo que se hace como ya he mencionado anteriormente es estimar todas esas cosas con la muestra de datos. De ahí la importancia de que sea una muestra limpia o se utilicen estimadores que sean robustos a la presencia de atípicos. Entonces recordemos cuáles son los estimadores muestrales para la media y la matriz de covarianza, teniendo en cuenta que estamos en el espacio multivariante, y no en el univariante, es decir, estamos considerando todas las variables explicativas a la vez. Para cada grupo k tenemos que estimar la probabilidad a priori, que se estima con el porcentaje de datos muestrales que pertenecen al grupo k:

$$\hat{\pi}_k = \frac{|\{i: y_i = k\}|}{n}$$

Y las estimaciones del vector de medias y de la matriz de covarianza en el caso de LDA serían:

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|\{i: y_i = k\}|} \sum_{i:y_i=k} \mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=0}^{K-1} \sum_{i:y_i=k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

Donde lo anterior se llama matriz de covarianza agrupada, porque es como un promedio de las estimaciones de covarianza entre todos los grupos. Para poder obtener solo una estimación de matriz de covarianza que es lo que necesitamos en LDA. En QDA esto no es necesario y la estimación sería para cada grupo k y sin hacer el promedio.

La regla discriminante bayesiana asigna la observación $\mathbf{x} \in \mathbb{R}^p$ al grupo k para el cual el $\ln(\pi_K f_k(x))$ se maximiza. Donde f_k que es la distribución de los datos del grupo k se asume que sigue una distribución Normal. Por ejemplo en el caso de QDA:

$$f_k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

La regla se reduce a maximizar en el caso de QDA:

$$d_k^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k)^T \hat{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_k) + \ln(\hat{\pi}_k)$$

Y en el caso de LDA:

$$d_k^L(\mathbf{x}) = \hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1}\mathbf{x} - \frac{1}{2}\hat{\boldsymbol{\mu}}_k^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_k + \ln(\hat{\pi}_k)$$

El método clasifica a la nueva observación $\mathbf{x}$ en el grupo m si $d_m^Q(\mathbf{x}) > d_k^Q(\mathbf{x})$ para todo $k \neq m$. En el caso de LDA lo mismo pero con $d_m^L(\mathbf{x})$.

Ejemplo:

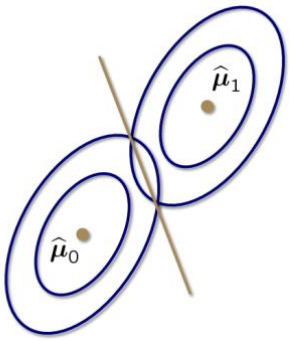
Imagina que tenemos dos clases o dos grupos, por ejemplo, perros y gatos. Entonces $K = 2$. Imagina que tenemos una variable dependiente Y que nos dice a qué grupo pertenece cada observación de mi conjunto de datos de entrenamiento. Imagina que medimos tres variables explicativas o características sobre los datos que tenemos, por ejemplo, la altura del animal, el peso, y el tamaño de la cabeza. Imagina que nos interesa aplicar LDA para hallar una regla entrenada con nuestros datos que nos permita clasificar a las nuevas observaciones en uno de los grupos, es decir, que nos llegue información sobre un nuevo animal y quisiéramos clasificarlo como gato o como perro.

Entonces lo primero que hacemos es hallar qué porcentaje de mis datos de entrenamiento pertenece a cada grupo y esas serán mis probabilidades a priori $\hat{\pi}_0$ y $\hat{\pi}_1$. Es decir, qué proporción de gatos tengo en mis datos y qué proporción de perros. Luego tendríamos que estimar el vector de medias y la matriz de covarianza según las fórmulas que vimos anteriormente, teniendo en cuenta que cada fila o cada dato es multidimensional porque está caracterizado por tres variables como dijimos antes.

