Tarea-Examen 4: Métodos de Remuestreo

Curso Avanzado de Estadística. Profa. Guillermina Eslava Gómez.

Aldo Sayeg Pasos Trejo. César Cossio Guerrero.

Posgrado en Ciencias Matemáticas. Universidad Nacional Autónoma de México.

15 de abril de 2020

1. Problema 1

Buscamos evaluar los modelos presentados en el problema 1 de la tarea-examen 3 con los métodos de Bootstrap y repeated Training/Test. Para no tener que acudir al otro reporte, la tabla 1 muestra los modelos agrupados por el método de clasificación junto con las variables que usan.

Método	Variables en el modelo	Notas extra
Análisis de Discriminante Lineal	"npreg", "glu", "bp", "skin", "bmi", "ped", "age", "type", "ped*age"	-
Naive Bayes	"npreg", "glu", "bp", "skin", "bmi", "ped", "age", "type", "ped*bp"	-
Regresión logística	"'glu", "bmi", "ped", "age", "age^2"	-
Support Vector Machines	"npreg", "glu", "bp", "skin", "bmi", "ped", "age", "type"	Kernel Polynomial de grado 5

Tabla 1: Variables de los modelos

1.1. Bootstrap

La tabla 2 muestra las tasas de error obtenidas mediante el método de Bootstrap para cada modelo. Se promedio sobre $B = 500$ iteraciones.

Model	Global	Class Yes	Class No
Analisis de Discriminante Lineal	0.1011	0.1989	0.0523
Naive Bayes	0.1131	0.1730	0.0834
Regresión Logística	0.1006	0.1910	0.0560
Support Vector Machine	0.1082	0.2377	0.0442

Tabla 2: Errores aparentes globales y locales obtenidos mediante Bootstrap

1.2. Training/Test

La tabla 3 muestra las tasas de error obtenidas mediante el método de Bootstrap para cada modelo. Ya que en el trabajo anterior con modelo se mostró que para fracciones de entrenamiento mayores al 50 % la variación del error aparente era pequeña, se tomó como fracción de entrenamiento el 75 % de la muestra y se valido con el restante 25 %. Los conjuntos de entrenamiento y prueba fueron separados manteniendo proporcionalidad de las clases. Se promedio sobre $B = 500$ iteraciones.

Model	Global	Class Yes	Class No
Analisis de Discriminante Lineal	0.2154	0.4099	0.1063
Naive Bayes	0.2339	0.3479	0.1701
Regresión Logística	0.2081	0.3839	0.1144
Support Vector Machine	0.2238	0.4876	0.0875

Tabla 3: Errores no aparentes globales y locales obtenidos mediante separación en training test

1.3. Conclusiones

Para tener una mejor idea de los resultados anterior, podemos compararlos con los resultados obtenidos previamente respecto al error aparente, que se muestran en la tabla 4

Model	Global	Class Yes	Class No
Analisis de Discriminante Lineal	0.2049	0.4124	0.1014
Naive Bayes	0.2199	0.3446	0.1577
Regresión Logística	0.2068	0.3898	0.1155
Support Vector Machine	0.2162	0.4802	0.0845

Tabla 4: Errores no aparentes globales y locales obtenidos mediante separación en training test

Es claro que para bootstrap las tasas de errores parecen ser mucho menores. Sin embargo, si se analiza la desviación estandar de la muestra de 500 tasas de errores, esta es muy alta, casi del orden de magnitud de dichas tasas, por lo que no podemos tomarlas como las más optimas. Por otro lado, las tasas obtenidas por training/test son mucho más cercanas a las tasas aparentes. En ese sentido, podemos confiar mucho más en ese método para validar la muestra.

El hecho de que todas las tasas de error hayan sido más altas en los métodos usados nos da la idea de que si hay una utilidad en evaluar a los modelos con estos métodos, pues dichas tasas de error pueden resultar más estrictas con nuestro modelo que simplemente la tasa aparente.

2. Problema 2

Evaluaremos los modelos presentados en el problema 3 de la tarea-examen 3 con los métodos de repeated Training/Test y Cross Validation. Para no tener que acudir al otro reporte, la tabla 1 muestra los modelos agrupados por el método de clasificación junto con las variables que usan.

Método	Variables en el modelo	Notas extra
Naive Bayes	"Sex", "AngPec", "AMI", "QWave", "QWavecode", "STcode", "STchange", "SuffHeartF", "Hyptertrophi", "Hyperchol", "Smoker", "Inherit", "Heartfail", "CAD", "Sex*AMI"	-
Regresión logística	"AngPec","AMI", "STcode","STchange", "Hyperchol"'	-
Support Vector Machines	"Sex", "AngPec", "AMI", "QWave", "QWavecode", "STcode", "STchange", "SuffHeartF", "Hyptertrophi", "Hyperchol", "Smoker", "Inherit", "Heartfail", "CAD"	Kernel Polynomial de grado 5

Tabla 5: Variables de los modelos

2.1. Validación cruzada

Se obtuvieron errores no aparentes mediante validación cruzada dividiendo la muestra en $k = 5$ partes iguales, cada una manteniendo la proporcionalidad original entre las clases de las observaciones. La decisión del valor de k nuevamente se tomó teniendo en cuenta que eso permitia que la fracción de entrenamiento consistiera del 80 % del conjuto original, que sabemos, por el trabajo anterior, es suficiente para entrenar al dicho modelo. Se realizaron $B = 500$ repeticiones de este proceso. La tabla 6 muestra los resultados obtenidos

Model	Global	Class Yes	Class No
Naive Bayes	0.1724	0.1909	0.1571
Regresión Logística	0.1853	0.2122	0.1628
Support Vector Machine	0.1437	0.1835	0.1106

Tabla 6: Errores no aparentes globales y locales obtenidos mediante validación cruzada para $k = 5$

2.2. Training/Test

La tabla 7 muestra las tasas de error obtenidas mediante el método de training/test‘ para cada modelo. Se tomó como fracción de entrenamiento el 75 % de la muestra y se valido con el restante 25 %. Los conjuntos de entrenamiento y prueba fueron separados manteniendo proporcionalidad de las clases. Se promedio sobre $B = 500$ iteraciones.

