Universidad del Rosario, Facultad de Economía
Introduction to Data Science
Examen

Junio 28, 2021

Primera Parte

Ejercicios Teóricos (25%)

1. Supongamos que usted tiene un conjunto de datos de 100 personas que han tomado un préstamo con el banco para el cual trabaja. La información incluye el género, con 50 hombres y 50 mujeres. También el estado civil, con 50 casados y 50 solteros. De esas 100 personas, 50 hicieron default. Usted está entrenando un ábol de decisión para predecir default, con base en género y estado civil. Supongamos que en el grupo de los hombres, el 70% hizo default, frente a un 30% en las mujeres. En los casados, el 40% hizo default, frente a un 60% en los solteros. Con base en estos datos, y suponiendo que se utiliza el índice de entropía para decidir las divisiones, ¿en qué variable se alcanza la mayor ganancia en información y de cuánto es dicha ganancia? Justifique su respuesta

2. Considere los siguientes datos:

Datos Originales	1	1	1	1	1	1	1	1	2	2	2	4	4	4	4
Split en Variable A	1	1	1	1	1	1	1	1	2	2	2	4	4	4	4
Split en Variable B	1	1	1	1	1	1	1	1	2	2	2	4	4	4	4

   En el contexto de los árboles de regresión y árboles modelo, supongamos que el algoritmo debe decidir entre hacer el split en la variable A o en la variable B, con los subconjuntos resultantes que muestra la tabla. Con esta información, resuelva:

   (a) ¿Cuál es el cambio en la pureza si se divide en A? ¿Cuál si se divide en B? ¿En qué variable debe entonces hacerse la división?

   (b) Explique, claramente, qué procedimiento debe seguirse si quiere utilizarse la técnica de los árboles de regresión para clasificar objetos en este contexto.

   (c) Explique, claramente, qué procedimiento debe seguirse si quiere utilizarse la técnica de los árboles modelo para clasificar objetos en este contexto.

3. Considere la matriz de confusión representada en la Tabla 1, para un modelo que intenta predecir existencia de células cancerígenas. Con base en estos resultados:

   (a) Determine el nivel de precisión y la tasa del error de modelo.

   (b) Determine las medidas de sensibilidad (*sensitivity*) y especificidad (*specificity*).

   (c) ¿Qué miden los estadísticos de machine learning *precision*, *recall* y *F-Score*? ¿Cuánto dan dichos estadísticos en este caso?

Table 1: Matriz de Confusión para Predicción de Cáncer

	Prediccion		
Diagnostico	Benigna	Maligna	Total Fila
Benigna	77	0	77
Maligna	2	21	23
Total Columna	79	21	100

Table 2: Observaciones para un ejemplo hipotético

Obs.	X_1	X_2	Y
1	3	4	Rojo
2	2	2	Rojo
3	4	4	Rojo
4	1	4	Rojo
5	2	1	Azul
6	4	3	Azul
7	4	1	Azul

4. El objetivo de este ejercicio es construir un modelo de *Support Vector Machine* (SVM) basándonos en los datos de la Tabla 2.

   (a) En este ejemplo hay $n = 7$ observaciones en $p = 2$ dimensiones y sabemos la categoría a la que pertenece cada observación. Grafique estas observaciones.

   (b) En la gráfica, muestre cuál sería el hiperplano separador óptimo. Identifique tambinés los vectores de soporte. Ayuda: identifique el *maximum margin hyperplane* (MMH).

   (c) Escriba la ecuación del MMH. Ayuda: Identifique entre qué observaciones debe estar el plano y deduzca su correspondiente ecuación.

   (d) Con base en su respuesta al numeral anterior, caracterice el algoritmo SVM en este ejemplo. ¿Bajo qué circunstancias un objeto es clasificado como "rojo" y bajo qué circunstancias como "azul" de acuerdo con el algoritmo?

Ejercicios Empíricos (25%)

1. **Árboles de Decisión, *Bagging* y *Random Forests***

   El propósito de este ejercicio es predecir presencia de células cancerosas usando árboles de decisión, *bagging* y *random forests*. Para ello, usaremos la base `wisc_bc_data.csv` que utilizamos anteriormente.

