Universidad Zaragoza

Trabajo Fin de Grado

Un estudio de asociación genómica basado en aprendizaje automático para la caracterización de la enfermedad de Alzheimer

A genomic-wide association study based in machine learning for the characterization of Alzheimer's Disease

Autor

Eduardo Alonso Monge

Directoras

Elvira Mayordomo Cámara
Mónica Hernández Giménez

Ingeniería Informática - Computación

Escuela de Ingeniería y Arquitectura
2020

Resumen

En el estudio reciente de imágenes genéticas se está haciendo uso de la relación entre polimorfismos de nucleótido único (SNP) y rasgos cuantitativos (QT). Para llevarlos a cabo se utilizan técnicas de selección de variables, como Lasso (least absolute shrinkage and selection operator, por sus siglas en inglés), tratando de seleccionar los SNPs que están más relacionados a los QT. En este Trabajo de Fin de Grado se va a comparar Lasso con distintas variantes, cómo Elastic net y Group Lasso, para tratar de resolver ciertas limitaciones que presenta Lasso. Además, se va a hacer una comparación entre dos implementaciones distintas para Lasso y Elastic net, la técnica de descenso de coordenadas y la técnica de búsqueda lineal de Armijo con descendo de gradiente. El método de Group Lasso permite crear una estructura jerarquizada de los datos de tal forma que se pueda generar un árbol genético. Se presentarán los distintos métodos con la intención de identificar asociaciones entre variantes genéticas (SNPs) y medidas derivadas de volúmenes del hipocampo en imágenes de resonancia magnética (MRI). Los experimentos realizados se llevarán a cabo con datos extraídos de la base de datos ADNI (Alzheimer's Disease Neuroimaging Initiative), que contiene una gran cantidad de datos, tanto genéticos como imágenes MRI, de sujetos de Estados Unidos y Canadá en diferentes fases de la enfermedad de Alzheimer, lo que permite corroboar con un nuevo estudio que el gen APOE está altamente relacionado con la enfermedad.

Índice

Capítulo 1

Introducción

La enfermedad de Alzheimer[2] es un trastorno progresivo que hace que las células del cerebro se degeneren y mueran. El Alzheimer es la causa más común de demencia, una disminución continua de las habilidades cognitivas y sociales que altera la capacidad de una persona para desempeñarse de manera independiente. La enfermedad se manifiesta como un deterioro cognitivo y no existe actualmente un tratamiento que pueda curarla, pero detectarla a tiempo puede ayudar disminuir la velocidad del deterioro. Suele aparecer a partir de los 60 años, tan solo el 10 % de personas con Alzheimer la padecen antes, y se estima que a partir de los 85 años hasta un 50 % de la población puede padecer algún tipo de deterioro cognitivo que acabe degenerando en Alzheimer. Algunas personas mayores sufren una condición denominada deterioro cognitivo leve (Mild Cognitive Impairement, MCI por sus siglas en inglés), esto puede ser una señal temprana de Alzheimer, pero no todos los que sufren MCI desarrollarán la enfermedad. La principal diferencia entre una persona con MCI de una con Alzheimer es que puede seguir actuando en entornos laborales o sociales de forma autónoma.

El Alzheimer es una enfermedad que el riesgo de sufrirlo aumenta considerablemente si algún familiar la ha tenido previamente, aunque la mayoría de la gente que la padece no es debido a una herencia genética. Un factor genético altamente conocido es una forma del gen apoliproteina E, una variante del gen APOE que aumenta considerablemente el riesgo de padecer la enfermedad, aunque no todas las personas con dicha variación la desarrollan. En los últimos años se han descubierto mutaciones en 2 genes más, ABCA7 y SORL1, lo que en muchas ocasiones garantizan que la persona que herede dichas mutaciones padecerá la enfermedad. Sin embargo estas mutaciones genéticas son raras, dado que están presentes en menos del 1 % de las personas con Alzheimer.

Tanto personas con Alzheimer, como personas con MCI y personas sanas pueden participar en estudios clínicos para ayudar a combatir la enfermedad. Por ejemplo, la Iniciativa de Neuroimagen de la Enfermedad de Alzheimer[6] (Alzheimer's Disease Neuroimaging Initiative, ADNI) conforma un consorcio de universidades y centros médicos en los Estados Unidos y Canadá establecidos para desarrollar técnicas estandarizadas de imagen y procedimientos de biomarcadores en sujetos normales, sujetos con MCI y sujetos con AD (Alzheimer Disease). Los principales objetivos de ADNI son desarrollar métodos que conduzcan a estándares uniformes para la adquisición de datos longitudinales, de imágenes por resonancia magnética (MRI) y tomografía por emisión de positrones (PET) para desarrollar un repositorio de datos

accesible que describa los cambios longitudinales en la estructura del cerebro y el metabolismo mientras se obtienen datos clínicos, cognitivos y bioquímicos de forma paralela para desarrollar métodos que puedan determinar los efectos del tratamiento en ensayos clínicos, y, para probar una serie de hipótesis basadas en datos clínicos y de biomarcadores. Entre los datos del repositorio de ADNI se pueden encontrar a participantes de entre 55 y 90 años que son obtenidos en 57 lugares diferentes de los Estados Unidos y Canadá, a los que se evalúa a través de pruebas periódicas. Existen distintos conjuntos de datos que han sido creados a lo largo de los años, inicialmente, el conjunto ADNI-1 constaba de 200 participantes sanos, 400 participantes con MCI y 200 participantes con AD. ADNI-GO, ADNI-2 y ADNI-3 agregaron participantes adicionales para aumentar la cohorte, para un tamaño de cohorte final de más de 1000 participantes. Hasta la fecha, más de 1000 publicaciones científicas han utilizado datos de ADNI. Lleva existiendo desde 2004 y está financiado hasta el año 2021.

En genética, un estudio de asociación del genoma completo (Genome-wide association study, GWAS por sus siglas en inglés) es un análisis de cada variación genética a lo largo de todo el genoma humano con el objetivo de poder asociarlas a un fenotipo (asma, cáncer, diabetes, etc.). Los GWAS suelen centrarse en asociaciones entre los polimorfismos de un solo nucleótido (SNPs) y enfermedades de interés en salud pública, como el Alzheimer. Un SNP es una variación genética en la secuencia del ADN que afecta a una sola base (adenina (A), timina (T), citosina (C) o guanina (G)) de una secuencia del genoma. Para que la variación sea considerada como SNP debe darse en al menos un 1 % de la población, si no es considerada como una mutación puntual. Se dice que una variación está asociada a una enfermedad cuando es más frecuente en una población con un tipo de enfermedad determinada. Para ello se deben estudiar una gran cantidad de individuos, de modo que se puedan comparar datos genéticos entre ellos. Estas variaciones han permitido descubrir que ciertos genes están asociados a enfermedades, como el caso del APOE y el Alzheimer.

Los GWAS presentan ciertas limitaciones, entre los problemas más comunes se encuentran la difícil selección de individuos con respecto a la enfermedad que se quiere estudiar o la insuficiencia del tamaño de la muestra. Estos estudios además generan una gran cantidad de falsos positivos, por lo que se requieren sistemas informáticos mas generales y flexibles que los GWAS y que sean capaces de manejar correctamente la gran cantidad de información que hay que procesar. Es aquí donde entran en juego los sistemas analíticos y de aprendizaje automático.

El estudio de datos biológicos está creando un gran interés en el campo de la investigación, buscando continuamente nuevos métodos para almacenar, manipular y tratar datos que, durante los últimos años, están mostrando un crecimiento exponencial. Las técnicas de aprendizaje automático sirven como métodos analíticos para extraer conocimiento de los datos, permitiendo crear modelos para detectar patrones como estructuras de proteínas, cadenas de genomas, mutaciones de ADN, etc. Sin embargo, los estudios son complejos dada la gran cantidad de datos heterogéneos que se generan en los procesos biológicos. También puede haber fallos en fases previas al análisis, desde posibles pruebas contaminadas, fallos en la extracción física de los datos biológicos, en los métodos de secuenciación, etc. Es por ello que los estudios que se realizan deben tener en cuenta todos estos factores a la hora de desarrollar nuevas aplicaciones y algoritmos.

