Informe funciones discriminantes

Diego Hernández Jiménez

1. Introducción

Los accidentes cerebrovasculares o ictus constituyen la segunda causa de muerte en la población general española, siendo además la primera causa de muerte en mujeres, (Consejo Interterritorial del Sistema Nacional de Salud, 2009). En otros países desarrollados la situación es similar (en Estados Unidos, por ejemplo, es la quinta causa de muerte, Center of Disease Control and Prevention, 2018). Debido al enorme impacto que tienen a nivel social, resulta de gran interés investigar acerca de sus determinantes, factores de riesgo, o simplemente predictores. Para este propósito resulta indispensable el modelado estadístico, y son especialmente útiles las técnicas de aprendizaje automático o *machine learning*.

En este estudio se adopta un enfoque basado en aprendizaje automático y se fija como objetivo la clasificación precisa de una muestra de sujetos en dos clases, "individuos que han sufrido ictus" (*Yes*) e "individuos que no han sufrido ictus" (*No*). Para lograrlo se propone un modelo de funciones discriminantes.

La muestra empleada para construir y evaluar el modelo fue publicada por McKinsey & Company y contiene 43.000 observaciones. Cada caso tiene puntuaciones en 10 atributos distintos. Por requisitos del análisis discriminante, las variables incialmente consideradas son aquellas de tipo cuantitativo, a saber, edad (*age*), índice de masa corporal (*bmi*) y nivel promedio de glucosa en sangre (*avg_glucose_level*)

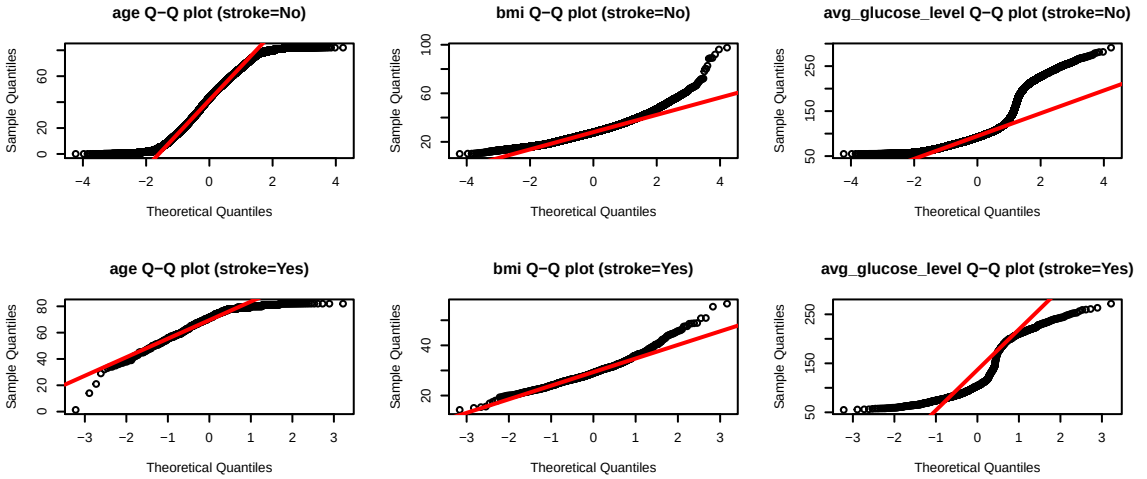
Un análisis exploratorio revela que no existen casos perdidos en ninguna variable salvo en *bmi*, que tiene 1462. Representan un 3.4% de la muestra, pero no serán imputados, sino que se eliminarán en posteriores análisis. Otro aspecto a destacar es el desequilibrio en el número de casos por clase: 98.2% pertenece a la clase *No*, mientras que solo hay 783 casos en la clase *Yes*. Por último, a partir de la inspección visual de la matriz de correlaciones y de diagramas de puntos puede observarse que los predictores no son completamente independientes, aunque el nivel de correlación no parece preocupante. No obstante, la correlación de éstos con la variable criterio (codificada como 0=No, 1=Yes) tampoco es muy elevada y hay bastante solapamiento entre las variables, lo que anticipa que quizá un modelo basado en funciones lineales no clasifique con demasiado éxito.

