**Aprendizaje Automático Proyecto de la asignatura.**

**Curso 2019–2020**

**Aprendizaje automático en la evaluación del cáncer de pulmón**

Pablo Martín García**,** Aaron Reboredo Vázquez

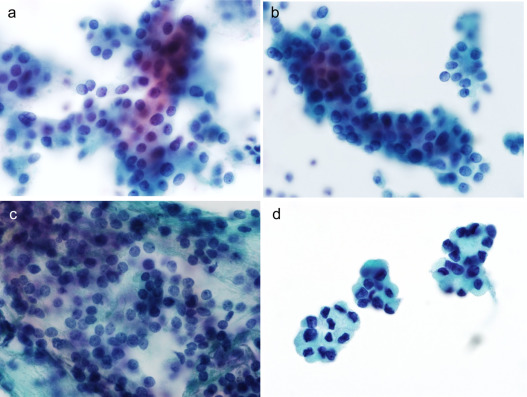
Esta memoria cubrirá las técnicas aplicadas a la hora de tratar el data set seleccionado para el proyecto final de la asignatura de Aprendizaje Automático.

**Motivación:**

“Cáncer” es una de las palabras más temidas hoy en día por el ser humano en el ámbito de la salud. Afortunadamente la ciencia y la medicina avanza y cada vez se están descubriendo formas menos abrasivas y más efectivas para combatirlo. Sin embargo, uno de los elementos cruciales para combatir la enfermedad es el rápido y preciso diagnóstico de la enfermedad para que pueda ser tratado con la mayor brevedad posible y con el tratamiento más adecuado. De esta manera nos resulta interesante el uso de algoritmos de aprendizaje automático que pueden ayudar a clasificar y a agilizar el proceso de diagnóstico de diversos tumores, en nuestro caso, el cáncer pulmonar.

El data set elegido contiene información relevante en la investigación del cáncer de pulmón para un estudio realizado en Wisconsin. En el data set se recogen los atributos de las células de una muestra de pulmón previamente analizadas digitalmente por imagen y les asocia un grado de Malignidad o Benignidad resultante de los análisis correspondientes.

El objetivo es comprobar que se puede generar un sistema capaz de predecir la malignidad o benignidad de un tumor pulmonar en función de los parámetros de entrada que se analizan bajo microscopio y solo utilizando el sistema de análisis por imagen con cierta efectividad. Como comentamos en este caso esas imágenes ya están procesadas y nosotros trabajaremos directamente con los datos extraídos de ese análisis por imagen.



**Imagen digitalizada de extracción masa pulmonar vía aspiración por aguja.**

**El dataset:**

Nuestro data set está compuesto por 569 casos de análisis de los cuales 357 se corresponden con valores para los cuales el resultado es benigno y 212 se corresponden con casos malignos. De ellos conocemos 30 valores que se corresponderán con nuestras variables y que hacen referencia a la media, errores estándar y peores o mayores valores. De esta manera y cuando leamos el data set lo procesaremos para trabajarlo como una matriz de 569 x 30 con los valores de las variables y un vector resultado de dimensión 30 con 0 para el resultado benigno y 1 para el caso de haber obtenido un resultado maligno.

Nuestro set de datos se compone de:

Variables:

Los campos sobre los que vamos a realizar nuestro análisis o sobre los que vamos a ejecutar nuestros sistemas de predicción son los siguientes:

* Radio: Hace referencia a la distancia del punto central a un punto del perímetro del tumor.
* Textura: Hace referencia a la variedad de tonalidades de la imagen del tumor.
* Perímetro: Hace referencia a la longitud del borde del tumor o las aglutinaciones de este.
* Área: Hace referencia a la superficie que abarca el tumor.
* Suavidad de los bordes: Hacer referencia a la forma de los bordes del tumor Esto es la variación de longitudes de los radios.
* Compacidad: Es una medida a partir del perímetro del tumor y de la superficie que ocupa.
* Grado de simetría: Hace referencia a la equidad en las distancias de los radios a partir del eje perpendicular a los radios analizados.
* Concavidad: Hace referencia a la severidad de las porciones cóncavas del contorno del tumor.
* Cantidad de concavidades: Hace referencia al número total de concavidades visibles en la muestra.
* Dimensión fractal: Es una medida de la proximidad de las irregularidades de los bordes. Da una idea de la forma que tiene el perímetro del tumor.

Se analizan las medias de:

El radio, textura media de la mancha, perímetro medio, área, suavidad de los bordes, la compacidad, la concavidad, número de concavidades, grado de simetría, dimensión fractal.

Se analiza el error estándar en:

El radio, la textura, el perímetro, el área, la suavidad de los bordes, la compacidad, la concavidad, número de puntos cóncavos, simetría y dimensión fractal.

Se analizan también y se tienen en cuenta para el análisis los peores casos de (medidas más grandes y preocupantes dentro de las variables descritas):

Radio, textura, compacidad, perímetro, área, suavidad, textura, radio, concavidad, mayor cantidad de número de puntos cóncavos, peor simetría y peor dimensión fractal.