   (a) Cargue la base en un data frame llamado `cancer`. Inspeccione la base con la función `str()`. ¿Cuántas observaciones tiene la base? ¿Qué nombre tiene el outcome que nos interesa predecir?

   (b) Limpie los datos para poder implementar el algoritmo. Asegúrese de eliminar la variable de identificación. Además, convierta el outcome de interés en un factor. Por medio de una tabla de proporciones, determine qué porporción de los casos son benignos y malignos.

   (c) Divida la base de datos en dos: 80% para entrenamiento y 20% para prueba. Para esto, aleatorice las filas que formarán parte de la base de entrenamiento y las de prueba. Fije la semilla como los últimos tres dígitos de su documento de identidad. Llame a sus bases de entrenamiento y prueba `cancer_train` y `cancer_test`. Inspeccione estas dos bases con la función `str()` y construya tablas de proporción para los outcomes en cada base. ¿Encuentra coherencia con las proporciones de la base original?

(d) Utilizando el paquete `C50` construya un árbol de decisión. ¿cuáles son las características de las biopsias más importantes al momento construir los subgrupos? ¿Qué porcentaje de precisión tiene el algoritmo sobre la base de entrenamiento? Realice la predicción sobre la base de prueba. Construya una tabla en la que represente en el eje X el verdadero diagnóstico de cada biopsia y en el eje Y las predicciones del modelo. ¿Qué tan preciso es el modelo? ¿Cuál es la proporción de falsos positivos? ¿De falsos negativos?

(e) Repita el ejercicio anterior, pero esta vez usando *bagging*. ¿Mejora la precisión del modelo si se utiliza este procedimiento?

(f) Ahora repita el ejercicio anterior, pero esta vez usando *random forests*. ¿Mejora la precisión del modelo si se utiliza este procedimiento?

2. **Support Vector Machines**

Ahora buscamos predecir presencia de células cancerosas usando el algoritmo SVM. Para ello, usaremos nuevamente la base `wisc_bc_data.csv`.

(a) Cargue la base en un data frame llamado `cancer`. Divida la base de datos en dos: entrenamiento y prueba. Para esto, aleatorice las filas que formarán parte de la base de entrenamiento y las de prueba. Fije la semilla en su número de identificación. Llame a sus bases de entrenamiento y prueba `cancer_train` y `cancer_test`.

(b) Utilizando el paquete `kernlab` construya un modelo SVM utilizando un kernel lineal. Construya la matriz de confusión utilizando el paquete `caret`. ¿Qué porcentaje de precisión tiene el algoritmo sobre la base de entrenamiento? ¿En qué cree usted que influye el uso de un kernel lineal?

(c) Nuevamente, usando el paquete `kernlab` realice la predicción sobre la base de prueba, pero en esta ocasión pruebe un kernel NO lineal. Construya la matriz de confusión utilizando el paquete `caret` ¿Qué porcentaje de precisión tiene el algoritmo sobre la base de entrenamiento? ¿Cuál es la proporción de falsos positivos? ¿De falsos negativos? ¿El cambio de kernel ayudó a mejorar el desempeño del modelo? Explique por qué los resultados cambian al modificar el tipo de kernel.

(d) Finalmente, entrene un nuevo modelo pero esta vez usando la técnica de *10-Fold Cross-Validation*. Construya la matriz de confusión y determine si el desempeño del modelo mejoró en comparación con los modelos anteriores.

3. **Vecino más Cercano**

Para predecir el éxito o fracaso de una campaña de marketing de un banco portugués que ofrece CDTs a sus clientes, trabaje con los datos `bank_sample.csv`. Esta base tiene muchas menos observaciones que la original y algunas de sus variables han sido modificadas. Nuevamente, el outcome de interés, `y`, indica si el cliente adquiere el producto o no luego de que se le ofrece telefónicamente.

(a) Lea los datos por medio de un objeto llamado `bank`. Usé la función `str()` para entender la estructura básica de los datos. ¿Cuántas observaciones hay? ¿Cuántas variables se incluyen?

(b) Use las funciones `table()`, `summary()` y otras que considere pertinentes para entender las variables. ¿Qué proporción de los clientes adquieren el producto?