Si quieres más información sobre estos métodos y muchos otros más, así como si quisieras afianzar mejor estos conceptos que estamos viendo por arriba, del espacio multivariante, te aconsejo que veas el curso de estadística multivariante.

Volviendo al tema. Una vez que tenemos las tres cosas estimadas, junto a $\mathbf{x}$ que es el nuevo dato, procedemos a sustituir todo en la ecuación de la regla y estimar $d_k^L(\mathbf{x})$ para cada grupo, es decir, obtendremos una estimación para el grupo de los perros $d_0^L(\mathbf{x})$ y otra para el grupo de los gatos $d_1^L(\mathbf{x})$ y en la que tenga mayor valor, procederemos a clasificar la nueva observación en ese grupo.

Visualmente se podría interpretar el problema de esta manera, tenemos dos grupos y lo que hace la regla es hallar la separación lineal entre ellos tal que si mi nueva observación está del lado izquierdo pertenecería al grupo de la izquierda y si está en la derecha pues al derecho.

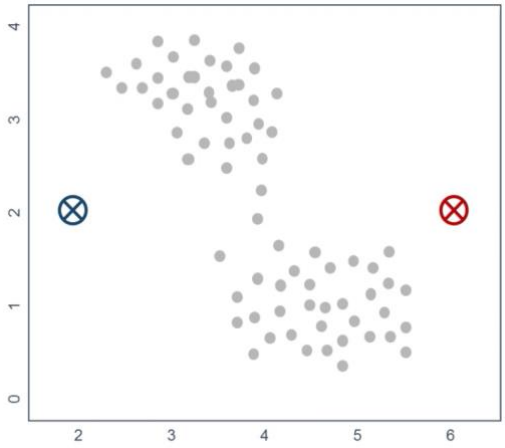


En el caso cuadrático como su nombre lo indica la separación no es lineal sino cuadrática:

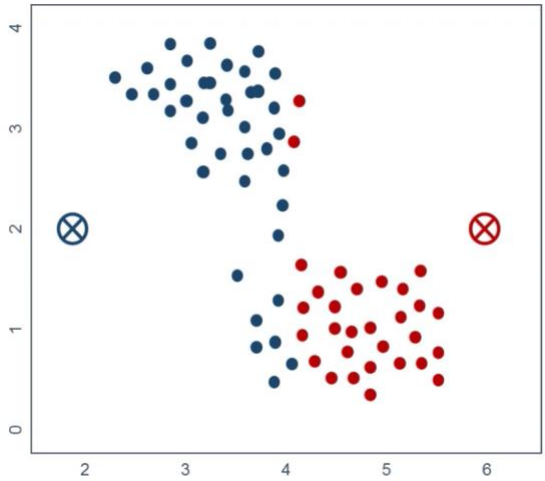
MODELOS DE MIXTURA GAUSSIANOS (GAUSSIAN MIXTURE MODELS)

Anteriormente hablamos de clasificación supervisada, ahora vamos a hablar de clasificación no supervisada lo que se conoce como análisis clúster, donde el objetivo es el mismo, es agrupar los datos en grupos diferentes para poder clasificar en alguno de esos grupos las observaciones futuras pero ahora no disponemos de la pertenencia a los grupos de nuestros datos de entrenamiento.

Antes de ver el algoritmo que se basa en el enfoque bayesiano vamos a recordar cómo funciona el algoritmo más básico y más usado del análisis frecuentista en el análisis clúster que es el algoritmo de K-means, para luego entender la diferencia con el bayesiano. Supongamos que tenemos estos datos de aquí de color gris porque aunque veamos visualmente que forman dos grupos, no sabemos cuáles observaciones exactamente pertenecen a cada grupo.