Numerosos estudios de bioinformática están basados en seleccionar característi-

cas biológicas y determinar su importancia dentro de un gran número de muestras. La selección de características (feature selection) es un proceso en el cual se busca reducir el número de características para mejorar la calidad del resultado. Lo que se intenta hacer con este método es desechar aquellas características que son redundantes o irrelevantes. Otro método generalmente utilizado para evaluar la correlación entre múltiples variables es el conocido como *ridge regresison*[7].

El objetivo de este Trabajo de Final de Grado se basa en utilizar una técnica de selección de variables conocida como Lasso[10] (Least Absolute Shrinkage and Selection Operator, por sus siglas en inglés), tratando de seleccionar los SNPs que están mas relacionados con la enfermedad del Alzheimer a través de medidas derivadas de volúmenes del hipocampo en imágenes por resonancia magnética (MRI), basado en los árticulos [8] y [12]. Además se va a comparar dicha técnica con dos variantes, conocidas como Elastic net[3] y Group Lasso[5], para tratar de resolver ciertas limitaciones que presenta Lasso. Tanto para Lasso como Elastic net se va a hacer una comparación entre dos implementaciones diferentes para cada método, la técnica de descenso de coordenadas y la técnica de búsqueda lineal de Armiijo con descenso de gradiente acelerado, a excepción de Group Lasso que solo se implementará con la última técnica citada.

La técnica de Lasso utiliza la norma de regularización L1 para poder identificar un conjunto de variables (en el caso de este estudio, SNPs) para su posterior análisis de asociación genética. Por otra parte, Elastic net combina las normas de regularización L1 y L2 para abordar el problema de la alta dimensión de las variables del estudio. Por último, Group Lasso utiliza la norma de regularización L2, y resulta altamente interesante en este trabajo dado que permite agrupar los SNPs de forma jerárquica, permitiendo crear una estructura arborescente que simula un árbol genético agrupado por genes, en el que cada hoja se identifica como un SNP perteneciente a uno de los genes que se van a utilizar en el estudio.

Este trabajo utiliza los conjuntos de datos de los repositorios de ADNI-1 y ADNI-2, lo que aumenta la cantidad de los mismos y permite comparar los resultados con mas exhaustividad entre sí y con los estudios en los que se basa este trabajo, obteneniendo de esta manera unas conclusiones más precisas y relevantes.

En los resultados finales obtenidos se observa cómo todos los métodos generan unos resultados similares entre ellos, especialmente Lasso y Elastc net. Entre las distintas implementaciones existe más variación en los resultados que entre los métodos, debido a las características propias de cada una de ellas. Cabe destacar que aparece un mismo SNP seleccionado perteneciente al APOE, independientemente de los distintos conjuntos de datos, métodos de aprendizaje, técnicas de regresión y regiones del hipocampo que han sido analizados, lo que reafirma que el gen APOE está altamente relacionado con el Alzheimer, como muchos otras investigaciones y GWAS han demostrado previamente. Otro dato interesante extraído es que los genes ABCA7 y EPHA1 parecen estar altamente relacionados con el hipocampo derecho, al ser seleccionados también entre los SNPs más frecuentes en los distintos métodos e implementaciones en dicha región.

Capítulo 2

Métodos

2.1. Lasso

Tibshirani introdujo Lasso[10] (least absolute shrinkage and selection operator) como un método estadístico de análisis mediante una regresión lineal que minimiza la suma residual de cuadrados. El método está sujeto a que la suma del valor absoluto de los coeficientes sea menor que una constante. La naturaleza de dicha restricción crea una tendencia en el método para proporcionar coeficientes que son exactamente cero, por lo que se suele utilizar Lasso como un método de selección de variables.

Suponiendo que se tienen unos datos $(x_i, y_i),\ i = 1, 2, ..., N$ donde $x_i = (x_{i1}, ..., x_{iK})$ representa las variables de entrada o predictores, e y_i las respuestas. Se puede denotar X como la matriz de entrada N x K, e $Y = (y_1, ..., y_N)^T$ como el vector de salida de dimensión N. Un modelo de regresión lineal puede ser formulado como:

$$\boldsymbol{Y = X\theta + \epsilon}$$

donde $\boldsymbol{\theta}$ es el vector de coeficientes asignados a cada variable y $\boldsymbol{\epsilon}$ un término de error. Para tratar la dispersión de las variables se utiliza la norma L_1 para regularizar el vector de los coeficientes de la regresión, quedando la ecuación de la siguiente manera:

$$\boldsymbol{\theta = argmin\ ||Y - X\theta||^2 + \lambda||\theta||_1}$$

donde $\boldsymbol{\lambda}$ es un parámetro de regularización que controla la dispersión de la solución. Los elementos iguales a cero en $\boldsymbol{\theta}$ indican que dichas variables son irrelevantes y pueden ser desechadas.

Esto se puede ver como la minimización de dos términos: $\boldsymbol{OLS + L_1}$.

- El primer término, denominado mínimos cuadrados ordinarios (OLS) se puede escribir como $(y - X\theta)^T(y - X\theta)$ que da lugar a una elipse centrada alrededor del estimador de máxima verosimilitud (Maximum Likelihood Estimator, MLE).
- El segundo término L_1 es la ecuación de un diamante centrado alrededor de 0 (o un romboide en dimensiones más altas)

- La solución a la optimización con restricciones se encuentra en la intersección entre los contornos de las dos funciones, y esta intersección varía en función de λ. Para $\lambda = 0$ la solución es el MLE y para $\lambda = \infty$ la solución está en [0,0].

- Dado que en los vértices del diamante, una o muchas de las variables tienen un valor 0, existe la probabilidad de que una o muchas de las características tengan un valor exactamente igual a 0.

Intuitivamente, a medida que aumentamos λ de 0 a ∞, se espera que la solución de Lasso se mueva de la solución OLS a la solución L_1 que es 0. En la figura 2.1 se pueden ver la explicación de forma gráfica.

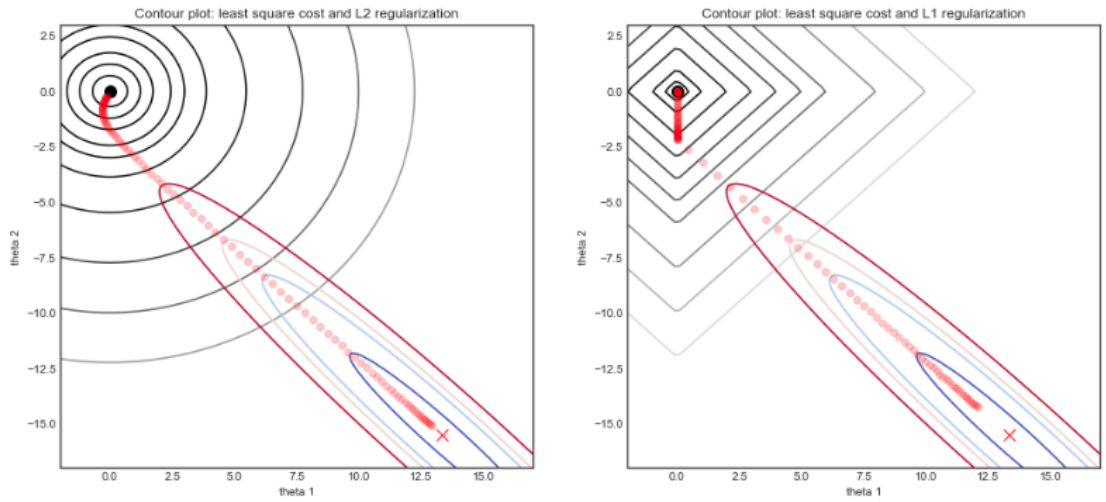


Figura 2.1: Formas de las regiones de restricción para ridge regression y lasso

2.1.1. Implementaciones del método Lasso

A continuación se van a presentar las dos implementaciones distintas que se van a utilizar en este Trabajo de Fin de Grado.