(c) Reescale las variables de la base por medio de una normalización min-max de los datos. Para ello, cree una función que permita normalizar los datos. Tenga cuidado de hacer esto solamente para las variables independientes del modelo. Bautice a su nueva base de datos, tras la normalización, `bank_n`. Verifique, por medio de estadística descriptiva, que los datos estén en la misma escala y sean comparables.

(d) Cree un par de vectores, llamados `bank_train_y` y `bank_test_y`, que representen los outcomes de interés de los grupos de entrenamiento y prueba, respectivamente.

(e) Divida la base `bank_n` en dos partes: una con 3521 observaciones elegidas al azar para entrenar el modelo, y las restantes para probar el modelo. Llame a estas dos bases `bank_n_train` y `bank_n_test`, respectivamente. Fije como semilla los últimos tres dígitos de su documento de identidad para hacer esta aleatorización.

(f) Usando la heurística más utilizada para elegir k en este tipo de modelos, haga clasificación usando vecino más cercano (kNN). Presente en una tabla el resumen de las predicciones hechas para la base de prueba.

(g) ¿Qué tan precisos son los resultados? Considera usted que en este caso es preferible minimizar el número de falsos positivos o de falsos negativos? ¿Por qué? ¿Cree usted que el algoritmo kNN es el más apropiado para este tipo de datos?

Segunda Parte

Ejercicios Teóricos (25%)

1. Considere una red neuronal con 4 inputs, una capa intermedia con 2 nodos y una capa de salida con 3 nodos. Sugerincia: dibuje la red antes de comenzar el ejercicio

   (a) Cuántas matrices de pesos tendremos? Cuáles son sus dimensiones?

   (b) Utilizando notación matricial escriba como se lleva a cabo la propagación hacia adelante

   (c) Utilizando notación matricial escriba como se lleva a cabo la propagación hacia atras

   (d) Considere las siguientes matrices de pesos, función de activación sigmoid y bias igual a cero en toda la red. Si consideramos el vector $x = [2, 3, 1, 4]$, cuál es el valor del vector $\hat{y}$

$$W_1 = \begin{bmatrix} -2 & 1 \\ 2 & 1 \\ -2 & 2 \\ 2 & 1 \end{bmatrix} \qquad W_2 = \begin{bmatrix} 1 & 2 & -2 \\ -3 & -1 & -3 \end{bmatrix}$$

   (e) Asuma que la función de costos es cuadrática y que la y asociada a la x está dada por $y = [1, 0, 2]$. Calcule la función de costos en este punto

   (f) Asuma que la tasa de aprendizaje está dada por $a = 0.2$. Calcule una iteración hacia atrás para actualizar los pesos. Cuáles son los pesos nuevos? Cuál es la nueva $\hat{y}$? Cuál es el nuevo valor de la función de costos?

   (g) Proponga al menos dos modificaciones a este ejemplo que podrían mejorar la red. No es necesario resolver nada so argumentar

2. Suponga que una compañía de crédito está decidiendo crear un nuevo sistema para evaluar a sus posibles clientes. El nuevo sistema utiliza una red neuronal hacia adelante supervisada. Sugiera que necesita el banco antes de implementar su nuevo algoritmo. Discuta que problemas podrían surgir durante el proceso y las consecuencias de los mismos

3. Suponga que el gobierno le pide ayuda para seguir tendencias de Twitter durante la pandemia. El objetivo del gobierno es conocer el sentir de la gente respecto a distintas políticas. Describa brevemente los pasos que tendría que seguir para:

   - obtener los datos, incluyendo que clase de términos y como restringir la búsqueda en Twitter
   - limpiar los datos

   Recuerde tomar en cuenta el objetivo final, un análisis de sentimiento, para decidir como limpiar los datos. Discuta posibles problemas u omisiones de su método de búsqueda y limpieza.

