Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Otro ejemplo con SVM

Docente: Abraham Sánchez López

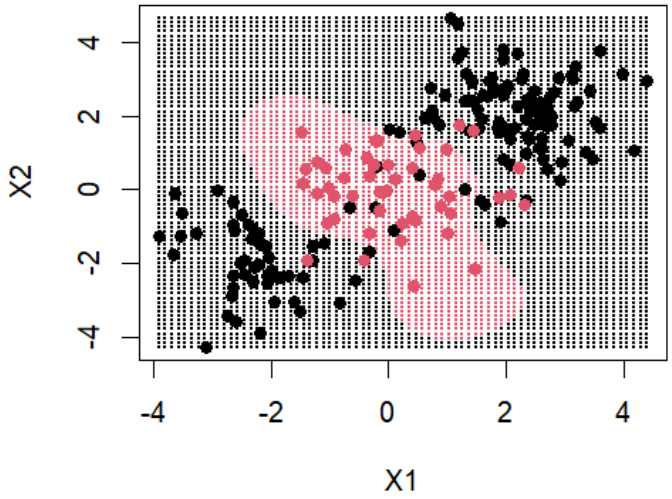
Alumno	**Matrícula**
Taisen Romero Bañuelos	202055209

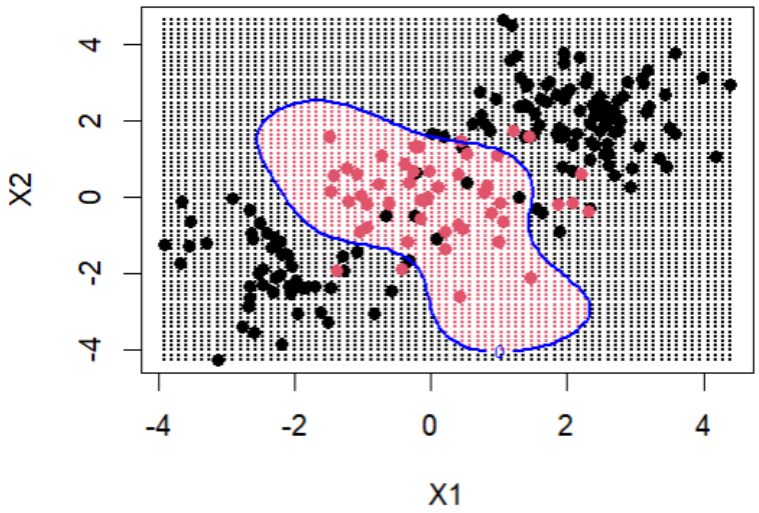
SVM

A manera de síntesis, en el PDF se crea un conjunto de datos ficticio para poder hacer un grafo de dispersión. Esta dispersión es lineal porque se pueden separar las dos clases (grupos de datos) mediante una línea recta. Se hicieron algunas modificaciones en los ploteos pero lo interesante empieza cuando se nos presenta un conjunto de datos mezclado. Mi primera impresión fue que los datos eran iguales pero luego de unos comandos se hizo clara la distinción de clases. Me resultó sencillo interpretar esa distribución mezclada de la información luego de recordar el ejemplo de la montaña, en donde si se consideraba una dimensión más los datos adquirían una representación más intuitiva. Este método de SVM se centra más en delimitar las clases usando curvas, por lo que supongo que la adición de una dimensión no forma parte de la metodología usada, sin embargo, considero que tener conocimiento de otros métodos puede ayudar a comprender los datos incluso si no forman parte del proceso de resolución del problema planteado.

En fin, para nuestro segundo grafo de dispersión hay una distribución no lineal por lo que se optó por usar un kernel radial. Inicialmente se sombreó de negro el límite de decisión de Bayes (la frontera teóricamente óptima que separa dos clases de manera probabilística, minimizando la tasa de error) y después se trazó una línea azul que representa la frontera del SVM, que es como el límite de Bayes pero con la diferencia de que solo busca una separación clara entre clases (de forma determinista, es o no es), sin considerar probabilidades. Dado que el límite de Bayes considera zonas de incertidumbre la línea azul encierra un área mucho menor que la sombreada de negro, lo que me lleva a pensar que un límite generaliza más que el otro y también que se puede correr el riesgo de que la línea azul se sobreajuste.

Sobre la parte final del PDF, nada más viendo el código uno se puede hacer a la idea de cómo son los datos, pues, se ve como unos datos se desplazan arriba a la derecha y otros abajo a la izquierda, pero ploteando la información se puede confirmar que se trata de un conjunto de datos con una distribución no lineal, por lo que lo más adecuado sería usar un kernel radial (RBF).





A primera vista parece que todo está bien, pero el que la zona sombreada de rojo y el límite azul parecen coincidir casi perfectamente me parece extraño. Podría ser un signo de sobre ajuste, pero como la frontera azul es fluida y sin patrones erráticos creo que podemos descartar el sobreajuste. Entonces, ¿qué podría significar ese detalle?, quizás signifique que el modelo está clasificando con mucha seguridad, lo cual puede ser bueno pero a la vez una señal de sobreajuste. Lo ideal sería hacer una evaluación al modelo para revisar su precisión, pero como no contamos con datos de prueba (sólo de entrenamiento) creo que no será algo posible.

Si intentamos resolver este problema con regresión logística tendríamos dificultades porque la regresión logística traza una línea recta, por lo que no sería suficiente para separar las clases de este problema.