- La primera implementación trata del algoritmo de descenso de coordenadas, implementado en Python a través de la biblioteca Scikit. En este repositorio público no existe ninguna implementación de Group Lasso por lo que solo se estudiarán los resultados obtenidos de Lasso y Elastic net.

- La segunda implementación se basa en el algoritmo de búsqueda lineal de Armijo con un descenso de gradiente acelerado. El código de la implementación proviene del repositorio público de los autores del árticulo original en el que se basa este trabajo [12].

Algoritmo descenso de coordenadas

Para la implementación de Lasso en Python se utiliza una técnica llamada descenso de coordenadas (coordinate descent), que está basada en los métodos de optimización mediante sub-gradientes. En cada paso del algoritmo se actualiza cada componente de forma individual, de forma ordenada e independiente, en lugar de aplicarlo sobre todas ellas de forma conjunta, minimizando la función de coste con

respecto a cada coordenada.

Algorithm 1 Coordinate descent algorithm

Initialize: $\theta = 0$
while not converged **do**
  for $j = 0$ **to** K **do**
    compute: $p_j = \sum_{i=1}^{N} x_j(y_j^{(i)} - \hat{y}^{(i)} + \theta_j x_j^{(i)})$

    set: $\theta_j = S(p_j, \lambda)$
  end for
end while

La función de coste de Lasso tiene una solución cerrada en el caso especial del descenso de coordenadas por ser un problema de optimización de una sola variable. La solución para datos normalizados se define en términos de la función de umbral $\boldsymbol{S(p_j, \lambda)}$ (soft threshold function)

$$\boldsymbol{S(p_j, \lambda)} = \begin{cases} \frac{p_j+\theta}{z_j} & \text{si } p_j < -\lambda \\ 0 & \text{si } -\lambda \leq p_j \leq \lambda \\ \frac{p_j-\theta}{z_j} & \text{si } p_j > \lambda \end{cases} \tag{2.1}$$

donde λ representa el parámetro de regularización y $\boldsymbol{p_j}$ representa la diferencia entre el resultado real y el resultado predicho, considerando todas las componentes excepto la j-ésima. Si el resultado es pequeño, significa que el algoritmo es capaz de predecir el resultado sin la j-ésima componente y por tanto puede ser desechada de la ecuación estableciendo su coeficiente en cero.

Algoritmo búsqueda lineal (Armijo) + descenso de gradiente

Para la implementación en Matlab se utiliza la técnica de búsqueda lineal Armijo Goldstein[1] con descenso de gradiente acelerado. A diferencia del descenso de coordenadas, en el descenso de gradiente se actualizan todas las componentes simultáneamente. La condición de Armijo asegura que habrá una disminución suficiente en el valor de la función objetivo después de cada iteración para una longitud de paso adecuada. Se trata de, dada una posición inicial **s** y una dirección **p**, encontrar el tamaño adecuado de paso con $\lambda > 0$ que reduce la función objetivo, para encontrar un valor de λ que reduzca $f(x + \lambda p)$ relativo a $f(x)$. La desigualdad de Armijo se ve de la siguiente manera:

$$\boldsymbol{f(x_k + \lambda p_k) \leq f(x_k) + c_1 \lambda \nabla f(x_k)^T p_k}$$

Esta condición puede asegurar que la longitud de paso en cada iteración no es demasiado grande cuando se usa en una búsqueda lineal, Sin embargo, esta condición no es suficiente para asegurar que la longitud de paso del algoritmo es óptima, dado que cualquier valor de λ suficientemente pequeño satisface dicha condición.

2.1.2. Limitaciones de Lasso

Pese a que Lasso ha demostrado ser exitoso en muchos casos, tiene algunas limitaciones. Por ejemplo:

- En el caso en que el número de variables sea mucho mayor que el número de entradas ($p >> n$), Lasso selecciona como máximo n variables antes de saturarse debido a la naturaleza de los problemas de optimización convexos. Esto supone una gran limitación en un proceso de selección de variables.
- Si hay un grupo de variables altamente correladas, Lasso tiende a seleccionar solamente una variable del grupo sin tener en cuenta cual de ellas selecciona y desechando el resto.
- Para los casos en que el número de entradas es muy superior al número de variables ($n >> p$), si hay una alta correlación entre los predictores, se ha observado empíricamente que el rendimiento de predicción de Lasso está dominado por ridge regresion.

Teniendo en cuenta que en este trabajo se está buscando seleccionar los SNPs relacionados con la enfermedad del Alzheimer, si por ejemplo tomamos un problema de selección de genes dentro de un array de datos, generalmente tenemos que el número de predictores (genes) suele ser de muchos miles, mientras que el número de muestras (sujetos) suele ser inferior a mil. Además, para aquellos genes que comparten la misma cadena biológica, la correlación entre ellos puede ser muy alta.

Debido a estas limitaciones, se han propuesto en los últimos años distintos métodos para tratar de corregir dichos problemas. Entre los más destacados se encuentran dos variantes de Lasso conocidas como Elastic net y Group Lasso, que utilizan la norma de regularización L2 usada en la ridge regresion. En este trabajo se van a comparar dichas variantes con Lasso para tratar de eludir las limitaciones que se han descrito previamente.

2.2. Group Lasso

El método de Lasso no siempre proporciona resultados satisfactorios ya que solo selecciona variables individualmente en lugar de factores completos, como podría ser el caso de un gen. Además, en el caso de variables categóricas, la solución depende de cómo se codifican las variables, generalmente se suele crear una variable numérica para cada uno de los posibles valores categóricos, por lo que en el caso de tener, por ejemplo, una variable categórica con 5 niveles, Lasso podría seleccionar 3 de ellos y dejar 2 fuera, cuando todos pertenecen a la misma variable. El método de Group Lasso fue introducido por Yi Lin y Ming Yuan [5] para superar estos problemas al introducir una extensión en la penalización de Lasso.

$$\boldsymbol{\theta = argmin\ ||y - \sum_{j=1}^{J} X_j \theta_j||^2 + \lambda \sum_{j=1}^{J} ||\theta_j||_{K_j}}$$

Donde la matriz de datos $\mathbf{X}$ y el vector de covariables $\boldsymbol{\theta}$ se reemplazan por una colección de matrices $\boldsymbol{X_j}$ y vectores $\boldsymbol{\theta_j}$, para cada uno los J grupos. Adicionalmente,

el término de penalización ahora es una suma sobre la norma L_2 definidas por las matrices positivas $\boldsymbol{K_j}$. Si cada covariable está en su propio grupo y $K_j = I$, la función queda reducida al Lasso estándar. Dado que la penalización se reduce a una norma L_2 en los subespacios definidos por cada grupo, no puede seleccionarse solo algunas de las covariables de un grupo, al igual que la ridge regresion.

De este modo se puede simular una estructura jerárquica en forma de árbol genético estructurado en tres niveles. Cada hoja del árbol representa un SNP, cada nodo intermedio un gen y la raíz el sujeto, como se representa en la figura 2.2.