Para trabajar con nuestra base de datos aplicaremos principalmente tres técnicas de análisis que nos permitirán predecir y clasificar los casos de análisis y comprobar mediante los valores de prueba (el data set), la fiabilidad o porcentaje de acierto de los procedimientos.

* Regresión logística
* Redes neuronales
* SVM

**Regresión lineal y gradiente descendiente:**

De esta manera empezaremos aplicando regresión lineal. A continuación, estudiaremos el caso mediante una regresión logística y terminaremos por implementar una red neuronal que nos permita discernir y clasificar los casos de estudio en función de los valores de entrada descritos anteriormente.

En el primer caso trabajaremos con una regresión lineal con varias variables, que se corresponden con cada una de las columnas de nuestro data set y con las descritas anteriormente. Haremos la aproximación por descenso de gradiente y comprobaremos tanto su similitud en las predicciones a los resultados obtenidos en la ecuación normal como su similitud con los valores reales para así poder extraer un porcentaje de acierto del análisis.

Es posible trabajar con este tipo de regresión puesto que tras un primer análisis o representación de las variables y su malignidad y benignidad nos hemos dado cuenta que dos a dos existía una correlación que se podría aproximar a lo lineal entre los variables analizadas y el diagnóstico obtenido.

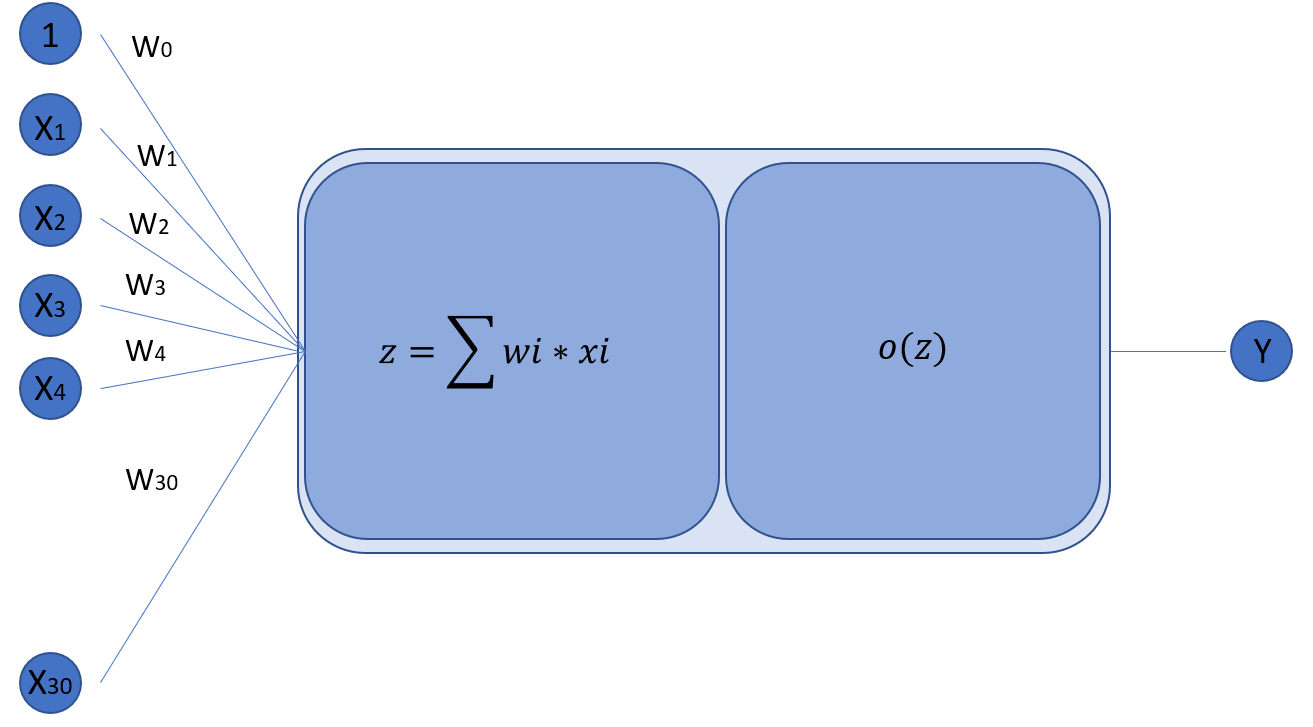
Los valores predichos por el procedimiento del gradiente descendiente se alejan de los valores reales que deberíamos obtener (recordemos que trabajábamos con 0 y con 1), pero que podemos modelizar o ajustar de tal manera que todos aquellos valores que no superen el 0.5 sean tomados como valor 0 (benignidad) y todos aquellos que superen el 0.5 sean considerados como 1(malignidad). Si practicamos este ajuste tanto a la predicción por gradiente descendiente como a la predicción normal obtenemos un 100 % de equivalencia entere ambas predicciones. Si comparamos los resultados predichos con los resultados reales obtenemos un 87% de coincidencia en las predicciones para el uso de regresión lineal mediante gradiente descendiente, lo que nos pone en una predicción no ideal, pero en cierta medida favorable.