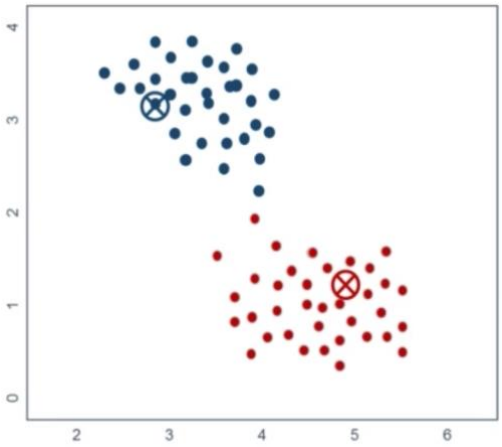
Lo que nos gustaría hacer ahora es segmentar esos datos en dos o más grupos de tal manera que podamos encontrar algún tipo de estructura.



El resultado ideal de este procedimiento debería ser que encontremos dos grupos diferentes, uno en la esquina superior izquierda que muestra valores bajos de la variable en el eje horizontal y valores altos de la del eje vertical, y otro grupo en la esquina inferior derecha con valores altos de la variable en el eje horizontal y valores bajos de la del eje vertical,

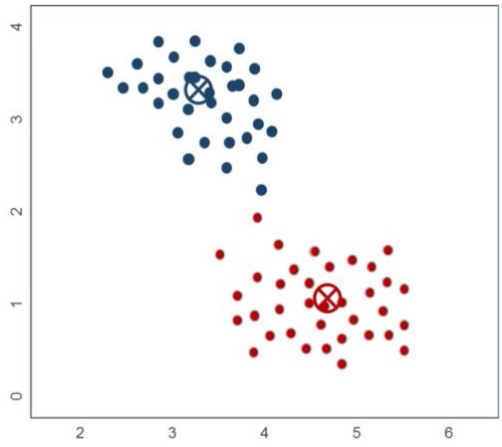
Entonces, un enfoque clásico para agrupar un conjunto de datos es el llamado algoritmo de agrupación de K means o K medias, donde K describe el número de grupos que nos gustaría encontrar entre los datos. Es decir, de antemano tenemos que hacer una suposición ciega de qué valor puede tomar K, que sería dos en este caso.

La forma en la que el algoritmo procede a hacer la agrupación es poner dos centros de agrupación aleatorios que son ese círculo con una cruz adentro, uno azul y otro rojo. Cada uno de esos centros va a caracterizar a un grupo. Y a continuación procede a asignar cada punto de datos al grupo más cercano, que sería esta clasificación:



Es decir, mide las distancias de cada punto de mis datos hacia los centros y le asigna el que esté más cerca entre los dos. Por eso aquí hay una división entre datos azules en la parte izquierda y datos rojos en la derecha. Posteriormente, el centro del clúster o grupo, se vuelve a calcular haciendo el promedio de todos los puntos de datos que se asignan a cada grupo. Entonces, el centro del grupo azul se moverá hacia arriba y el centro del clúster rojo, se moverá hacia abajo. Es decir, ambos se moverán más hacia el centro de su grupo correspondiente.

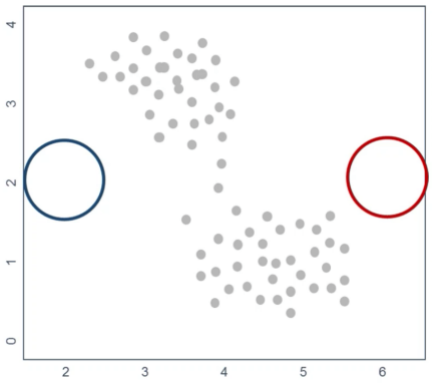
A continuación, asignamos nuevamente cada punto de datos al centro del clúster más cercano. Y repetimos de nuevo el procedimiento de mover el centro del clúster hacia el centro de todos los puntos. Y volver a asignar los puntos al centro, y así sucesivamente, hasta que no veamos un cambio en varias iteraciones, lo que significaría que ha convergido y hemos terminado.