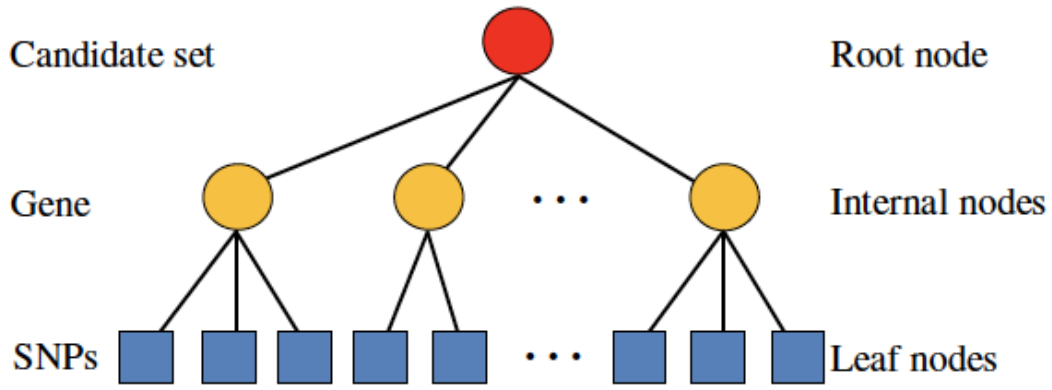


Figura 2.2: Relación jerárquica en forma arborescente entre SNPs y genes

2.3. Elastic Net

En 2005, Zou y Hastie [3] introdujeron la red elástica (Elastic net) para intentar superar algunas de las limitaciones de Lasso expuestas previamente. Se trata de un método de regresión regularizado que combina linealmente las penalizaciones L_1 y L_2 de los métodos Lasso y ridge regression. Su principal objetivo era tratar el caso en que el número de variables sea mucho mayor que el número de entradas ($p >> n$), donde Lasso selecciona como máximo n variables. También en el caso en el cual exista un grupo de variables altamente correladas, donde Lasso tiende a seleccionar una variable de un grupo e ignorar el resto. Siguiendo la notación de los apartados anteriores, la estimación de la red elástica está definida por:

$$\boldsymbol{\theta = argmin\ ||y - X\theta||^2 + \lambda_1||\theta||^2 + \lambda_2||\theta||_1}$$

donde

$$\boldsymbol{||\theta||^2} = \sum_{n=1}^{K} \theta_n^2$$

$$\boldsymbol{||\theta||_1} = \sum_{n=1}^{K} |\theta_n|$$

representan las normas L_1 y L_2, las cuales tienen asociadas los parámetros de regulación λ_1 y λ_2 de forma independiente para corregir la dispersión de cada una de ellas.

En la Figura 2.3 se puede observar la diferencia de las restricciones entre los tres métodos formulados, donde aparecen las distintas formas geometricas que dan lugar a las distintas normas de regularización.

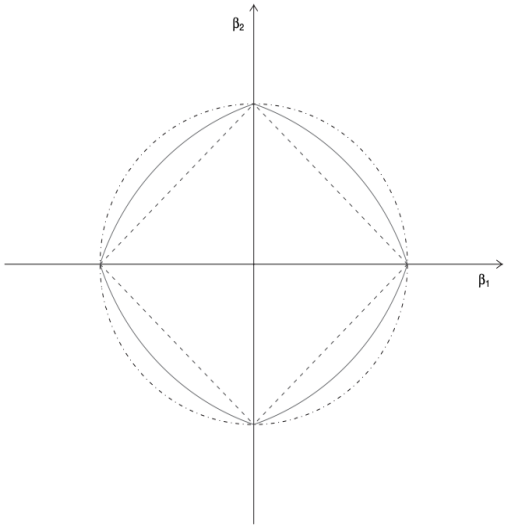


Figura 2.3: Formas de las regiones de restricción para lasso (línea de rayas), ridge regression o group lasso (línea de puntos) y elastic net (línea continua)

Capítulo 3

Materiales y métodos

En esta sección se van a evaluar los conjuntos de datos ADNI-1 y ADNI-2, realizando un estudio del número de SNPs y los individuos que contienen cada uno de ellos, calculando el número de individuos que hay de cada tipo e indicando las medias de edad y educación de cada conjunto, así como el número de hombres y mujeres dentro de los mismos. El conjunto ADNI-1 consta de 620907 SNPS y un total de 757 participantes, en la tabla 3.1 se pueden observar como se dividen entre 368 con deterioro cognitivo leve (LMCI), 175 con la enfermedad del Alzheimer (AD) y 214 individuos sanos (HC).

	LMCI	**AD**	**HC**
Número	368	175	214
Edad (mean±std)	74.77±7.28	75.39±7.38	75.67±4.89
Educación (mean±std)	15.69±3.93	14.61±3.16	16.07±2.79
Género (H/M)	241/127	93/82	115/99

Tabla 3.1: Características de los sujetos ADNI1

El conjunto ADNI-2 consta de 730531 SNPs y un total de 791 participantes, de los cuales 99 con significant memory concern (SMC) , 277 con early mild cognitive impairment (EMCI), 134 con late mild cognitive impairment (LMCI), 126 con la enfermedad del Alzheimer (AD) y 155 individuos sanos (HC), como muestra la tabla 3.2.

	SMC	EMCI	LMCI	AD	HC
Número	99	277	134	126	155
Edad (mean±std)	74.77±7.29	71.14±7.04	72.24±7.78	74.53±7.75	74.00±7.17
Educación (mean±std)	16.81±2.50	15.93±2.61	16.43±2.62	15.78±2.69	16.41±2.48
Género (H/M)	40/59	156/121	73/61	75/51	80/75

Tabla 3.2: Características de los sujetos ADNI-2

HC=Healthy Control, SMC=Significant Memory Concern, EMCI=Early Mild Cognitive Impairment, LMCI=Late Mild Cognitive Impairment, AD=Alzheimer's disease.

3.1. Genotipado y pre-procesamiento

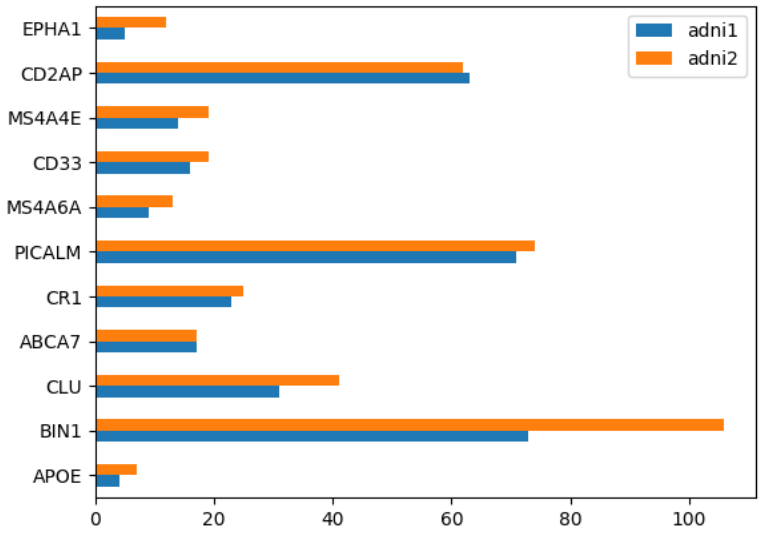
Los datos de genotipado del genoma obtenidos de la base de datos ADNI están en formato PLINK (.bed, .bim, .fam) y contienen los cromosomas de los alelos de cada SNP. Cada alelo es una de las formas alternativas que puede tener un mismo gen, que se diferencian en su secuencia y que se puede manifestar en modificaciones concretas de la función de ese gen. Están representados por las bases nitrogenadas que componene el ADN, la adenina (A), timina (T), guanina (G) y citosina (C).

Para manejar los datos genéticos a gran escala se utiliza una herramienta de análisis genético conocida como PLINK [9] que permite transformar los datos categóricos en numéricos a partir del componente aditivo y dominante, indicando el número de alelos menores. También permite filtrar los datos genotipados utilizando diferentes criterios como por ejemplo: SNPs raros (frecuencia alélica (MAF) $< 0{,}05$), violaciones del equilibrio Hardy-Weinberg (HWE $p < 10^{-6}$), verificaciones de género, etc. El proceso para realizar todos estos análisis y transformaciones está explicado en el Anexo 1.1.