Este modelo de predicción, sin embargo, no es el ideal, puesto que tenemos que modelizar los resultados obtenidos. La manera más eficaz para llevar un análisis de clasificación como el que queremos hacer tenemos otros modelos bastante más precisos y correctos como pueden ser la regresión logística, redes neuronales o las SVM, que nos brindan modelos más adecuados al tipo de resultado que queremos predecir (clasificación) y como veremos, más precisos.

**Regresión logística:**

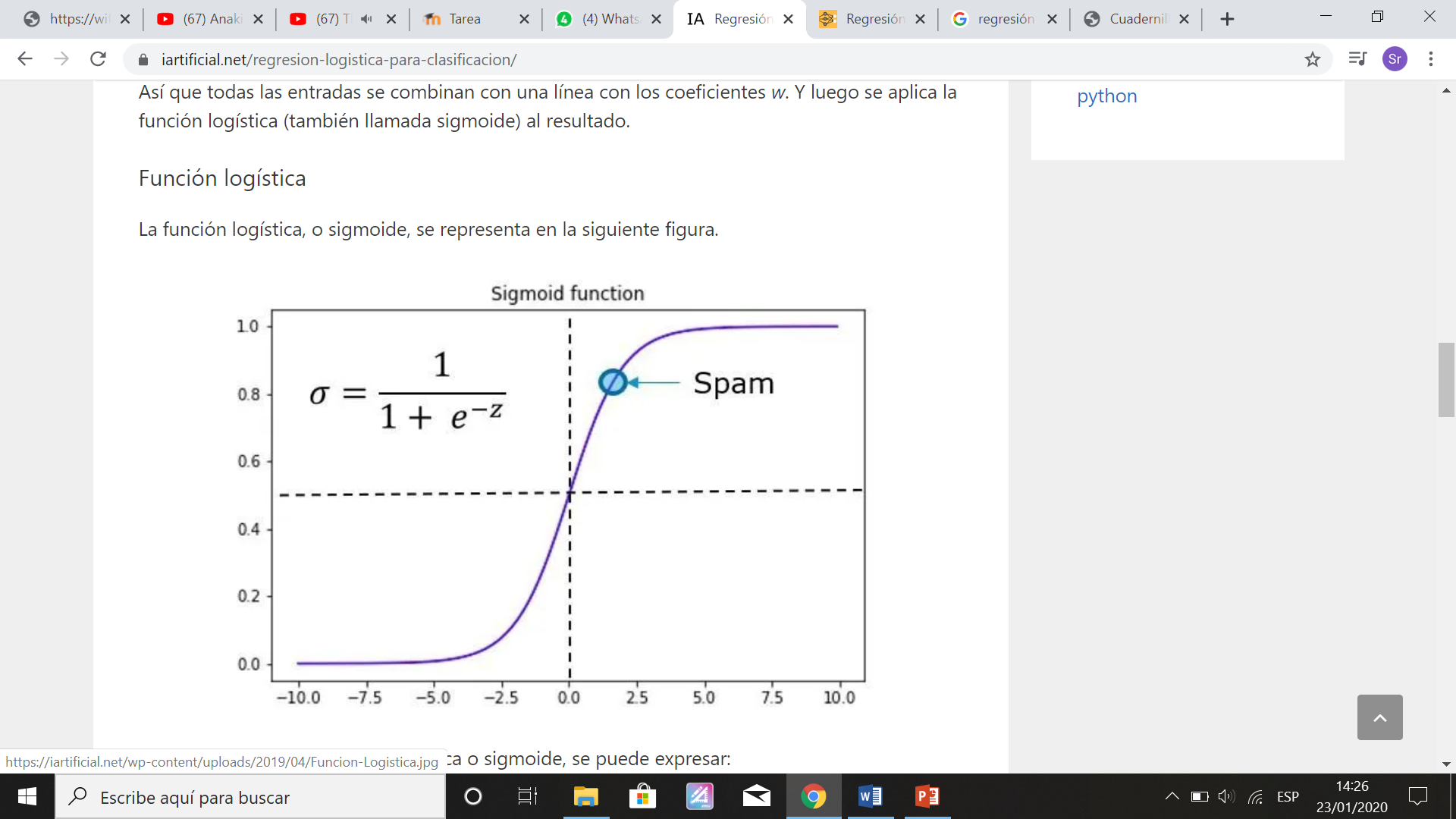
El primer paso lógico a la hora de lanzar nuestro análisis es empezar por el más sencillo de los tres que nos hemos planteado. Esta sencillez facilita y acelera la implementación de un sistema capaz de clasificar los valores de nuestro data set. Sin embargo, todo apunta a que esta sencillez nos hará pagar en precisión en la predicción.