Model	Global	Class Yes	Class No
Naive Bayes	0.1760	0.1670	0.1537
Regresión Logística	0.1850	0.1910	0.1452
Support Vector Machine	0.1423	0.1192	0.0776

Tabla 7: Errores no aparentes globales y locales obtenidos mediante separación en training test

2.3. Conclusiones

Nuevamente, para evaluar estas tasas, podemos comparar con las tasas no aparentes obtenidas anteriormente, que se muestran en la tabla 8

Model	Global	Class Yes	Class No
Naive Bayes	0.1441	0.1495	0.1395
Regresión Logística	0.1695	0.2430	0.1085
Support Vector Machine	0.0890	0.1121	0.0698

Tabla 8: Errores no aparentes globales y locales obtenidos mediante separación en training test

En este caso, las tasas obtenidas tanto por validación cruzada como por training/test son más altas que las tasa aparentes. Esto se puede explicar debido a que el modelo ajustado y evaluado con todos los datos es más preciso debido a que tiene más observaciones sobre cuales ajustar. Si las observaciones son lo suficientemente “sencillas” para clasificarlas bien, siempre será mejor tener más datos para evaluar el modelo.

En particular, para el caso de la support vector machine, la tasa aparente global es mucho menor que las tasas obtenidas por los otros métodos. Es particularmente interesante que eso suceda

para este modelo pues es el que presenta la tasa de clasificación más baja. En ese sentido, podemos concluir que los métodos de remuestreo presentaron evaluaron peor a nuestros modelos en las tasas locales y globales. Podemos utilizarlas

3. Problema 3

Buscamos evaluar el modelo predictivo presentado en el problema 3 de la tarea-examen 3. La tabla 9 muestra las variables del modelo predictivo. Debemos recordar que el modelo utilizaba el método de regresión logística multinomial, tomando como referencia la clase "Chemical" del conjunto de datos.

Modelo	Variables	Notas
Predictivo	"InsulinResp", "Fglucose*InsulinResp" , "GlucoseInt*InsulinResp"	-

Tabla 9: Modelo predictivo

3.1. Bootstrap

La tabla 10 muestra las tasas de error por Bootstrap para $B = 500$ repeticiones

Model	Global	Class 1	Class 2	Class 3
Regresión logística	0.0029	0.0000	0.0031	0.0042

Tabla 10: Errores no aparentes globales y locales obtenidos mediante Bootstrap

3.2. Validación cruzada

Para la validación cruzada, separamos a la muestra en $k = 5$ partes iguales, usando 4 para entrenar al modelo y la restante para evaluarlo. Dicha separación mantenía la proporcionalidad en las clases de los conjuntos particionados. Se realizaron 500 iteraciones de dicho proceso para obtener las tasas de error que se muestran en la tabla 11

Model	Global	Class 1	Class 2	Class 3
Regresión logística	0.0300	0.0272	0.0295	0.0315

Tabla 11: Errores no aparentes globales y locales obtenidos mediante validación cruzada para $k = 5$

3.3. Training/Test

Para training/test, se dividió al conjunto de datos en un conjunto de entrenamiento de 75 % del tamaño original y se entrenó con el restante 25 % Los porcentajes se escogieron con el mismo argumento de que, como se mostró en la tarea 3, la tasa de no aparente presenta estabilidad a partir del 50 %. Nuevamente, esta división se hizo para siempre mantener la proporcionalidad de las clases entre los conjuntos separados .Las tasas de error que se muestran en la tabla 12

Model	Global	Class 1	Class 2	Class 3
Regresión logística	0.0282	0.0076	0.0167	0.0147

Tabla 12: Errores no aparentes globales y locales obtenidos mediante training/test

3.4. Conclusiones

Podemos realizar la comparación con las tasas aparentes obtenidas anteriormente, que se muestran en la tabla 13

Model	Global	Class 1	Class 2	Class 3
Regresión logística	0.0138	0.0000	0.0278	0.0132

Tabla 13: Errores no aparentes globales y locales obtenidos mediante training/test

Solo en el caso de Bootstrap, el error global aparece menor que el error global aparente. En todos los demás, los errores aparecen mayores a los aparentes. Esto quiere indicar que quizá Bootstrap no sea el mejor método para evaluar a nuestro modelo. Los otros métodos presentan lo esperado: un alza en las tasas de error, por lo que si podríamos utilizarlos para evaluar con mayor rigurosidad a nuestro modelo.

Ya que el modelo presentado nos interesa solamente en términos predictivos, podríamos utilizar la validación cruzada o el training/test como evaluación del modelo en un algoritmo que seleccione modelos óptimos.