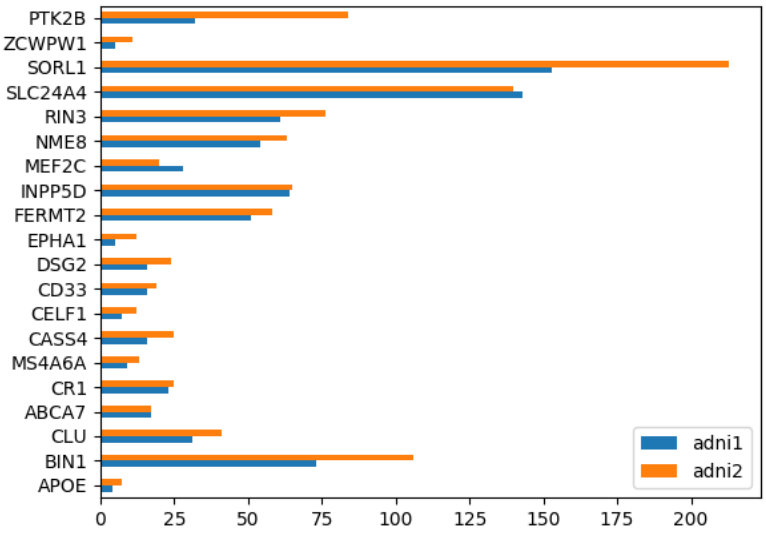
El siguiente paso es reducir el número de variables y trabajar con los SNPs de los genes más relevantes a la hora de que una persona padezca Alzheimer en caso de haber mutaciones en ellos, los cuales podemos encontrar en la base de datos de Alzheimer (www.alzgene.org)[4]. Para ello se hará uso de la herramienta ANNOVAR [11], la cual permite anotar cada SNP con su gen correspondiente y de esta forma crear posteriormente la estructura jerárquica en forma de árbol genético. En el Anexo 1.2 vienen explicados los pasos que hay que realizar.

En la Figura 3.1 se muestra la distribución final de los SNPs que van a ser utilizados en el estudio una vez han sido procesados por las herramientas previamente descritas. Para el top 10 de genes se tendría un total de 318 SNPs para ADNI-1 y 387 SNPs para ADNI-2, mientras que para el top 20 se tendrían un total de 789 y 1017 SNPs para cada uno de los conjuntos de datos respectivamente.

Una vez clasificado cada SNP con su gen correspondiente y teniendo los datos en formato numérico, procedemos a crear la matriz de datos que se utilizará en el



(a) Top 10 AD genes



(b) Top 20 AD genes

Figura 3.1: Número de SNPs dentro del top 10 y 20 AD genes

estudio. Para ello falta un último dato, el volumen de cada región del hipocampo de cada sujeto, obtenido a través de las imágenes MRI. Para ello se ha hecho uso de los estudios realizados por distintas entidades y que están recogidos en la base de datos de ADNI en los que aparecen volúmenes de diferentes zonas y regiones del cerebro, extraídos de las imágenes de los individuos de cada conjunto. Para ADNI-1 se ha hecho uso del estudio realizado por la Universidad de California, San Diego (UCSD), mientras que para ADNI-2 se ha hecho uso del de la Universidad de San Francisco (USCF), en los cuales aparecen los valores del volumen de los hipocampos derecho e izquierdo para cada individuo de los correspondientes estudios.

Capítulo 4

Resultados

4.1. Datos del estudio

En esta sección se presenta el estudio realizado de los métodos Lasso, Elastic Net y Group Lasso. Se van a comparar los resultados obtenidos para los conjuntos de datos ADNI-1 y ADNI-2, así como para el top 10 y 20 de genes más relacionados con el Alzheimer, de tal modo que se puedan comparar con los artículos en los que se basa este trabajo [8], [12]. Además se llevará a cabo una comparativa entre las implementaciones de Python (basada en descenso de coordenadas) y Matlab (basada en búsqueda lineal + descenso de gradiente acelerado). El estudio ha sido realizado mediante una partición de los datos de entrada, divididos en un 80 % para datos de entrenamiento y un 20 % para datos de test. Los datos de entrenamiento han sido subdivididos en datos de entrenamiento (90 %) y validación (10 %). El rendimiento de los distintos métodos se ha calculado mediante la raíz del error cuadrático medio (RMSE) un criterio usado ampliamente en la evaluación de sistemas analíticos. La media de los valores RMSE, véase figura 4.1, ha sido calculada por el método de validación cruzada con 10 pliegues. Finalmente, para que los resultados fuesen más robustos, se han realizado un total de 100 experimentos, haciendo un estudio de la frecuencia de selección de los SNPs. En cada experimento se ha realizado una validación cruzada para buscar el valor del parámetro de regularización λ con el que seleccionar, mediante los datos de test, un total de 100 y 200 SNPs para el top 10 y 20 de genes respectivamente.

4.2. Comparación de implementaciones en estudio de frecuencias

En este apartado se va a hacer una comparación entre las dos implementaciones con los resultados obtenidos en los métodos de Lasso y Elastic net. Para presentar los múltiples experimentos realizados, y dada la diferencia existente en los datos de entrada, se van a exponer divididos en los conjuntos de datos ADNI-1 y ADNI-2, puesto que los SNPs de cada conjunto son diferentes.

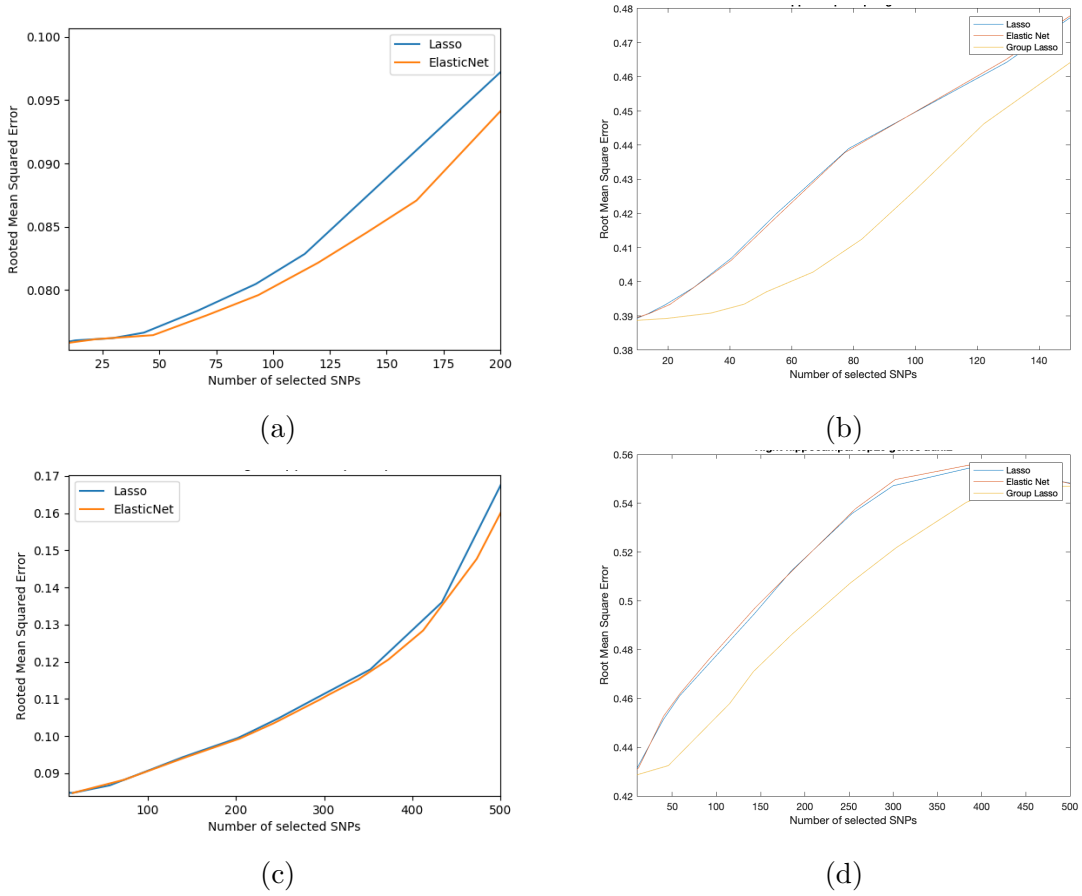


Figura 4.1: Errores de validación para datos de entrenamiento mediante k-fold cross validation. a) Python Left Hippocampal top 10 genes ADNI-1, b) Matlab Left Hippocampal top 10 genes ADNI-1, c) Python Rigth Hippocampal top 20 genes ADNI-2, d) Matlab Right Hippocampal top 20 genes ADNI-2

