



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

Propuesta de proyecto:

Sleepy Bayesian

Fernanda Weiss - 18406890

Ignacio Espinoza - 18406882

**IIC3695 Tópicos Avanzados en
Inteligencia de Máquina**

Profesor: Karim Pichara Baksai

**Ayudantes : Francisco Pérez Galarce,
Orlando Vásquez**



1. Definición problema

Descansar lo suficiente es importante para llevar una vida saludable, pues durante el sueño el cuerpo realiza procesos importantes que nos permiten mantenernos sanos, por ejemplo se relaja la columna, se restauran los músculos, se regulan los niveles de azúcar y apetito, entre otros efectos¹. Por lo tanto, malos hábitos de sueño, como dormir poco, dormir mucho, o dormir sin un patrón estable de sueño aumenta el riesgo de sufrir enfermedades de distinta gravedad, desde enfermedades leves como mayor probabilidad de desarrollar un resfriado hasta enfermedades crónicas como problemas al corazón, diabetes, obesidad, etc². Además, esta actividad nos ayuda en otras tareas. Por ejemplo, el cerebro fija los recuerdos y lo aprendido durante el día, repone del cansancio para poder realizar las funciones óptimamente como concentración, pensar con claridad y reaccionar rápido, pues hemos visto accidentes de tránsito que ocurren por la falta de sueño.

Debido a lo explicado anteriormente es que consideramos importante que cada persona pueda obtener su patrón de sueño, pues si presenta alguna enfermedad con esta información un especialista tendría más antecedentes para detectar dicha enfermedad y también para saber cómo enfrentarla. No obstante, de igual modo esta información puede ser útil para que uno haga uso de ella, pues podrá regular cuánto duerme cada día y cuándo le es más provechoso dormir. Esta información para estudiantes como nosotros sería beneficiosa, ya que al no tener una rutina clara con los diferentes horarios y quehaceres no hay un claro patrón de sueño, lo que claramente nos lleva a enfermarnos, y utilizando los resultados de este estudio uno se podría organizar mejor para cuidar nuestra salud.

El tracker de actividad inteligente Vivosmart HR posee la capacidad de medir la frecuencia cardiaca en la muñeca las 24 horas del día, durante los 7 días de la semana. Entrega información de calorías quemadas y cuantifica la intensidad de las actividades físicas. Estudiando el ritmo cardíaco, junto con el trackeo del movimiento, se puede determinar cuando una persona se queda dormida y se despierta.

El proyecto está enfocado al estudio de datos del sueño para crear un modelo que determine la probabilidad a posterior de que una persona se duerma en un determinado horario. A su vez, determinar la probabilidad a posterior de que una

¹ <http://www.bbc.com/mundo/noticias-41811949>

² <https://hipertextual.com/2013/10/efectos-sueno-corazon-diabetes>



persona se despierte. Ambas probabilidades se determinan en base al historial de sueño de una persona en específico.

Como el tiempo es una variable continua, calcular la distribución a posterior a cabalidad no es posible analíticamente. Por esto es necesario utilizar métodos para simular la distribución que modela el tiempo. Dentro de esos métodos se encuentran los algoritmos de *Markov Chain Monte Carlo (MCMC)*. MCMC son métodos numéricos usados para calcular, aproximar y simular expresiones o sistemas matemáticos complejos y difíciles (a veces imposibles) de evaluar.

2. Características de los datos

Los datos a utilizar fueron recopilados por William Koehrsen³ por medio de su trackeador VivoSmart. Las mediciones cardíacas fueron transformadas en entradas que indican la hora y fecha en que fueron tomadas, una variable booleana que indica si la persona estaba dormida o despierta y un valor de *offset* indicando la cantidad de minutos de diferencia con la hora habitual en que él se duerme (10 p.m.) y con la hora que habitualmente se despierta (6 a.m.).

Se trabajará el dataset `sleep_wake.csv` que cada una de sus entrada marca el minuto, desde las horas habituales de dormir y despertar mencionadas anteriormente, en la que el tracker detecta que la persona se durmió y el momento en que esta despertó. Además, se registra el día de la medición. En total el dataset tiene 65 registros, donde cada uno representa un día. Por lo tanto, esta información se debe procesar para obtener datos manejables entre las transiciones despierto-dormido y dormido-despierto. Así, se crearán dos datasets: uno que describa los datos cuando la persona se va a dormir y el segundo cuando la persona se vaya a despertar.

