Ingeniería Matemática
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Tarea 5

MA5203 – Aprendizaje de Máquinas Probabilístico

Nombre: Sebastián Parra

Profesor: Felipe Tobar

Auxiliares: Alejandro Cuevas
Cristóbal Valenzuela
Lerko Araya
Mauricio Campos
Nicolás Aramayo

Fecha: 15 de agosto de 2018

Resultados: Parte 1

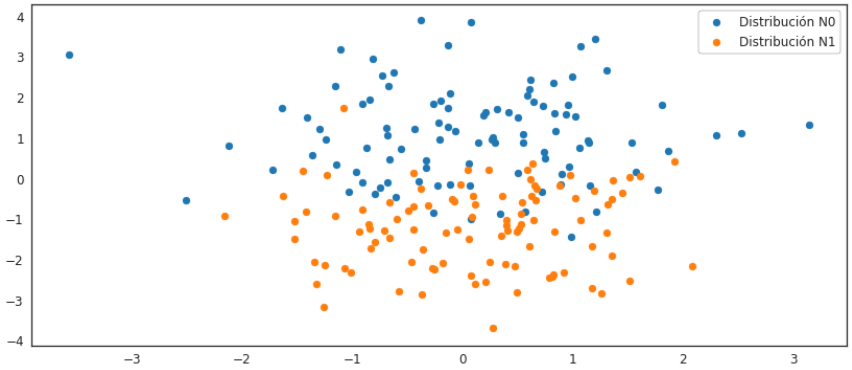

Figura 1: Datos generados para n=100, λ=1.2

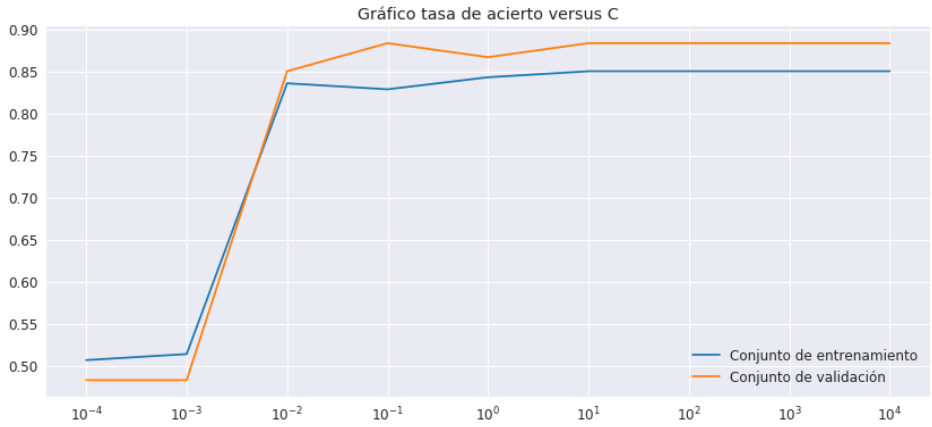

Figura 2: Tasa de acierto obtenida sobre los conjuntos de entrenamiento y validación, para una SVM con distintos valores de C

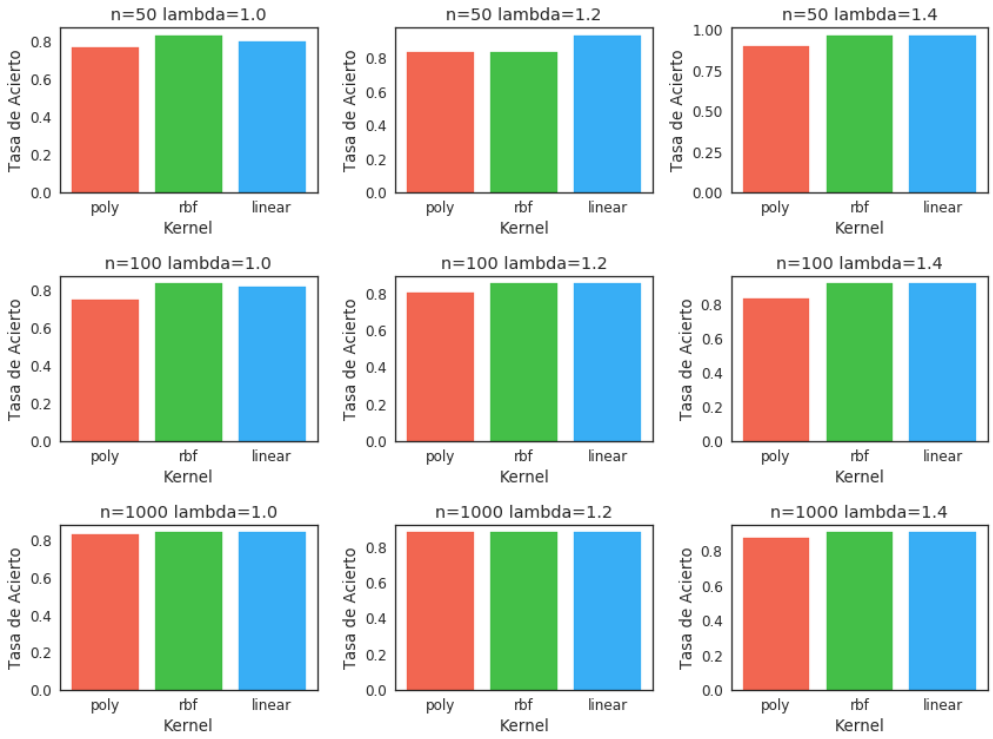

Figura 3: Tasas de acierto obtenidas para distintos kernels, con distintos valores de n y lambda.