4. En está pregunta construiremos un modelo del lenguaje con bi-gramas. Considera las siguientes 4 oraciones de una canción de Bad Bunny como el corpus.
   Considere las siguientes oraciones (de una canción de Bad Bunny) como el corpus

   Que cantamos bien borrachos Que bailamos bien borrachos Nos besamos bien borrachos los dos

   (a) Calcule la probabilidad de que la palabra bien aparezca después de la palabra cantamos, es decir $p(bien|cantamos)$

   (b) Qué supuestos utilizó para responder la pregunta anterior

   (c) Calcule la matriz de probabilidades, normalizándolas y aplicando el logaritmo natural

   (d) Cuál es la probabilidad de la frase 'bailamos los dos bien borrachos'?

Ejercicios Empíricos (25%)

1. **Redes Neuronales**

   Utilizaremos la red neuronal de una sola capa que programamos en clase y la base de datos de los billetes falsificados. En todos los ejemplos utilizar la varianza y la entropía como x's a menos que se indique explícitamente lo contrario.

   (a) Sustituya la función de activación sigmoid por una lineal. Recuerde modificar la propagación hacia adelante y hacia atrás. Cuál es la accurancy para este ejemplo? Mantenga la tase de aprendizaje $a = 0.01$ y el número de iteraciones iguales a $10,000$

   (b) Sustituya la función de costos por la función de entropía cruzada

   $$C = -\frac{1}{n}\sum_x [y * ln(\hat{y}) + (1-y)ln(1-a)]$$

   , conserve la función de activación sigmoid, tasa de aprendizaje $a = 0.01$ e iteraciones iguales a $10,000$. Cuál es la nueva accuracy? Escriba un par de oraciones describiendo por que pasa esto

   (c) Modifique la red neuronal base de manera que ahora tenga dos nodos de salida, uno para la probabilidad que el billete sea una falsificación y otro para la probabilidad que no lo sea. Transforme los datos de la manera necesaria para poder probar la red

   (d) Utilice 'keras' para evaluar la misma red (con dos nodos de salida), compare sus resultados con la red escrita a mano

2. **Topic Modeling**

   En este ejercicio encontraremos los temas de una base de datos de críticas de hoteles en Trip Advisor. La base de datos se llama *datos_trips.csv*. Se encuentra como un archivo comprimido en el github de la clase

   (a) Explore la base de datos. Cuántas reviews hay? Qué información tenemos de cada una de ellas?

   (b) Limpie el texto de cada una de las reviews. Recuerde que el objetivo del análisis y que datos hay determinaran este paso

   (c) Construya una nube de palabras con las 50 palabras más utilizadas después de la limpieza de los datos

   (d) Tokenize los datos y muestre una gráfica o tabla con la distribución de los token, es decir que porcentaje aparece una vez, cual dos, etc

   (e) Decida si necesitamos quitar algunos tokens. Justifiqué el proceso con un par de lineas

   (f) Corra el LDA, justifiqué su elección inicial de k. En un par de oraciones describa que hace el algoritmo de LDA y cuales son las dos matrices que calcula

(g) Construya una gráfica con las 10 palabras más probables por tema

(h) Pruebe con un par de k's distintas. Escriba lo que encuentra cuando modifica la k

(i) Escoja uno de los modelos que corrió e intente describir que representa cada uno de los temas

3. **Análisis de Componentes Principales**

Utilizaremos la base de datos USArrests, dentro de tidyverse.

(a) Explore la base de datos. Cuántas observaciones y cuántas variables tenemos?

(b) Utilice la libraría tidymodels para llevar a cabo un análisis de componentes principales, recuerde normalizar las variables

(c) Cuántos componentes principales hay?

(d) Muestre sus resultados en una gráfica que incluya el nombre de cada uno de los Estados

(e) Cómo podemos interpretar el componente principal 1 y el dos?

(f) Cómo interpretamos Estados cerca del centro de esta gráfica?

(g) Muestre en una gráfica que porcentaje de la varianza explica cada uno de los componentes

(h) Discuta los beneficios de esta técnica y si seria razonable en nuestro ejemplo usar solo los PC y no todas las variables

Fecha de entrega: Julio 3 de 2021, hasta las 11:59pm. **NO** se recibirán exámenes por fuera de esta fecha. Por favor entregar la parte práctica del examen en formato de Script de R, listo para ser ejecutado. **El examen es individual, cualquier forma de copia y fraude serán sancionadas severamente**. Enviar el examen por correo electrónico a los profesores del curso.