Puesto que se manejan pocos datos será difícil que los modelos a entrenar puedan predecir bien datos futuro pues no podrán generalizar bien y existirá overfitting. La solución a esto se hará un preprocesamiento de datos que permita agregar información real a cada dataset. Se tomará como punto de inflexión la hora común de ir a dormir (10 p.m) para la transición despierto dormida y como hora común de despertar (6 a.m) para la transición inversa. Con este punto de inflexión se generarán tantas filas por minutos, en un intervalo de tiempo alrededor a este, para analizar cuánto es que varía su comportamiento de sueño, así se podrán obtener distintos dataset. Por ejemplo, solo analizar +/- 1 hora desde el punto de inflexión, y por otro lado analizar +/- 2 horas desde el punto de inflexión. Cada fila luego del

³ <https://medium.com/@williamkoehrsen>

preprocesamiento de los datos de entrada indican la hora y fecha en que fueron tomadas, una variable booleana que indica si la persona estaba dormida o despierta y un valor de offset indicando la cantidad de minutos de diferencia con la hora habitual en que él se duerme (10 p.m.) y con la hora que habitualmente se despierta (6 a.m.). Así, al agregar +/- 1 hora se tendrán 15.600 para cada dataset (31.200 en total), y con +/- 2 horas 31.200 datos para cada dataset (62.400 en total).

3. Modelos a utilizar

Puesto que los datos son binarios, y buscamos una transición suavizada entre los dos estado (durmiendo y despierto), se debe encontrar una función que modele la probabilidad de transición entre estados. Para resolver esta tarea postulamos una función logística⁴, cuyo dominio es $[0,1]$ con una transición lineal entre ambos límites. En ambos extremos tiene colas pesadas que aseguran que datos muy alejados tomen probabilidades aproximadas a cero (por la izquierda) como aproximadas a uno (por la derecha), como lo muestra la Imagen 1.

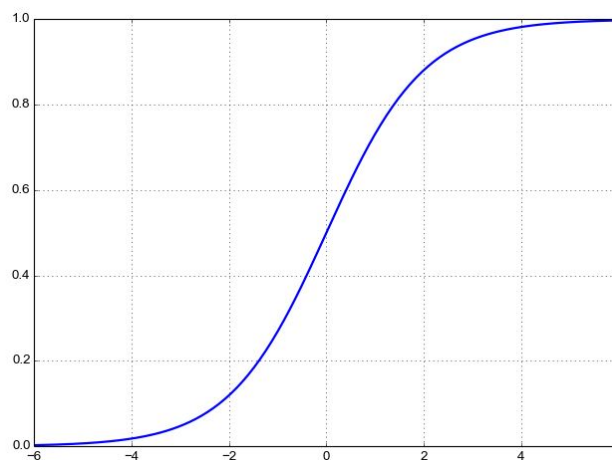


Imagen 1: Función Logística

Dado que los parámetros de esta función tienen que ser adaptados a los datos, debe haber un proceso de aprendizaje. Para este proceso se probará con tres diferentes algoritmos de MCMC para simular parámetros que mejor describan los datos. Estos algoritmos nos permitirán sacar muestras tanto de la distribución a posteriori de que una persona se duerma o se despierte. Así, con dichas muestras se podrá estimar el valor óptimo de los parámetros.

⁴ https://en.wikipedia.org/wiki/Logistic_function



4. Evaluación de los resultados

Se comparará los resultados de los tres algoritmos a utilizar: Metropolis Hastings, Gibbs Sampling, Simulated Annealing y Slice Sampling, los cuales serán evaluados bajo el criterio de convergencia a su distribución estacionaria, que será calculado mediante el método Gelman-Rubin, Running means, y mediante métodos gráficos.

Primero, se realizará un pre-procesamiento de los datos junto con su visualización para comprender la naturaleza de estos. Luego, se realizará un conjunto de pruebas para cada algoritmo usando diferentes configuraciones para generar las muestras. Se visualizarán y analizarán las muestras obtenidas, para determinar los distintos efectos que pueden tener los parámetros sobre los modelos de los datos. Finalmente, se probará la convergencia de cada algoritmo con los métodos anteriormente nombrados.