La regresión logística es una técnica de aprendizaje automático relativamente sencilla. Es una de las técnicas más aplicadas en la industria debido a sus buenos resultados cuando se trabaja con muchos datos con interrelaciones sencillas. Además, el modelo de regresión logística nos permite aplicar la regularización, lo que nos permite generalizar y obtener mejores resultados.



Una regresión logística es en resumidas cuentas una pequeña red neuronal constituida por una sola neurona.

Cuando trabajamos con una regresión logística los datos se “preprocesan” y se combinan de manera lineal para posteriormente aplicarle la función logística y que en nuestro caso será la sigmoide.



Para la predicción de valores debemos pasar por el cálculo de los costes y los gradientes regularizados de trabajar con nuestra base de datos y procesarlos de tal manera que podamos obtener los parámetros óptimos, los pesos de nuestra pequeña red neuronal, que llamaremos thetas, para la regresión logística implementada. Esto pasa por hacer los cálculos de ambos y procesarlos.

El proceso del que hablamos es una minimización. Para ello existen diversos modelos de minimizado. Durante el curso hemos utilizado del modelo TNC, sin embargo, en nuestro caso hemos encontrado ciertos problemas a la hora de procesar nuestro data set, de tal manera que hemos decidido aplicar otro que nos ha brindado el cálculo de unos thetas o pesos óptimos para el problema que queremos resolver. El modelo elegido ha sido el “Nelder\_Mead. Hemos probado con otros modelos de minimización pero los resultados obtenidos no han sido tan convincente y se alcanzaba la excepción “divided by zero” demasiado proto en el ciclo de cálculo de los pesos, obteniendo resultados no deseados.

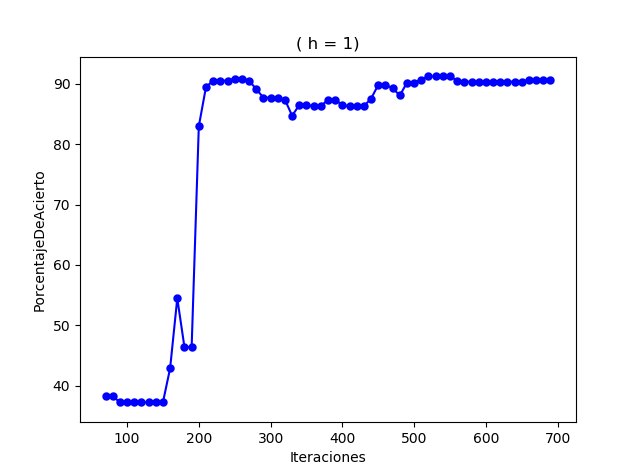
Este modelo tiene una peculiaridad que difiere del modelo TNC y es que se basa en el número de iteraciones para hacer los cálculos pertinentes para calcular las thetas óptimas para la predicción. De esta manera hemos creído conveniente experimentar por la afinidad de la predicción en función del número de iteraciones lanzadas para el algoritmo de minimización.

Como hemos mencionado anteriormente, nuestra regresión logística está regularizada, de tal manera que existe un parámetro de regularización que podría tener un gran impacto en la predicción de resultados, sin embargo, el modelo de minimización implementado y debido a la naturaleza de nuestro data set este tiene muy poco impacto en los resultados obtenidos. De esta manera lo hemos fijado a 1 y hemos hecho el análisis de eficacia en la predicción variando únicamente el número de iteraciones para el algoritmo de minimización.

Una vez tenemos los pesos óptimos podemos utilizarlos para calcular la predicción de nuestra regresión logística procesándolos con la función de normalización y posteriormente por la sigmoide.

Comparamos directamente el resultado con los valores reales

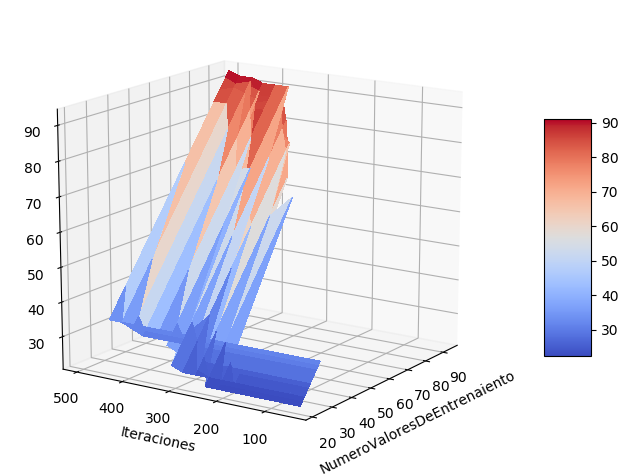
Los resultados obtenidos son los siguientes:



Podemos comprobar que para números bajos de iteraciones los pesos o thetas óptimos calculados nos llevan a unos porcentajes de aciertos muy bajos, pero a medida que aumentamos encontramos un punto a partir del cual los porcentajes son bastante prometedores y estabilizándose llegados a las 500 iteraciones y brindándonos unos porcentajes de acierto en torno al 90%.