4.2.1. ADNI-1

En la tabla 4.1 se muestra el número de coincidencias en las tablas de frecuencias de ambas implementaciones. Se puede observar como en la mayoría de los casos los mismos SNPs son seleccionados con un porcentaje mayor o cercano al 50 %. En especial, para el hipocampo izquierdo los SNPs **rs405509-APOE, rs439401-APOE, rs10779339-CR1** y **rs9314349-CLU** aparecen seleccionados en todos los experimentos de ADNI-1. En el hipocampo derecho aparecen seleccionados los SNPs **rs439401-APOE, rs405509-APOE, rs4726618 -EPHA1, rs3752237-ABCA7, rs2242601-EPHA1** y **rs9314349-CLU**. Lo que da como resultado final que los SNPs **rs405509-APOE, rs439401-APOE,** y **rs9314349-CLU** aparecen tanto en el hipocampo derecho como izquierdo seleccionados dentro del Top 50 de frecuencias de selección de ambas implementaciones y en ambos conjuntos de genes. Estos resultados coinciden con el artículo [8], en el que se ha basado este Trabajo de Fin de Grado, donde ambos SNPs del APOE aparecían seleccionados dentro del conjunto de datos ADNI-1, enfatizando todavía mas en el hecho de que también aparecen seleccionados dentro de conjunto de 20 genes, cuando en el artículo hacia referencia únicamente al de 10.

		Top 10 freq		Top 25 freq		Top 50 freq	
Genes	Hip	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
10	Der	3	3	15	15	28	29
	Izq	5	4	11	11	30	34
20	Der	2	3	9	9	23	22
	Izq	3	4	8	7	21	21

Tabla 4.1: Estudio Frecuencias ADNI-1. Número de coincidencias entre implementaciones para el top 10 y 20 de genes en los hipocampos derecho e izquierdo

4.2.2. ADNI-2

Para el conjunto de datos ADNI-2 se muestran en la tabla 4.2 los resultados del estudio de frecuencias. En este caso los resultados muestran un poco más de disparidad entre sí que en ADNI-1, especialmente en el conjunto de 20 genes, debido a que el número de datos es mayor y hay mas posibilidad de que seleccione otros SNPs en cada experimento del estudio. En el hipocampo izquierdo se observa que los SNPs **rs7533408-CR1, rs405509-APOE, rs1866235-BIN1** y **rs3852865-CD33** son seleccionados dentro del top 50 de frecuencias en ambas implementaciones mientras que en el hipocampo derecho los SNPs que aparecen seleccionados en ambas implementaciones son: **rs405509-APOE, rs7533408-CR1, rs4726618-EPHA1, rs1469979-BIN1, rs3752237-ABCA7, rs3852865-CD33** y **rs3795065-ABCA7**. Resultando finalmente que los SNPs **rs7533408-CR1, rs405509-APOE** y **rs3852865-CD33** aparecen seleccionados en ambos hipocampos para ambas implementaciones y los distintos conjuntos de genes dentro del Top 50 de frecuencias.

		Top 10 freq		Top 25 freq		Top 50 freq	
Genes	Hip	Lasso	ElasticNet	Lasso	ElasticNet	Lasso	ElasticNet
10	Der	2	1	10	11	26	28
	Izq	4	4	13	15	31	30
20	Der	4	2	7	9	18	17
	Izq	1	1	5	4	18	17

Tabla 4.2: Estudio Frecuencias ADNI-2. Número de coincidencias entre implementaciones para el top 10 y 20 de genes en los hipocampos derecho e izquierdo

4.2.3. Resultados comunes

Entre todos los experimentos realizados, cabe destacar el SNP **rs405509** del gen **APOE** dado que ha sido el único seleccionado dentro del Top 50 de frecuencias en todos ellos, independientemente del número de genes, SNPs o región del hipocampo de los estudios. De hecho en la mayoría de ellos aparecen incluso dentro del Top 10 de frecuencias. Esto enfatiza una vez más el hecho de que el APOE haya sido un gen

		Top 10 freq		Top 25 freq		Top 50 freq	
Genes	Hipoc	ADNI-1	ADNI-2	ADNI-1	ADNI-2	ADNI-1	ADNI-2
10	Der	9	8	23	22	47	46
	Izq	8	8	22	23	47	45
20	Der	8	8	23	18	45	37
	Izq	9	7	19	21	38	42

Tabla 4.3: Estudio Frecuencias Python. Número de coincidencias entre Lasso y Elastic net para el top 10 y 20 de genes en los hipocampos derecho e izquierdo

reconocido durante mucho tiempo y en numerosos estudios como uno de los principales genes relacionados con la enfermedad del Alzheimer. Otro dato destacable son los SNP **rs3752237** del gen **ABCA7** y **rs4726618** del gen **EPHA1**, que resultan seleccionados en el hipocampo derecho en todos los métodos e implementaciones, y que han aparecido en estudios recientes como genes de referencia, en especial el gen **ABCA7**.

4.3. Comparación de métodos en estudio de frecuencias

En este apartado se va a hacer una comparación entre los diferentes métodos con los resultados obtenidos en cada implementación. Cabe destacar que los SNPs seleccionados por los métodos de Lasso y Elastic net son muy similares.

4.3.1. Descenso de coordenadas

En la tabla 4.3 se pueden observar los resultados obtenidos con la implementación de descenso de coordenadas por los métodos Lasso y Elastic net. La tasa de coincidentes es muy alta, superior al 90 % en la mayoría de los casos, por lo que ambos métodos tienden a seleccionar los mismos SNPs. La diferencia entre los índices de frecuencia es muy pequeña, y pese a que la posición en que aparecen en cada implementación varía, lo hace en muy pocas posiciones, por lo que encontrarlos dentro de los tops de frecuencias es muy común.

Dentro del top 50 de frecuencias se encuentran para ADNI-1 un total de 13 coincidencias en el hipocampo izquierdo, 11 en el derecho y 5 entre ambos. En ADNI-2 el total de coincidencias es de 13 para el hipocampo izquierdo, 14 para el derecho y 9 en común. Al igual que en el estudio de frecuencias entre métodos discutido en la sección anterior, destacan los SNPs **rs405509** y **rs439401** del gen **APOE**, que aparecen seleccionados en ambos hipocampos y en ambos conjuntos de datos. En el hipocampo derecho nuevamente los SNPs **rs3752237** del gen **ABCA7** y **rs4726618** del gen **EPHA1** son seleccionados como en el estudio anterior.

		Top 10 freq		Top 25 freq		Top 50 freq	
Genes	Hipoc	ADNI-1	ADNI-2	ADNI-1	ADNI-2	ADNI-1	ADNI-2
10	Der	10	8	23	23	46	46
	Izq	10	9	24	23	47	47
20	Der	8	9	22	23	44	43
	Izq	8	8	21	24	43	46

Tabla 4.4: Estudio Frecuencias Matlab. Número de coincidencias entre Lasso y Elastic net para el top 10 y 20 de genes en los hipocampos derecho e izquierdo

4.3.2. Búsqueda lineal Armijo

En la implementación de Armijo con descenso de gradiente los resultados que se muestran a continuación se dividen en dos conjuntos. Por un lado las coincidencias entre Lasso y Elastic net, y por otro los resultados obtenidos entre estos dos métodos y Group Lasso. En las tablas 4.4 y 4.5 se muestran los distintos resultados.

Como sucedía con el descenso de coordenadas, las coincidencias entre Lasso y Elastic net son muy altas, superiores al 90 %, lo que indica que el número de SNPs seleccionados para realizar el estudio no es muy elevado, y la diferencia entre Lasso y Elastic net se verá principalmente para los casos en los que el número de variables es muy superior al número de muestras.