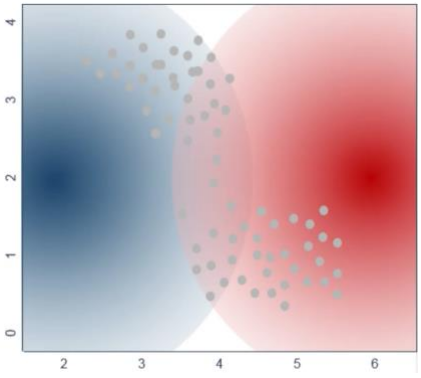
Gaussian Mixture Models (GMM):

Ahora vamos a ver a continuación, el algoritmo de agrupación en clústeres llamado Gaussian Mixture Models, que se basa en estadísticas bayesianas. Este algoritmo sigue un procedimiento similar a la agrupación por k-medias, pero hay una diferencia clave. Recordemos que en K means, asignamos cada punto de datos en cada interacción al centro de clúster más cercano. Entonces, hacemos una asignación de cada punto de datos a uno y solo un centro del clúster. Por el contrario, la agrupación que hace el método de Gaussian Mixture Model es una agrupación más suave en el sentido de que observa cuál es la probabilidad de que un punto de datos pueda pertenecer a un centro de grupo en comparación con el otro grupo. Esto nos flexibiliza en el sentido de que cada dato puede estar en cualquier grupo pero si tenemos calculadas las probabilidades de pertenencia a cada grupo, lo asignaremos al grupo en el que tenga mayor probabilidad de pertenecer.

Para calcular esta probabilidad, no solo definimos un único centro de clúster como un único punto de datos, sino que definimos una distribución para cada uno de los grupos.



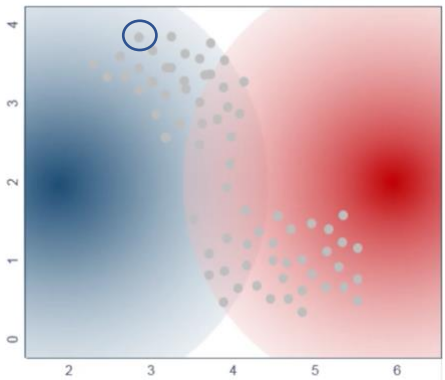
Así que para empezar, iniciamos de manera similar, con una media aleatoria, donde estará el centro de esta distribución y una varianza que me diga cómo deberían dispersarse los datos en cada grupo. Además, asignamos a cada grupo la probabilidad a priori, donde nuestra mejor suposición al principio, cuando hablamos de dos grupos, sería asignar a cada grupo una probabilidad a priori del 50%.



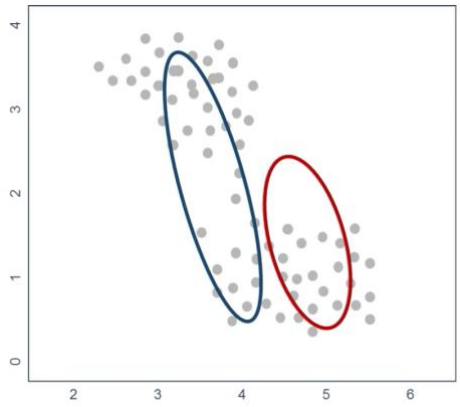
Una vez que hemos inicializado medias y varianzas aleatorias y la probabilidad a priori de pertenencia a cada grupo, podemos calcular cuál es la probabilidad de pertenencia a cada grupo para todos los datos. Como hemos visto anteriormente esto lo podemos calcular haciendo uso del Teorema de Bayes:

$$P(dato_i \in grupo_j | valores)$$

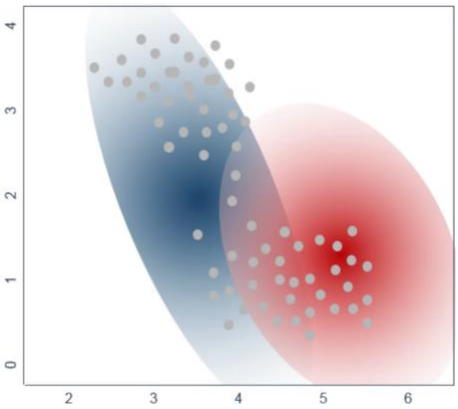
Para ser precisos, para cada dato i tendríamos calculada la probabilidad de pertenecer al grupo 1 y la probabilidad de pertenecer al grupo 2, en nuestro caso. Supongamos que para este dato de aquí la probabilidad de pertenecer al grupo 1 (azul) es 90% y al grupo 2 (rojo) de 10%.



