**Aprendizaje Automático Proyecto de la asignatura.**

**Curso 2019–2020**

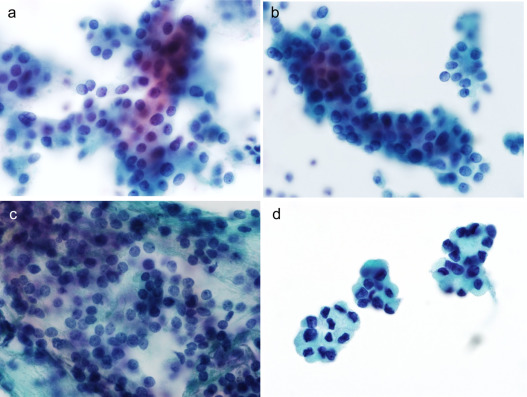
**Aprendizaje automático en la evaluación del cáncer de pulmón**

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.

El data set elegido contiene información relevante en la investigación del cáncer de pulmón en 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 data 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 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 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.

**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 nos ceñiremos 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.

En cuanto a la tasa de aprendizaje utilizaremos un valor ya utilizado en redes neuronales anteriores y que nos ha aportado buenos resultados, quedando el valor fijado a 1.