Avance 2: Entrenamiento, adecuación y evaluación de modelos

Estudiantes

Santiago Escobar Casas 1214746431
Santiago Otálvaro Ospina 1020492578

Tutor

Javier Fernando Botia Valderrama

2508205 - Modelos de sistemas

Universidad de Antioquia
Ingeniería de Sistemas
2020-1

Introducción

Para este avance se tomarán los diferentes datasets obtenidos en el avance 1 y se procederá a buscar el mejor o los mejores modelos de clasificación para predecir la posibilidad de intento de suicidio recurrente.

En un principio, se hizo uso de la herramienta Classification Learner de Matlab para hacer pruebas a diferentes modelos, debido a su rapidez para probar múltiples modelos a la vez.

Se tomarán los pares dataset-modelo que den mejores resultados, preferiblemente por encima del 70% de exactitud, para seguir desarrollando la parte de selección de parámetros y optimización de hiper parámetros.

Para la selección de estos elementos (parámetros y ajustes de hiper parámetros) se usarán las herramientas interactivas o automáticas que provee Matlab.

Como métodos de validación y calificación de los modelos se pretenden usar: Exactitud, Matriz de confusión y ROC curve.

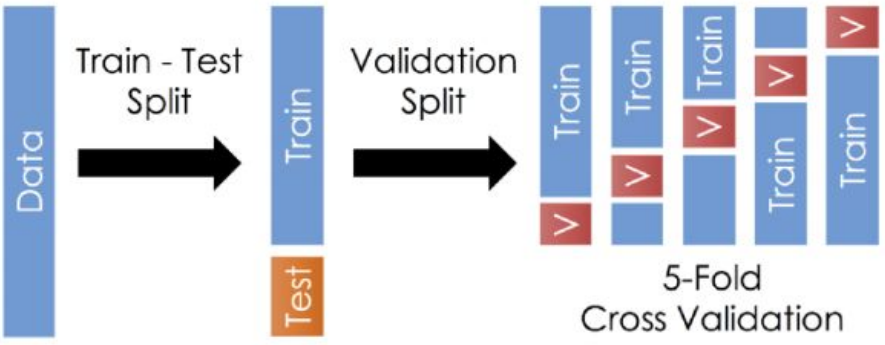
Datasets de entrada

Previamente en el avance 1, se obtuvieron 4 data sets después de un proceso de limpieza y que ahora se usarán como datos de entrada. Estos eran:

- cds_imputed : dataset con 33 características y 4146 registros
- cds : dataset con 28 características 4146 registros
- cds_few : dataset 33 características y 655 registros
- cds_fem_minus_alcohol: dataset 32 caracteristicas 1690 registros.

Posterior a un proceso de entrenamiento exploratorio, se continuará con los modelos más prometedores, es decir, en un primer momento se tomará en cuenta la exactitud de los modelos.

Es de suma importancia tener en cuenta que para el entrenamiento de los modelos fue usada una validación cruzada con "kfolds" (k=5), así, el valor de la exactitud presentado corresponde a la exactitud de validación y esta sirve como un estimado del desempeño del modelo en nuevos datos, comparados con el conjunto de entrenamiento.



Resultados resumidos de los Dataset

Por cada dataset se presentan los 3 modelos con mejores resultados (el mejor en amarillo):

Dataset	Categoría	Exactitud
cds	Logistic Regression	67.4%
cds	Linear Discriminant	66.9%
cds	SVM (Coarse Gaussian)	66.8%
cds_imputed	Linear Discriminant	67.9%
cds_imputed	Logistic Regression	67.8%
cds_imputed	Ensemble(Subspace Discriminant)	67.7%
cds_few	Tree (Coarse)	64.4%
cds_few	SVM(Medium Gaussian)	63.7%
cds_few	Ensamble(Subspace Discriminant)	63.4%
cds_few_minus_alcohol	Linear Discriminant(Covariance Structure)	60.2%
cds_few_minus_alcohol	Ensamble (Boosted Tres)	59.2%
cds_few_minus_alcohol	Logistic Regression	59.0%

Como se puede observar, ningún par dataset-modelo obtuvo una precisión mayor al 70% tal y como se había definido inicialmente para su aceptación. Por este motivo, se tomará aquel dataset que produjo el modelo con la mayor precisión (cds_imputed) y los mejores modelos obtenidos a partir de este.

Feature selection

En la búsqueda de reducir la dimensionalidad se pretende realizar un proceso de selección de características usando el algoritmo MRMR (Minimum Redundancy Maximum Relevance), del cual se puede obtener el "ranking" de importancia de los predictores teniendo en cuenta la respuesta.

