Universidad del Rosario, Facultad de Economía
Big Data for Public and Private Sectors
Parcial

June 27, 2017

Primera Parte: Teoría (50%)

1. Para los numerales (a) hasta (c), indique cuál(es) de i. a iv. son correctas. Justifique su respuesta.

   (a) El modelo LASSO, en relación a mínimos cuadrados, es:

      i. Más flexible y por ende garantiza precisión predictiva mejorada cuando su incremento en el sesgo es menor que su disminución en la varianza
      ii. Más flexible y por ende garantiza precisión predictiva mejorada cuando su incremento en la varianza es menor que su disminución en el sesgo
      iii. Menos flexible y por ende garantiza precisión predictiva mejorada cuando su incremento en el sesgo es menor que su disminución en la varianza
      iv. Menos flexible y por ende garantiza precisión predictiva mejorada cuando su incremento en la varianza es menor que su disminución en el sesgo

   (b) Repita (a) pero para RIDGE relativo a mínimos cuadrados

2. Supongamos que recolectamos datos de un grupo de estudiantes de una clase de estadística y medimos las variables X_1 = horas estudiadas, X_2 = Promedio de pregrado y Y = Saca 5. Estimamos una regresión logística y producimos unos coeficientes estimados $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) Estime la probabilidad de que un estudiante que estudia 40 horas y tiene un promedio de pregrado de 3.5 obtenga un 5 en esta materia.

   (b) ¿Cuántas horas necesita estudiar este estudiante para tener un chance del 50% de obtener un 5 en la materia?

3. Una empresa de bebidas contrata a una agencia de mercadeo para estudiar las preferencias de los consumidores respecto a nuevos productos. Las bebidas son descritas de acuerdo con dos características: nivel de acidez y nivel de espesor. Las dos características se miden en una escala de 1 a 10, donde 1 significa un nivel bajo y 10 un nivel alto. Los resultados del estudio hecho a 4 bebidas permiten establecer si los consumidores clasifican cada una como una bebida "rica" o "fea", y los resultados, así como las características de cada caso, los resume la tabla 1:

   La empresa duda de si introducir al mercado la bebida 5, pues desconoce el veredicto del público. Se le pide predecir la categoría a la que pertenece esta bebida, usando el algoritmo de vecinos más cercanos (kNN).

   (a) Usando el parámetro $k = 1$, ¿cuál es la predicción de cómo será clasificada la bebida? Justifique su respuesta.

   (b) ¿Cuál es la predicción si $k = 3$? Justifique su respuesta

Table 1: Estudio de Mercado

Producto	**Acidez**	**Espesor**	**Dictamen**
Bebida 1	0	3	Rica
Bebida 2	7	10	Rica
Bebida 3	0	6	Fea
Bebida 4	10	6	Fea
Bebida 5	4	6	?

Table 2: "Soledad" en Diomedes

	Soledad		
Likelihood	**Si**	**No**	**Total**
Triste	40	20	60
Alegre	10	30	40
Total	50	50	100

4. Existen diferentes métodos para medir el desempeño de un modelo, incluyendo el estadístico kappa, los estadísticos de sensibilidad y especificidad, las curvas ROC y el método de *cross-validation*.

   (a) Explique en qué consiste el estadístico kappa. ¿En qué tipo de circunstancias es preferible examinar este estadístico en lugar de la precisión total del modelo? ¿Por qué?

   (b) Explique intuitivamente qué es una curva ROC. Encuentre un ejemplo numérico que permita entender la forma en la que se construye esta curva.

   (c) Explique cómo se implementa el método de *k-fold cross-validation*. ¿Cuáles son las ventajas y desventajas de este método en relación con el *hold-out method*? En el contexto de las regresiones RIDGE y LASSO, ¿cómo se implementa el método?

5. Usted desea clasificar las canciones de Diomedes Díaz (qepd) como "tristes" o "alegres", en función de las palabras contenidas en cada una. Por simplicidad, supongamos que la clasificación se hará únicamente usando la palabra "soledad". Asumamos que al entrenar un modelo de Naive Bayes usando 100 canciones del Cacique de la Junta, usted encuentra las frecuencias reportadas por la tabla 2. Con base en dicha información, si una canción no contenida en el conjunto de entrenamiento tiene la palabra "soledad", determine la probabilidad de que sea triste e indique la categoría en la que es clasificada.

Segunda Parte: Práctica (50%)

1. En el ejercicio que sigue trabajaremos con la base de datos `Hitters.csv` que se encuentra en la sección de RIDGE y LASSO. El objetivo es predecir qué jugadores resultan ser muy ricos, con base en su salario.

   (a) Lea los datos por medio de un objeto llamado `hitters`. Remueva de este objeto la información de identificación de los peloteros. Además, omita las observaciones que tienen datos ausentes en la variable `Salary`. Redefina su base de datos de acuerdo con estos cambios.

   (b) Use las funciones `table()`, `summary()`, `hist()` y otras que considere pertinentes para entender la estructura de los datos.

   (c) Discretice la variable de resultados `Salary`. Para esto use el percentil 75 de la distribución, de tal forma que quienes estén por encima de dicho valor serán considerados una `Star`, mientras que quienes están por debajo no lo son.

   (d) Reescale las variables de la base por medio de una estandarización (Z-score) de los datos. Bautice a su nueva base de datos, tras la estandarización, `hitters_z`. Verifique, por medio de estadística descriptiva, que los datos están en la misma escala y son comparables.

(e) Divida la base `hitters_z` en dos partes: una con el 50% de las observaciones para entrenar el modelo, y el otro 50% para probar el modelo. Llame a estas dos bases `hitters_z_train` y `hitters_z_test`, respectivamente.

(f) Usando $k = \sqrt{n}$ vecinos, haga clasificación usando vecino más cercano (kNN). Presente la matriz de confusión correspondiente a la base de prueba.

(g) ¿Cuál es el nivel de precisión alcanzado por el algoritmo? ¿Cuál es el estadístico kappa? ¿Cuáles son los estadísticos de sensibilidad y especificidad? Si ud trabaja para un equipo de béisbol y quiere medir el valor de un jugador en función de si es una *Star* o no, ¿cuál de estas medidas de desempeño es más relevante y por qué?

2. Compare sus resultados del ejercicio anterior con los que se obtendría si en lugar de usar la técnica kNN, ud entrena un árbol de decisión. ¿Cuál de los dos métodos es más preciso? ¿Con cuál de los dos métodos logra un mejor desempeño en su estadístico de interés? Haga los ajustes y fije los parámetros que sean necesarios para poder estimar el árbol.

3. ¿Cómo cambia la precisión y el desempeño de su modelo si en su lugar ensambla un modelo de *bagging*?

4. ¿Qué ocurre si en su lugar estima un *random forest*? ¿Cómo cambian la precisión y el desempeño de su modelo si implementa esta metodología? ¿Con cuál se queda y por qué?

Fecha de entrega: viernes 30 de junio, hasta las 11:59PM. No se recibirán exámenes por fuera de esta fecha. Para los ejercicios prácticos, entregar un archivo `.R` con los respectivos códigos y los comentarios que respondan las preguntas del examen. Enviar a jagallegod@gmail.com