Sin embargo, en este caso, hemos utilizado todos los valores de la data base como valores de entrenamiento y como valores de verificación, donde lo verdaderamente interesante sería ver como de buena es nuestra regresión logística prediciendo valores nuevos a partir de una serie de valores de entrenamiento.

De esta manera hemos lanzado la regresión logística variando el número de iteraciones y el número de valores de entrenamiento, tomados de nuestra data base, siendo los restantes, que no se han seleccionado para el entrenamiento, los valores de verificación.

Los resultados obtenidos los podemos representar en una gráfica 3D en el que se muestre el porcentaje de acierto de la red neuronal en función del número de iteraciones en la minimización y el número de valores de entrenamiento: 

Vemos como aumentando el número de iteraciones y el número de valores de entrenamiento aumenta el porcentaje de acierto, llegando a alcanzar valores al 90% como ocurría con anterioridad.

(Nótese que a la hora de representar la gráfica los valores mostrados para el número de iteraciones abarcan desde el 20 al 90, sin embargo estos números representan un rango que va desde los 50 valores de entrenamiento a los 500, esto puede ser debido a algún fallo a la hora de lanzar la representación de los ejes en 3D). “Ahora mismo hay un par de líneas ocultando los valores “

**Redes Neuronales:**

Uno de los métodos o aproximaciones más utilizadas dentro del aprendizaje automático son las redes neuronales. Estas son una representación artificial mediante programación y código de una red neuronal humana.

Una red neuronal consta de una serie de capas importantes y sobre las que se realizan las operaciones correspondientes para llegar a la predicción buscada en función de las variables de entrada.

En nuestro caso la red neuronal constará de una capa de entrada, una capa oculta y una capa de salida.

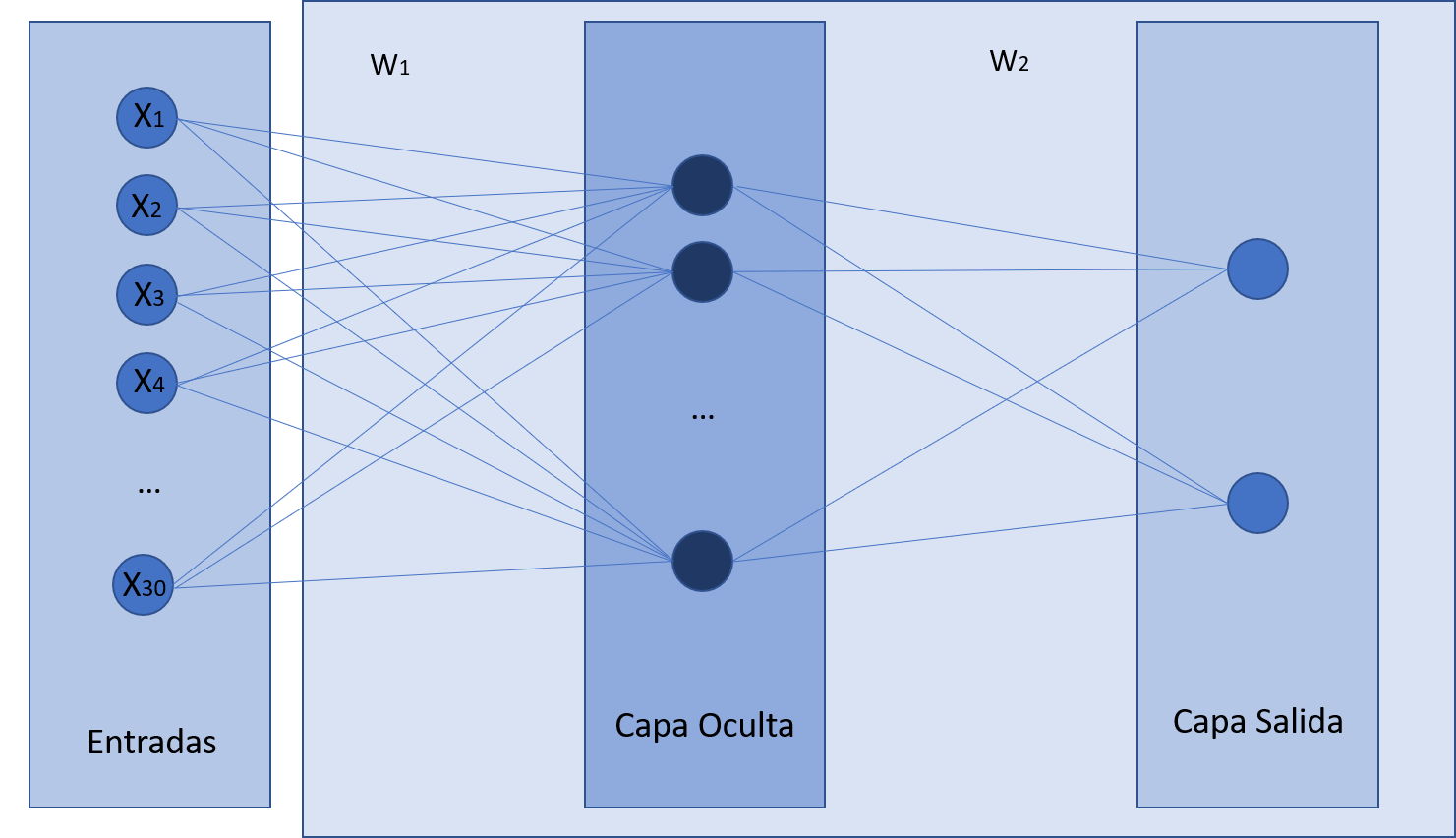
El tamaño de la capa de entrada se corresponde con el numero de variables que construyen la predicción de nuestro caso y que se corresponde con el número de columnas de nuestra base de datos.