Posteriormente se entrenan 2 modelos, uno con todas las características y, adicionalmente, otro con el conjunto de las 7 más importantes.

```
idx = fscmrmr(cds_imputed,'inten_prev');
most_signif_features = cds_imputed.Properties.VariableNames(idx(1:7)).'
```

```
most_signif_features = 7×1 cell
'antec_tran'
'hist_famil'
'muerte_fam'
'antec_v_a'
'prob_consu'
'plan_suici'
'gp_psiquia'
```

```
less_signif_features =cds_imputed.Properties.VariableNames(idx(end-4:end)).'
```

```
less_signif_features = 5×1 cell
'escolarid'
'esco_educ'
'tipo_ss_C'
'trab_socia'
'sexo_'
```

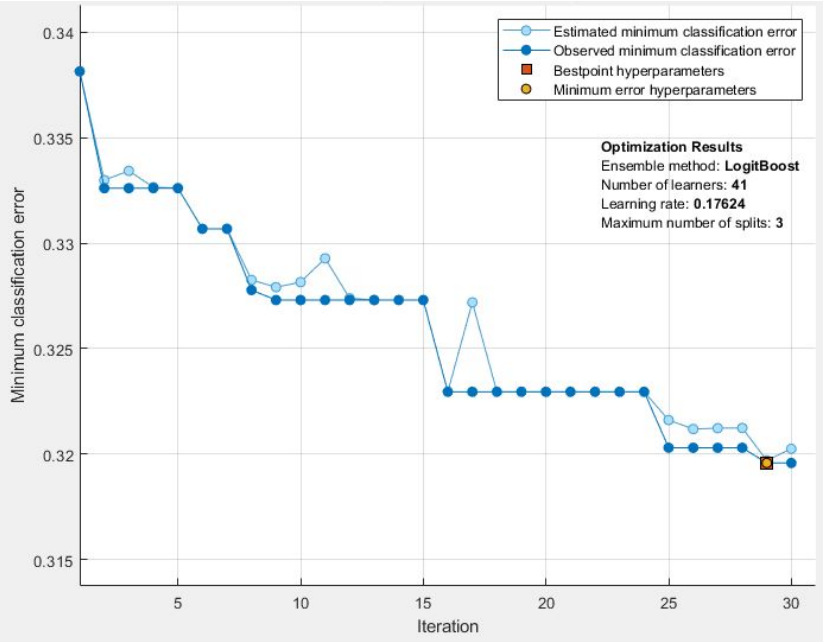
Optimización de hiper parámetros

Se utilizarán 2 enfoques: para los modelos simples (árboles de decisión) se realizará mediante GridSearch, mientras que para los más complejos (Ensamble, SVM) será utilizado un método de optimización bayesiana, el cual, a través de 30 iteraciones se va redirigiendo hacia aquellos hiper parámetros del espacio de búsqueda que proveen mejores resultados para el modelo. Este enfoque se toma debido a los costos computacionales y temporales elevados de realizar Grid o Random Search en modelos complejos.

Como parte del proceso de optimización se obtiene el gráfico de error de clasificación mínimo en el cual se encuentran principalmente:

- Los resultados de la optimización
- El mínimo error de clasificación observado (puntos azules) hasta la iteración actual
- Bestpoint hyperparameters (cuadrado rojo), indica la iteración que corresponde a los valores de los hiper parámetros optimizados.

A continuación, se presenta a modo de ejemplo uno de dichos gráficos obtenidos, correspondiente al proceso de optimización de un modelo Ensemble.



Comparación de modelos

Para este punto se tendrán en cuenta varias cosas:

- La exactitud del modelo: mientras mayor mejor, sin llegar a un caso de sobreajuste.
- El número de parámetros: en general es de interés obtener modelos que con un bajo número de parámetros sean capaces de cumplir con su objetivo a cabalidad, esto debido a que en un caso real es más difícil y costoso, en términos de dinero y tiempo, obtener una cantidad

grande de información. En este caso no se les dará mayor importancia a unos parámetros sobre otros, sólo será de interés el número de ellos.

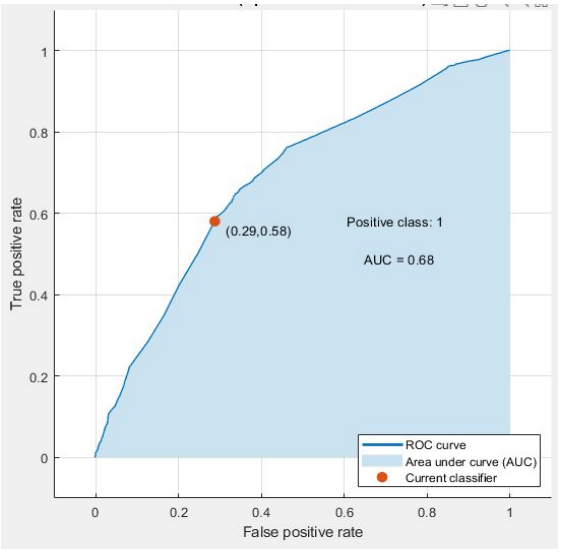
- La complejidad del modelo: se preferirán modelos más simples(e.g. árboles de decisión)
- Un último factor que se tendrá en cuenta para preferir un modelo sobre otro es la distribución de falsos negativos hacia cierta clase particular, por ejemplo, en este contexto no sería nada bueno identificar erróneamente a aquellas personas con tendencia repetitiva al intento de suicidio, mientras que identificar erróneamente a aquellos que en realidad no (falso positivo), sería más aceptable.