Resultados: Parte 2

Tabla 1: Características escogidas de la base de datos

Característica
Age
eduaction_num
hr_per_week
ocupation
race
sex

Tabla 2: Hiperparámetros de la SVM escogida

Hiperparámetro	**Valor**
C	10
Kernel	'rbf'
Gamma	0.1

Tabla 3: Medidas de desempeño de la SVM sobre el conjunto de prueba

Métrica	**Valor**
Precisión	0.79
F1 Score	0.79
Indice de Younen	0.39

Discusión: Parte 1

En primer lugar, a observar el gráfico de la figura 2 se puede notar que, tanto para el conjunto de entrenamiento como de validación, el clasificador comienza a tener mejores tasas de acierto para C grande (C > 10). Esto implicaría que para este caso particular es de mayor relevancia clasificar correctamente la mayor cantidad de puntos del conjunto de prueba, en vez de maximizar el margen de separación de clases. Además, observando los gráficos de la figura 3, se puede destacar que, en general, a medida que aumentan tanto n como lambda las tasas de acierto del clasificador aumentan, lo cual se explica sabiendo que valores grandes de n logran aproximar mejor las distribuciones muestreadas (facilitando el aprendizaje), mientras que valores grandes de lambda incrementan la separación entre clases. Finalmente, considerando que el mejor clasificador posible para las distribuciones generadas (no necesariamente los muestreos) es la recta y = 0, tiene sentido que en la mayoría de los casos el mejor clasificador sea el con kernel lineal, mientras que clasificadores que usan kernels más complejos como el rbf suelen presentar peores resultados debido a que se sobreajustaron al ruido del modelo.

Discusión: Parte 2

Para el experimento con datos reales se notaron las siguientes características para la base de datos: *Age, Workclass, education, eduaction_num, marital, ocupation, relationship, race, sex, capital gain, capital loss, hr_per_week, country,* y finalmente *income*, que corresponde a las etiquetas con las que se debe clasificar. Sin embargo, dada la naturaleza del problema (predecir los ingresos de una persona), se decidió eliminar las características *Workclass, marital, y relationship*, puesto que no se consideró que el estado civil/familiar o el sector de trabajo (público, privado, etc.) pudiese tener una correlación fuerte con lo que se buscaba clasificar. Luego, inspeccionando la base de datos también se percató que las columnas *capital gain* y *capital loss* presentaban una gran cantidad de filas con ceros, por lo que se concluye que no se podría sacar mucho provecho a esta información y también se retiran de la lista de características a considerar. Posteriormente, en un esfuerzo por descifrar el significado de la columna *education_num*, se compara con la característica *education*, notando que ambas columnas poseen el mismo número de clases, compartiendo también la frecuencia de cada clase. Teniendo esto en mente, se procede a visualizar ambas columnas de manera aislada, y se corrobora la hipótesis de que ambas columnas entregan la misma informacion, sólo que en formatos distintos, ante lo cual se decide eliminar la columna *education* ya que se prefiere manejar datos de tipo numérico, además que estos en particular poseen una relación de orden (i.e. un doctorado es más que educación primaria). Finalmente, se decide graficar las frecuencias de las clases de cada una de las características que quedan, notando así que la gran mayoría (casi el 90%) de los datos obtenidos son de individuos estadounidenses, y que demás las clases se encuentran desbalanceadas, teniéndose que cerca de un 75% de los datos son de individuos con ingresos menores a 50 mil dólares, por lo que cualquier clasificador creado deberá tener una tasa de acierto mayor a este valor. Debido a las frecuencias observadas para la columna *country*, se decide eliminar ya que su sesgo no permite que sea un buen criterio de clasificación. Además, se opta por conservar las columnas de raza y sexo, debido a que en Estados Unidos estas variables sí pueden influir en los ingresos de una persona, obteniéndose como resultado de todo este proceso la tabla de características mostrada en la Tabla 1.

Después de haber seleccionado las características, se procede a convertir las variables categóricas en numéricas utilizando codificación *one-hot*, puesto que no poseen relaciones de orden que puedan justificar una codificación por enteros. Posteriormente, se procede a construir un clasificador SVM, cuyos hiperparámetros se consiguen mediante un proceso de validación cruzada, probando los kernels lineal y rbf, con valores de C={0.1, 1, 10, 100}, y valores de gamma={0.1, 1, 10, 100} en el caso rbf, utilizando además el índice de Younen como métrica de desempeño. Los resultados obtenidos se pueden apreciar en la Tabla 2. Finalmente, en la tabla 3 se pueden apreciar distintas métricas de evaluación utilizadas para medir el desempeño del clasificador sobre el conjunto de prueba, utilizando los hiperparámetros obtenidos mediante validación cruzada.