**INTELIGENCIA ARTIFICIAL - INGENIERÍA INFORMÁTICA**

**PECL2**

Adrián Martínez Diaz

Raúl López Llana

Elena Martín Naranjo

Luis Ángel Barderas San Segundo

Tareas a realizar:

**1.** El problema consiste en una simulación de diagnóstico de cáncer, lo cual es en esencia, la clasificación de los datos según el tipo de tumor, benignos o malignos.

Para crear el modelo para el estudio y clasificación de los datos obtenidos de la página web se utilizarán técnicas de aprendizaje inductivo, analógico y bayesiano; estas serían k-vecinos próximos, árboles de clasificación y naive Bayes, las cuales permiten predecir la clase de un dato introducido en función de los valores de sus atributos.

Se debe escoger razonadamente un conjunto de entrenamiento y test adecuado para la creación del modelo, así como los valores seleccionados para los hiper-parámetros en busca del mejor ajuste posible de los modelos. También es necesario realizar una comparación entre las ventajas y desventajas de las distintas técnicas.

**2.** El tamaño de training data debe ser bastante grande para que al entrenar el modelo obtengamos las predicciones más exactas posibles.

Para el modelo knn debemos tener en cuentas además la “maldición de la dimensionalidad”, que provoca que cuanto más aumentemos el número de dimensiones de un conjunto de datos, más dificil será encontrar elementos vecinos próximos al elemento que el modelo trata de clasificar. En nuestro caso, la dimensionalidad de los datos es 30, lo que nos lleva a aumentar aún más el tamaño del training data.

En base a esto, hemos decidido tomar los siguientes conjuntos de datos:

training -> 80 % de los datos                   testing  -> 20 % de los datos

**3.** El hiper-parámetro k (número de vecinos**)** puede ser un número entre 1 y 455, que es el total de elementos del training data (80% del total de los datos originales).

Para determinar el hiper-parámetro k óptimo, se ha evaluado la precisión que tendría el modelo con cada valor de k posible a la hora de realizar predicciones sobre el testing data. Aquel modelo con la precisión máxima será el óptimo, y se debe por tanto guardar el valor de k que lo ha originado.

En este caso k = 9, con una precisión de 0.965, es el número de vecinos que maximiza la precisión del modelo.

**4.** En el caso de los árboles de clasificación puede darse un sobreajuste del modelo, que provocaría el aprendizaje de errores inherentes al training data que disminuirían la precisión del modelo al evaluarlo mediante el test data.

Por tanto, el sobreajuste es el factor a evitar. El hiper-parámetro que usaremos será max\_depth, que indica la profundidad máxima que puede tener el árbol. Reduciendo el valor de max\_depth se limita el desarrollo del árbol, evitando el sobreajuste.

En este caso se observa que se produce un sobreajuste del modelo (precisión del 100% sobre el training data pero del 91% sobre el test data) con una profundidad de 6 o más niveles. El valor óptimo sería max\_depth = 3, que otorga una precisión del 97% y 96%, la mejor posible respecto a los datos de entrenamiento y test.

**5.** Para naive Bayes el hiper-parámetro que permite ajustar el modelo será la distribución prior, que define la probabilidad previa de los sucesos que se están sometiendo a clasificación.

En este caso no contamos con información suficiente para estimar de forma razonada que un tumor sea maligno a benigno, por lo que usaremos priors no informativos, es decir, se estimará la misma probabilidad para ambas opciones, de modo que la distribución prior sea uniforme.

Por tanto, las probabilidades previas establecidas para los sucesos de tumor benigno y maligno son 0.5 ambas.

**6.** Comparar con algún criterio los resultados obtenidos con los 3 modelos (20 líneas máximo)

Para evaluar y comparar las distintas técnicas de aprendizaje automático hemos calculado el porcentaje de precisión de las predicciones de los 3 modelos, tanto para el conjunto de entrenamiento como para el conjunto de test.

En todos los casos se ha tratado de maximizar la precisión mediante el ajuste de los hiper-parámetros; no obstante, ese no es el único aspecto que se ha valorado. De igual modo es importante que la precisión de los conjuntos de entrenamiento y test sea parecida, puesto que un desequilibrio notable entre ambas puede indicar un sobreajuste del modelo al training data o el caso contrario, lo cual resulta en errores y una menor precisión del modelo al realizar predicciones sobre conjuntos de datos diferentes.

La precisión de las predicciones realizadas sobre el training data y el testing data respectivamente son las siguientes:

* Knn: 93% y 96%
* Árbol de clasificación: 97% y 96%
* Naive Bayes: 94% y 92%

Por tanto, teniendo en cuenta los criterios mencionados el mejor modelo sería el árbol de clasificación.

**7.** Mejoras o críticas (10 líneas)

- Mejora del algoritmo de cálculo de k óptima en la técnica de knn, puesto que el algoritmo actual, al evaluar todos los valores posibles de k, puede ser bastante lento en caso de que el tamaño del conjunto de datos sea mayor.

- Estudiar y evaluar el modelo mediante el uso de la “complejidad del error” para realizar un proceso de prunning óptimo, que produzca mejores resultados que el uso de max\_depth para limitar la complejidad (en niveles) del árbol.

- Determinar una distribución prior más precisa y ajustada a la realidad, puesto que el uso de una distribución prior uniforme, sesga la precisión del modelo de naive Bayes.

- Evaluación de los atributos del modelo, puesto que algunos atributos pueden tener distinta relevancia o ser dependientes entre ellos, lo que disminuye la precisión de los modelos, construidos asumiendo igual relevancia e independencia entre los atributos.

**8.** Replicar todos estos resultados, empleando obligatoriamente un código similar al utilizado en los puntos anteriores, sobre otro problema médico de características similares (clasificación binaria). Pueden consultarse bases de datos libres como las contenidas en el repositorio kaggle.com. (2 puntos, 40 líneas máximo).