Las coincidencias entre Lasso y Elastic net para ADNI-1 son de 12 para el hipocampo izquierdo, 13 para el derecho y 8 en común, mientras que en ADNI-2 son de 11 para el izquierdo, 9 para el derecho y 5 en común. Una vez más, el SNP **rs405509** del **APOE** es seleccionado entre los métodos de Lasso y Elastic net en ambos hipocampos y en el hipocampo derecho los SNPs **rs3752237** del gen **ABCA7** y **rs4726618** del gen **EPHA1** vuelven a ser seleccionados en ambos conjuntos de datos.

En la comparación con Group Lasso sin embargo no ocurre lo mismo. En ADNI-1 tan solo se encuentran 2 coincidencias para el hipocampo izquierdo, 3 para el derecho y 1 en común. En el conjunto de ADNI-2 únicamente se encuentra 1 coincidencia en el hipocampo izquierdo. Lo más destacable es que los SNPs en común pertenecen todos al gen **BIN1**, excepto uno al gen **CLU**. Esto es debido a que Group Lasso tiende a seleccionar los elementos de un conjunto o descartar dicho conjunto, y en los conjuntos con pocos SNPs, como puede ser el caso del APOE, hay particiones de los datos de entrenamiento en los que las frecuencias alélicas utilizadas para la creación de la matriz de datos dan columnas de ceros en ciertos SNPs del gen, por lo que en ese caso tiende a descartarlo.

4.4. Número de SNPs seleccionados por gen

Otra cuestión interesante es el estudio del número de SNPs seleccionados para cada gen. En la figura 4.2 se muestran los datos obtenidos en el estudio para el top 10 de genes. Se puede observar como en ambos conjuntos de datos los datos que se obtienen entre implementaciones son muy similares, pese a que se seleccionen distintos SNPs dentro de cada gen, el número total seleccionado en cada uno de

		Top 10 freq		Top 25 freq		Top 50 freq	
Genes	Hipoc	ADNI-1	ADNI-2	ADNI-1	ADNI-2	ADNI-1	ADNI-2
10	Der	3	1	13	9	33	29
	Izq	5	0	18	8	40	25
20	Der	2	2	10	8	16	20
	Izq	5	1	9	9	23	25

Tabla 4.5: Estudio Frecuencias Matlab. Número de coincidencias entre Lasso, Elastic Net y Group Lasso para el top 10 y 20 de genes en los hipocampos derecho e izquierdo

		Lasso		Elastic Net		Group Lasso	
		Izq	Der	Izq	Der	Izq	Der
APOE	Python	2	3	2	2		
	Matlab	2	2	2	2	2	2
BIN1	Python	29	37	37	28		
	Matlab	29	30	31	31	34	27
CLU	Python	10	13	16	8		
	Matlab	13	11	15	11	14	9
ABCA7	Python	6	6	4	5		
	Matlab	5	6	5	7	7	6
CR1	Python	4	6	4	2		
	Matlab	6	4	6	4	9	6
PICALM	Python	18	23	21	16		
	Matlab	20	17	20	18	24	18
MS4A6A	Python	3	1	2	1		
	Matlab	5	1	5	2	5	0
CD33	Python	3	4	5	4		
	Matlab	5	4	5	5	5	4
MS4A4E	Python	2	1	2	4		
	Matlab	2	1	3	1	3	0
CD2AP	Python	22	17	25	16		
	Matlab	21	21	21	24	20	26
EPHA1	Python	1	3	3	2		
	Matlab	3	4	3	4	3	5
TOTAL	Python	100	114	121	88		
	Matlab	111	101	116	109	126	103

(a) ADNI-1

		Lasso		Elastic Net		Group Lasso	
		Izq	Der	Izq	Der	Izq	Der
APOE	Python	5	6	5	3		
	Matlab	2	3	4	3	4	4
BIN1	Python	28	24	30	22		
	Matlab	24	28	35	31	31	28
CLU	Python	13	13	12	15		
	Matlab	10	13	13	12	9	9
ABCA7	Python	3	3	6	5		
	Matlab	6	8	7	8	7	9
CR1	Python	7	7	6	6		
	Matlab	6	7	10	8	5	10
PICALM	Python	12	17	20	20		
	Matlab	19	16	24	16	20	14
MS4A6A	Python	1	1	1	2		
	Matlab	1	3	1	3	0	0
CD33	Python	6	5	4	4		
	Matlab	5	4	5	4	8	6
MS4A4E	Python	6	2	4	1		
	Matlab	2	9	1	9	5	9
CD2AP	Python	15	18	16	13		
	Matlab	13	11	16	13	20	15
EPHA1	Python	7	5	4	4		
	Matlab	6	5	6	5	6	5
TOTAL	Python	103	101	108	95		
	Matlab	94	107	122	112	115	109

(b) ADNI-2

Figura 4.2: Número de SNPs seleccionados en el top 10 genes

ellos varía poco. Como Group Lasso tiende a seleccionar un conjunto de datos o no seleccionar ninguno dentro del conjunto, en los genes en los que Lasso y ElasticNet seleccionan un número muy pequeño de SNPs, Group Lasso tiende a no seleccionar ninguno, como en el ejemplo del gen **MS4A6A**.

En la figura 4.3 se muestran los resultados para el top 20 de genes. Nuevamente el número de SNPs seleccionados en cada gen es muy similar en todos los métodos y en ambas implementaciones. En Group Lasso se vuelve a dar el hecho de que haya genes en los que no seleccione ningún SNP dadas las características del propio método como se ha expuesto previamente. Con el método de Group Lasso se obtienen predicciones con menor error como se muestra en la figura 4.1 sirve mejor para conocer los genes

relevantes que los SNPs como tal.

Si comparamos los resultados se puede observar como los genes con un mayor número de SNPs en el conjunto de datos son claramente los que más selecciona, pero cabe destacar una vez más el **APOE**, dado que es el gen con menor número de SNPs dentro de los distintos conjuntos de datos y en todos los métodos son seleccionados la mayoría de ellos, incluido en el método de Group Lasso.

		Lasso		Elastic Net		Group Lasso	
		Izq	Der	Izq	Der	Izq	Der
APOE	Python	2	3	2	2		
	Matlab	1	1	1	2	0	2
BIN1	Python	19	17	21	25		
	Matlab	22	24	24	25	22	20
CLU	Python	11	10	13	10		
	Matlab	10	6	10	6	11	5
ABCA7	Python	6	2	1	2		
	Matlab	2	3	3	3	6	0
CR1	Python	5	4	3	5		
	Matlab	5	5	6	5	5	3
MS4A6A	Python	3	2	4	3		
	Matlab	1	2	1	2	0	8
CASS4	Python	2	5	7	5		
	Matlab	4	7	4	7	12	8
CELF1	Python	4	2	1	1		
	Matlab	1	4	1	4	0	6
CD33	Python	2	7	4	3		
	Matlab	4	6	4	6	0	6
DSG2	Python	5	3	3	5		
	Matlab	4	2	4	2	0	0
EPHA1	Python	5	4	4	4		
	Matlab	5	1	4	3	0	0
FERMT2	Python	11	8	11	8		
	Matlab	9	15	9	14	12	16
INPP5D	Python	15	15	15	18		
	Matlab	15	15	19	16	13	22
MEF2C	Python	7	2	5	1		
	Matlab	2	1	2	2	3	4
NME8	Python	9	12	11	10		
	Matlab	12	14	11	15	9	14
RIN3	Python	15	14	15	9		
	Matlab	21	16	23	17	19	14
SLC2A4	Python	26	23	26	22		
	Matlab	26	31	26	32	17	27
SORL1	Python	36	37	31	41		
	Matlab	43	42	45	44	41	39
ZCWPW1	Python	1	3	3	2		
	Matlab	3	1	3	2	0	0
PTK2B	Python	7	10	11	8		
	Matlab	12	12	14	12	7	11
TOTAL	Python	191	183	191	184		
	Matlab	202	208	214	219	177	205