La teoría nos dice que basta con utilizar una sola capa oculta y que la gran mayoría de problemas prácticos son resolubles con una única capa de esta naturaleza.

En cuanto a la elección del número de neuronas en la capa intermedia haremos una primera aproximación ciñéndonos a la topología de la regla de la pirámide geométrica, esto es, el número de neuronas en la capa oculta se corresponde con la raíz cuadrada del número de neuronas en la salida veces el número de neuronas en la entrada (). Números muy altos incrementan mucho el tiempo de entrenamiento y números demasiado pequeños pueden ser causantes de Underfitting, es decir, que no se llegue a resolver el problema.

En cuanto al número de neuronas en la capa de salida haremos corresponder una por cada resultado que podemos obtener, es decir, dos (si el tumor es maligno (1) o benigno (0)).

Así, si tenemos 30 neuronas en la entrada y dos en la salida, la regla nos dice que el número de neuronas en la capa oculta puede rondar las 8, que será el número con el que trabajemos.



Implementación de la red Neuronal:

Cuando pensamos en trabajar con una red neuronal pensamos principalmente en dos fases claramente diferenciadas. La fase de propagación o “Forward propagation” y el cómputo de gradiete o Backpropagation.

La fase de propagación hace referencia al recorrido de “izquierda a derecha“, desde la entrada de la red neuronal a la salida para obtener los valores de activación que van emitiendo los resultados a las siguientes conexiones hasta obtener los valores finales.

La fase de Backpropagation es la siguiente a tener en cuenta, esta a su vez se divide en otras dos fases: una primera de propagación y una segunda de actualización de pesos (que son los parámetros que definen el estado y en qué medida van a ser activadas las neuronas adyacentes a una dada).

En esa primera fase de propagación se hace una primera pasada hacia adelante para generar las activaciones de salida de la red para posteriormente hacer una pasada hacia atrás desde las salidas para generar los deltas de todas las neuronas de la salida y de las pertenecientes a la capa oculta.

Como hemos comentado, la segunda parte trata sobre la actualización de los pesos.

El primer conjunto de pesos se asigna de manera aleatoria y es la propia Red Neuronal, mediante el algoritmo, la que “aprenda” a ajustar el valor de los pesos para llegar al resultado o salida correcta.

Posteriormente iremos procesando los pesos hasta obtener los óptimos que nos reduciéndose en el proceso el coste e iterando hasta llegar al mínimo, nuestro objetivo.

Repetiremos este procedimiento hasta que el resultado obtenido sea el satisfactorio.

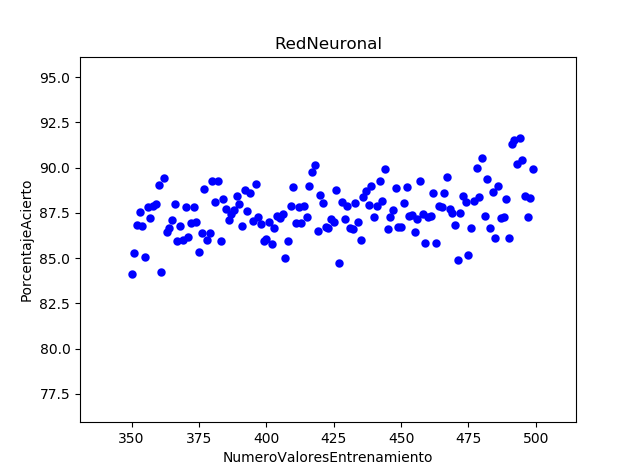
Debemos tener en cuenta que trabajaremos bajo la influencia de una tasa de aprendizaje, como mencionábamos anteriormente, y que va a influir en la velocidad y calidad de la red. Por motivos prácticos y debido a que ya hemos trabajado con redes neuronales anteriormente, la tasa de aprendizaje con la que trabajaremos al principio será equivalente a la utilizada en redes previas y que será de valor 1.

