Machine Learning

Tarea #2 Entrega: Noviembre 17 11:59.

---

1. En biología celular, la apoptosis es el fenómeno de muerte programada de las células. La alteración de este fenómeno es crucial en el desarrollo de enfermedades como el cáncer (en el que se inhibe la apoptosis) o el Alzheimer (donde la apoptosis se produce cuando no debiera). En el tratamiento de cáncer se busca producir la muerte de células malignas mediante la activación de la señal de apoptosis en estas células. Esto se puede lograr mediante la modificación de la expresión de uno o más genes que intervienen en el camino de la señal a nivel celular. La idea es producir la muerte de las células malignas sin afectar otras células, proceso que se denomina eliminación selectiva.

   El objetivo de este ejercicio es diseñar un clasificador que permita decidir si una combinación de drogas dada que actúa sobre un conjunto seleccionado de 15 genes en cierto dosaje produce o no eliminación selectiva de una clase de células malignas. Los datos a utilizar están en los archivos `xtrain.txt` y `ytrain.txt` adjuntos. Un dato de entrada consiste en la dosis aplicada de cada droga, y las etiquetas indican si la fila corresponde a eliminación selectiva (y=1) o no (y=-1). Se incluyen 2000 datos de entrada de prueba en el archivo `xtest.txt`. En cada uno de los siguientes casos su objetivo es obtener un modelo que tenga probabilidad de error tan baja como sea posible. Usted debe adjuntar a su reporte las etiquetas que cada uno de sus modelos finales calcula en los datos de prueba. Con estas etiquetas yo estimaré la probabilidad de error de sus modelos.

   a) Resuelva el problema de clasificación utilizando un SVM con kernel polinomial. Determine apropiadamente los valores del grado del polinomio, y la constante C o ν (dependiendo del problema de optmización que resuelva).

   b) Resuelva el problema de clasificación utilizando un SVM con kernel Gaussiano. Determine apropiadamente el ancho del kernel y la constante C o ν (dependiendo del problema de optmización que resuelva).

2. Cada línea del archivo `news.text` contiene una micronoticia[1] pertenecientes a la categoría de la línea correspondiente en el archivo `labels.txt`. Las categorías son: sports, business, entertainment, us, world, health, sci&tech. Usted debe resolver el problema de clasificación binaria de acuerdo a la siguiente tabla:

Último dígito de su código de estudiante	Clase 1	Clase 2
2,5,6 u 8	us + world	sports + entertainment
0 o 1	sports	sci&tech + business
3 o 4	us	health + entertainment
7 o 9	world	entertainment + sci&tech

   Resuelva el problema de clasificación utilizando un SVM con el kernel SSK (String Subsequence Kernel) visto en clase[2].

[1] Cada línea está terminada en `\n`, de manera que por ejemplo en Python, si `f` es el handle del archivo `news.text`, `f.readline()` retorna una micronoticia.

[2] Puede encontrar una implementación de este kernel para Python en el Shoguh Toolbox en http://shogun-toolbox.org