(a) ADNI-1

		Lasso		Elastic Net		Group Lasso	
		Izq	Der	Izq	Der	Izq	Der
APOE	Python	1	1	1	2		
	Matlab	2	1	2	1	2	0
BIN1	Python	19	16	20	16		
	Matlab	19	24	19	23	13	19
CLU	Python	7	8	8	8		
	Matlab	6	10	7	11	4	9
ABCA7	Python	3	2	6	1		
	Matlab	4	2	4	2	4	0
CR1	Python	4	2	6	3		
	Matlab	4	4	4	4	6	3
MS4A6A	Python	0	2	1	2		
	Matlab	2	2	2	2	0	0
CASS4	Python	3	5	7	8		
	Matlab	5	4	6	5	6	7
CELF1	Python	1	2	1	1		
	Matlab	0	0	0	0	0	0
CD33	Python	2	3	3	3		
	Matlab	5	3	5	4	8	0
DSG2	Python	3	4	3	3		
	Matlab	7	7	7	8	3	8
EPHA1	Python	3	1	4	3		
	Matlab	2	2	2	2	0	2
FERMT2	Python	12	16	11	11		
	Matlab	10	12	12	12	13	10
INPP5D	Python	12	22	21	15		
	Matlab	17	22	17	22	15	18
MEF2C	Python	6	8	5	8		
	Matlab	8	1	8	1	7	0
NME8	Python	6	10	13	13		
	Matlab	14	10	18	13	19	11
RIN3	Python	16	17	19	17		
	Matlab	21	19	23	19	20	17
SLC2A4	Python	31	20	24	31		
	Matlab	30	34	31	38	29	30
SORL1	Python	36	34	34	33		
	Matlab	30	29	32	32	37	28
ZCWPW1	Python	1	3	3	3		
	Matlab	1	2	1	2	2	1
PTK2B	Python	8	8	5	5		
	Matlab	5	5	5	8	4	6
TOTAL	Python	174	184	195	186		
	Matlab	192	193	205	209	192	169

(b) ADNI-2

Figura 4.3: Número de SNPs seleccionador en el top 20 genes

Capítulo 5

Conclusión

En este trabajo se han comparado diferentes variantes de Lasso utilizando distintas implementaciones. Los resultados obtenidos con distintos conjuntos de datos de la base de datos ADNI demuestran que la implementación de descenso de coordenadas obtiene unos resultados más robustos dado que tanto los porcentajes de frecuencias de selección de los SNPs son mucho más altos, así como los errores de validación de los estudios son menores. Los métodos de Lasso y Elastic net obtienen unos resultados muy similares debido a que el principal objetivo de mejora es el caso en que el número de variables es muy superior al número de muestras, y en este trabajo interesa conocer el número de variables seleccionadas en conjuntos de datos en los que el número de muestras es superior debido al preprocesamiento realizado. El método de Group Lasso presenta resultados mejores en las predicciones, pero su sus resultados sirven más para conocer los genes con mas relevancia en lugar de los SNPs relevantes como tal, dado que tiende a descartar grupos si hay elementos no seleccionados en ellos. Lo que queda destacado entre todos los distintos métodos e implementaciones es que el gen APOE, y en concreto el SNP rs405509, tienen una alta relación con la enfermedad del Alzheimer, al aparecer en el top de frecuencias y con altos coeficientes en todos ellos, a excepción de alguno de Group Lasso. También cabe destacar que los genes ABA7 y EPHA1 tiene una mayor relación con el hipocampo derecho, en especial los SNPs rs3752237 y rs4726618 correspondientes a cada gen respectivamente.

En la figura 5.1 se muestra un diagrama de Gantt con las diferentes tareas realizadas a lo largo del trabajo. Los principales problemas encontrados han estado en las fases de preprocesado de los ficheros para obtener las matrices de datos con los que realizar el estudio, puesto que se trata de ficheros con una gran cantidad de variables y encontrar las relaciones de los individuos con sus volúmenes de hipocampo para cada conjunto de datos no está bien definido dentro de ADNI. La fase de comparación de los datos también presenta ciertos problemas, dado que hay una gran cantidad de conjuntos de resultados disitintos que comparar (hasta 16 conjuntos diferentes) con una gran cantidad de marcadores que varían muy poco entre sí para determinar que SNPs son más o menos relevantes. Para resolver estos problemas se implementaron algoritmos que tratan tanto los datos de entrada como de salida de forma automática, pudiendo crear una gran cantidad de conjuntos de datos distintos para realizar estudios de una forma rápida con tan solo especificar ciertos parámetros de entrada, como ficheros o directorios.

Todos los métodos e implementaciones realizadas, desde los sistemas de apren-

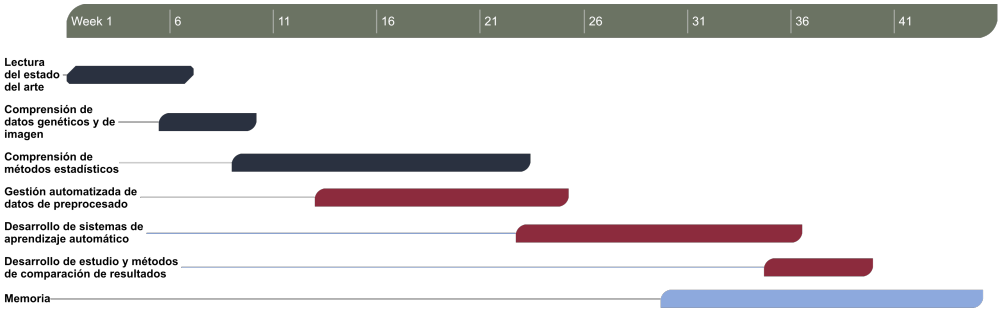


Figura 5.1: Diagrama temporal de las tareas del trabajo

dizaje automático como los sistemas de preprocesado y comparación de resultados están hechos de forma que puedan extenderse a otros trabajos futuros, como por poder aplicarlos para identificar asocaciones entre variantes genéticas del ADN mitocondrial y la enfermedad del Alzheimer, o incluso con otras enfermedades en caso de modificar los genes utilizados en el estudio.

Bibliografía

[1] Armijo, L. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific J. Math 16* (1996), 1–3.

[2] Burns A, Iliffe S. Alzheimer's disease. *BMJ* (2009), 338:b158.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Statist 67* (2005), 301–320.

[4] Lars Bertram et al. Systematic meta-analyses of alzheimer disease genetic association studies: the alzgene database. *Nature Genetics 39* (2007), 17–23.

[5] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist 68* (2006), 49–67.

[6] Mueller, Susanne G.; Weiner, Michael W.; Thal, Leon J.; Petersen, Ronald C.; Jack, Clifford; Jagust, William; Trojanowski, John Q.; Toga, Arthur W.; Beckett, Laurel. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America* (2017).

[7] O. Kohannim, D. P. Hibar, J. L. Stein, N. Jahanshad, C. R. Jack, M. W. Weiner, A. W. Toga, and P. M. Thompson. Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. *International Symposium on Biomedical Imaging: From Nano to Macro* (2011), 1855–1859.

[8] P. Golland et al. Identifying candidate genetic associations with mri-derived ad-related roi via tree-guided sparse learning. *The Medical Image Computing and Computer Assisted Intervention Society* (2014), 757–764.

[9] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC. Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetic* (2007), 81.

[10] R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Royal Statistical Society Series B-Statistical Methodology 73* (2011), 273–282.

[11] Wang K, Li M, Hakonarson H. Annovar: Functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Research* (2010), 38–164.

[12] Xiaoke Hao, Xiaohui Yao, Shannon L. Risacher, Andrew J. Saykin Saykin, Jintai Yu Yu, Huifu Wang, Lan Tan, Li Shen, Daoqiang Zhang Zhang, and for the Alzheimer's Disease Neuroimaging Initiative. Identifying candidate genetic associations with mri-derived ad-related roi via tree-guided sparse learning. *IEEE Transactions on Computational Biology and Bioinformatics* (2018), 1545–5963.