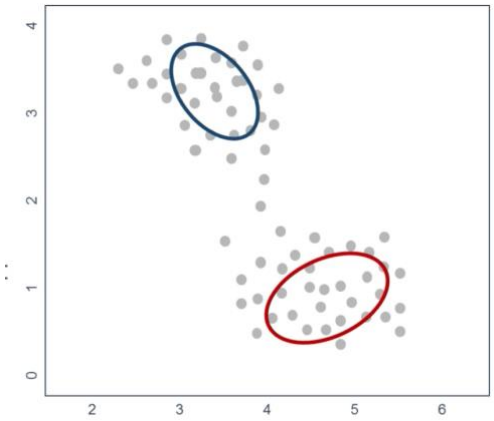
A esas probabilidades se les llamará responsabilidades (responsabilities) en el algoritmo de GMM. Entonces el siguiente paso que haremos, similar a k medias, es recalcular la media de las distribuciones, las varianzas y los priors. El prior nuevo actualizado será la probabilidad promedio sobre todos los puntos de datos de que un punto de dato pertenezca al clúster 1 o 2. Para las medias, el valor actualizado sería el valor promedio ponderado basado en las probabilidades de pertenencia al grupo 1 o 2. Y un procedimiento similar se usa para volver a estimar la varianza.



Ahora que hemos vuelto a estimar las distribuciones, volvemos a realizar la misma operación otra vez.

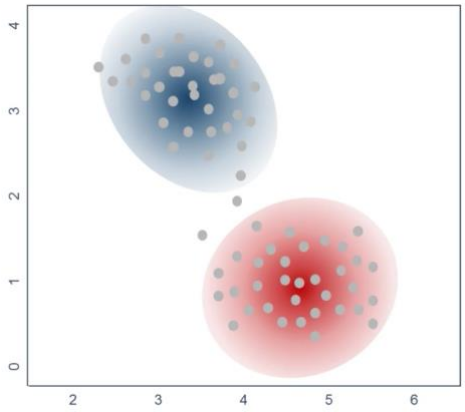


Es decir, calculamos nuevamente la probabilidad de que cada dato provenga de cada grupo y con ello actualizamos las medias, las varianzas, y los priors para luego pasar a la siguiente iteración. Y al igual que en K means, iteramos este procedimiento hasta que encontremos algún tipo de convergencia en el sentido de que las estimaciones que son las medias, las varianzas y los priors no cambien de manera significativa, durante un período largo de iteraciones.



Y una vez que ocurre esta convergencia, podemos incluso dibujar las elipses que rodean a los datos de cada grupo, usando la media y la varianza del ultimo paso.

Entonces así es como funciona este método desde el enfoque bayesiano. La diferencia clave con respecto al método frecuentista de K means es que este usa una asignación fuerte de cada dato a un grupo en concreto mientras que el GMM funciona con una asignación blanda donde esta asignación se basa en la probabilidad de pertenencia de cada punto de datos a cada clúster, dados sus valores.



MATERIAL COMPLEMENTARIO

Como material complementario o de soporte te recomiendo ver la Ruta de Aprendizaje de Aprende con Eli, donde puedes encontrar todos mis cursos de Estadística y Análisis de Datos en descuento. También te recomiendo pasarte por el blog donde hablamos y discutimos muchas de las cuestiones abordadas en este libro y durante el curso.