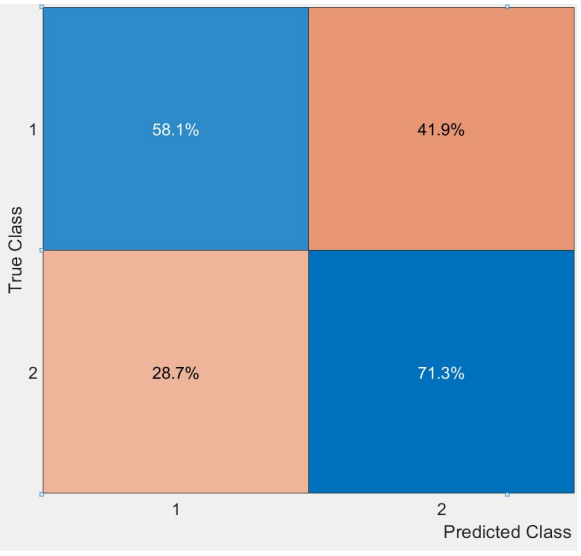
En cuestiones de exactitud se obtuvieron resultados que variaron entre diferentes entrenamientos de los modelos, pero en general se "destacaron" Ensemble y Logistic Regression con una exactitud alrededor de 68%.

En cuanto al análisis **ROC**, se tiene resultados poco satisfactorios, lo que quiere decir que la distinción entre clases no es óptima. En términos del valor AUC los mejores pertenecen a Discriminant, Logistic Regression y SVM.

Teniendo en cuenta el último factor que se mencionó para la comparación de los modelos, lo que se busca es maximizar la predicción correcta de la etiqueta 1 {si hay intentos previos de suicidio}, en este sentido la mayoría de los modelos son deficientes. Pero en general los de 7 características se comportan mejor que los completos. Así, uno de los mejores sería el de discriminante lineal con 7 características. ya que es el que más se acerca a lo requerido,pues tiene poco número de características y no hace parte de los modelos demasiado complejos.

Con fines ilustrativos, se presenta a continuación la curva ROC y la matriz del modelo "Linear Discriminant" entrenado solo con las 7 características más importantes.





Conclusiones

Como se puede observar, los resultados no juegan a nuestro favor. Esto se ve reflejado en la precisión de los modelos y en los árboles de decisión. Dicha situación puede ser debida a la gran cantidad de datos categóricos que poseemos, ya que, junto a una carencia de información, no nos entregan los resultados adecuados. Ejemplo de esto es, una de las características que mejor información debería brindarnos es la reincidencia, pero a la hora de revisar el parámetro, no es autosuficiente esta categoría debido a algunos factores, como el estado de vida actual de los individuos (si están vivos o muertos).

En cuanto a las características más relevantes que destacó Feature selection son bastante apropiadas, esto podría significar que se halló una mayor concurrencia o un patrón específico de estas señales en el recorrido de los datasets.

Sería de gran ayuda poder obtener un dataset con datos relacionados sobre los intentos previos de suicidio de las personas y que la información que esta contenga sea completa, al menos lo suficiente para poder alcanzar los objetivos que se propusieron desde el principio.

También se espera, ya sea en un futuro avance o en la versión final del proyecto, incluir predicciones no solo probabilísticas sino también determinísticas, además de una comparativa de los modelos no solo usando datos de validación sino también apartar datos para test y determinar resultados a partir de eso.

Se puede acceder al procedimiento completo desde el repositorio.

Referencias

[1] "Feature Selection and Feature Transformation Using Classification Learner App- MATLAB & Simulink", Mathworks.com, 2020. [Online]. Available: https://www.mathworks.com/help/stats/feature-selection-and-feature-transformation.html.

[2] "Train Classification Models in Classification Learner App- MATLAB & Simulink", Mathworks.com, 2020. [Online]. Available: https://www.mathworks.com/help/stats/train-classification-models-in-classification-learner-app.html.

[3] "Intuitive Hyperparameter Optimization : Grid Search, Random Search and Bayesian Search!", Medium, 2020. [Online]. Available: https://towardsdatascience.com/intuitive-hyperparameter-optimization-grid-search-random-search-and-bayesian-search-2102dbfaf5b.

[4] "Automated Machine Learning Hyperparameter Tuning in Python", Medium, 2020. [Online]. Available: https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a.

[5] "Workflow of a Machine Learning Project", Medium, 2020. [Online]. Available: https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94.

[6]"Introduction to Feature Selection- MATLAB & Simulink", Mathworks.com, 2020. [Online]. Available: https://www.mathworks.com/help/stats/feature-selection.html.

[7]"Understanding Confusion Matrix", Medium, 2020. [Online]. Available: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.

[8]"Understanding AUC - ROC Curve", Medium, 2020. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.