Trabajaremos por tanto con una red neuronal de 30 neuronas para la entrada (que se corresponde con el número de variables para tener en cuenta), con dos neuronas a la salida (que se corresponde con los resultados posibles, maligno o benigno) y con 8 neuronas en nuestra única capa oculta, siguiendo la regla de la pirámide.

Una vez aplicados los algoritmos pertinentes y tras haber estructurado la red neuronal hemos podido comprobar su efectividad. Hemos lanzado la red neuronal pasándole como parámetros de entrada todos nuestros casos de análisis para ver cuáles eran las perdiciones de nuestra red y el resultado lo hemos comparado con los resultados reales. El porcentaje de acierto para una tasa de aprendizaje de 1, 8 capas ocultas y el 100% del data set como valores de entrenamiento y como valores de análisis ronda en torno al 93%, que se corresponde con un resultado más favorable y preciso que el obtenido en algoritmos como el de descenso de gradiente y que nos harían pensar en el uso de una red neuronal como la programada para estudios como el que estamos llevando antes de cualquiera de los otros algoritmos utilizados.

Llegados a este punto sería interesante modificar ciertos parámetros de la red neuronal para ver como afecta a los resultados obtenidos y poder de esta manera elegir el mejor escenario para la predicción de valores.

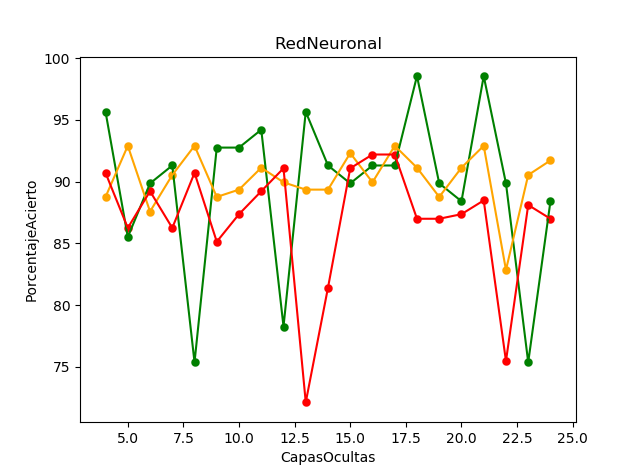
Un primer paso podría ser variar el número de valores de entrenamiento, y es que nuevamente el objetivo es entrenar la red neuronal con una serie de valores y posteriormente pasarle unos nuevos comprobando si predice adecuadamente los resultados. Hemos lanzado nuestra red neuronal variando el número de valores de entrenamiento utilizados de manera similar a como lo hemos hecho anteriormente, de tal manera que los valores de comprobación son los restantes que no hemos utilizado para entrenar la red neuronal:



Podemos ver como a medida que entrenamos más nuestra red neuronal (más valores de entrenamiento), esta nos da mejores resultados.

Teniendo esto en cuenta, el siguiente análisis lo haremos sobre una red entrenada con unos 450 valores de entrenamiento, utilizando los 110 restantes para comprobar la eficacia de nuestra red neuronal.

Nuestro primer análisis fue sobre una red que siguiendo la regla de la pirámide constaba de 8 capas ocultas. Lanzaremos ahora nuestra red neuronal variando el número total de capas ocultas y veremos cual es el resultado obtenido. Hemos aprovechado también en este paso para comparar las posibles diferencias que existirían al variar el número de iteraciones para nuestro modelo de minimización, los resultados son los siguientes:

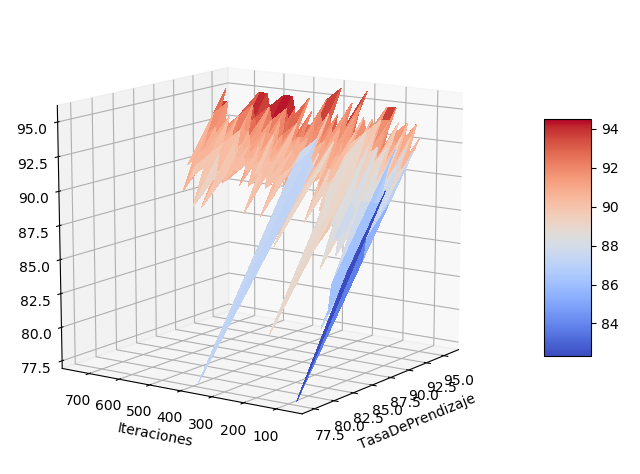


La gráfica verde, naranja y roja representan la variación en la precisión de acierto en función del número de capas ocultas de redes en las que se ha iterado un número de 700, 500 y 300 respectivamente veces a la hora de realizar la minimización para obtener los pesos óptimos.

Podemos comprobar que la gráfica verde predomina sobre las otras dos, indicándonos que a mayor número de iteraciones en nuestro método de cálculo de pesos óptimos obtendríamos mejores resultados, de la misma manera que los obtenemos al aumentar el número de capas ocultas. Si que es cierto que las gráficas también nos dicen que a medida que incrementamos mucho el número de iteraciones la dispersión en las predicciones obtenidas aumenta. Esto puede ser debido a que al aumentar el número de iteraciones nuestro modelo de minimización alcanza de manera más habitual situaciones de dividendo por 0 que pueden ser los que nos están dando esos porcentajes más bajos.

En un caso práctico eliminaríamos de la ecuación esas desviaciones aleatorias quedándonos con los valores más próximos al resto.

Por último, hemos decidido lanzar un estudio en el que se compruebe como afectan de manera conjunta el número de iteraciones y la tasa de aprendizaje



Podemos comprobar como a medida que aumenta el número de iteraciones y la tasa de aprendizaje ganamos precisión en la predicción, así como que la dispersión en los valores predichos se reduce, obteniéndose los valores más altos para las iteraciones más numerosas y nuestras tasas de aprendizaje más altas (analizado desde tasas de aprendizaje entre 0.2 y 5).

Utilizando los valores de iteraciones para la minimización, número de capas ocultas y número de valores de entrenamiento obtenemos unos porcentajes que alcanzan porcentajes en torno al 95%, superiores a los obtenidos anteriormente en nuestra regresión logística. Sin embargo, cabe resaltar que este incremento en la precisión se penaliza también con un incremento en los tiempos de ejecución y cálculo de resultados, que se ha hecho notable durante el desarrollo de la práctica y el entrenamiento de nuestras redes neuronales.

SVM:

Los SVM o máquinas de soporte vectorial son una serie de algoritmos de aprendizaje supervisado que están especialmente diseñados para resolver problemas de clasificación y regresión.

De la misma manera que hacíamos con nuestra red neuronal se debe utilizar una serie de valores para entrenarla y construir de esta manera un modelo capaz de predecir valores de la misma naturaleza que los que hemos utilizado como entrenamiento.

Será gracias a este modelo y a la aplicación de funciones Kernel como seremos capaces de elaborar un nuevo sistema para predecir resultados para nuestra base de datos.

La implementación a nivel de código es en este caso más sencilla, ya que la mayor parte de la algoritmia viene dada por la propia máquina de vectores.

Para empezar seleccionaremos un conjunto de valores de entrenamiento como hacíamos anteriormente, 400 para el entrenamiento y los restantes para comprobar el acierto en las predicciones.

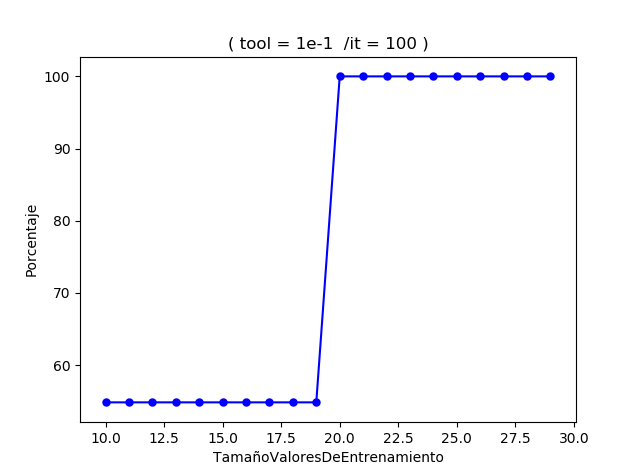
Con esos valores utilizamos los parámetros óptimos sigma y C de nuestra SVM y que a continuación pasaremos al modelo de entrenamiento para obtener la función gausiana óptima que nos permitirá hacer las predicciones correspondientes.

El modelo de entrenamiento elegido es en nuestro caso el gausiano que hace uso de la función kernel para realizar los cálculos correspondientes.

Después de obtener la función correspondiente podemos llamar directamente a su función de predicción a la que por medio del kernel gausiano y pasándole los valores de comprobación nos devolverá un vector de predicciones de la naturaleza de los resultados o variable dependiente de nuestro data set.

El valor obtenido es del 100%, un resultado que supera con creces al obtenido en la red neuronal y que nos da una fiabilidad excelente.

Sin embargo, tras obtener este resultado nos hemos propuesto investigar hasta que punto este sistema de predicción es eficaz y hemos querido comprobar el porcentaje de acierto dependiendo del número de valores de entrenamiento.

Los resultados son los siguientes:

Vemos como los resultados obtenidos son más que prometedores y es que basta con un número algo superior a los 20 valores de entrenamiento para obtener unas predicciones del muy cercanas al 100%.

Esto convierte a las SVM en el mejor modelo para la predicción de valores de la naturaleza de nuestra base de datos. Si que es cierto y necesario recalcar es que aquí es, aún si cabe, mucho más notable el coste en tiempo de cálculo frente a las Redes Neuronales y a la regresión logística, y es que estamos hablando de procesos que tardan entre los 2-5 minutos en mostrar resultados, que teniendo en cuenta la “simplicidad” de nuestra data set, son tiempos más que considerables.