Attributes	age	avg_glucose_level	bmi	stroke
age	1	0.238	0.359	0.156
avg_glucose_level	0.238	1	0.191	0.079
bmi	0.359	0.191	1	0.020
stroke	0.156	0.079	0.020	1

2. Métodos y herramientas

El análisis discriminante se asienta sobre dos supuestos fundamentales. La distribución conjunta de los predictores es normal multivariante para cada grupo k, $N \sim (\mu_k, \mathbf{\Sigma})$ siendo la matriz de covarianzas $\mathbf{\Sigma}$ igual

para las k clases (James et. al, 2013). El supuesto de normalidad se ha evaluado mediante la inspección visual de gráficos Q-Q, generados en R, para los diferentes predictores, basándose en el hecho de que "si alguna de las variables originales no se distribuye como una normal, entonces es seguro que todas las variables conjuntamente no se distribuirán como una normal multivariante" (de la Fuente, S., 2011).



Se aprecia una clara desviación de la normalidad de la variable *avg_glucose_level*, siendo aparentemente menos relevantes en los otros dos casos.

Por otro lado, el supuesto de homogeneidad de covarianzas se ha contrastado en SPSS mediante la prueba M de Box, arrojando la transformación en un estadístico F un valor de 125.82 ($p < 0.00$). No es en este contexto, sin embargo, la prueba más útil, pues el test es sensible a las desviaciones de la normalidad multivariante y al tamaño de la muestra, mostrándose más liberal con un elevado tamaño de ésta, como ocurre aquí. No obstante, ante esta situación resultaría más conveniente emplear un análisis cuadrático discriminante, pues no requiere el cumplimiento de este último supuesto.

En una siguiente etapa se procede a la selección de variables a incorporar en el modelo. Aunque existen criterios estadísticos, aquí el proceso está guiado más teóricamente. En primer lugar se ajusta un modelo con un único predictor, *age*. A continuación se introducen *avg_glucose_level* y *bmi*, que se mantendrán o no en el modelo final en función de la contribución que hagan a la reducción de los errores de clasificación.

Para el ajuste de modelos se emplea Rapidminer y Matlab y se sigue la siguiente estrategia:

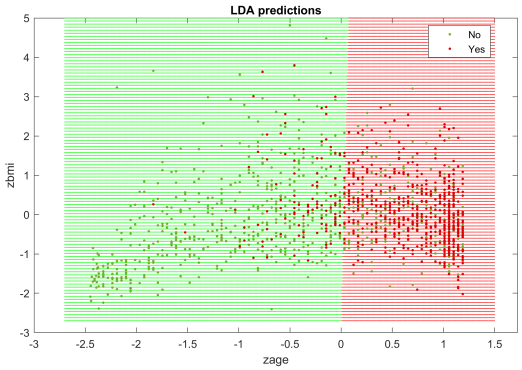
1. Debido a que se asume normalidad en la variables predictoras, se realiza una estandarización, transformándolas en puntuaciones típicas. A continuación, para solventar el problema del desequilibrio de las grupos (que produce un clasificador "trivial" que clasifica todos los casos como pertenecientes a la clase *No*), se filtra a los individuos de la clase *No* y se extrae una muestra aleatoria simple con tamaño igual al número de casos *Yes* (783). Entonces se une esta submuestra a la formada por las observaciones *Yes*. Esta nueva muestra está ahora equilibrada y las probabilidades a priori de cada clase es 0.5. Se divide entonces la muestra en un conjunto de datos de entrenamiento (70% del total) y otro de validación. Los modelos propuestos se ajustan en el primero y se evalúa su rendimiento en el segundo. La comparación entre modelos se se lleva a cabo mediante la prueba de McNemar, puesto

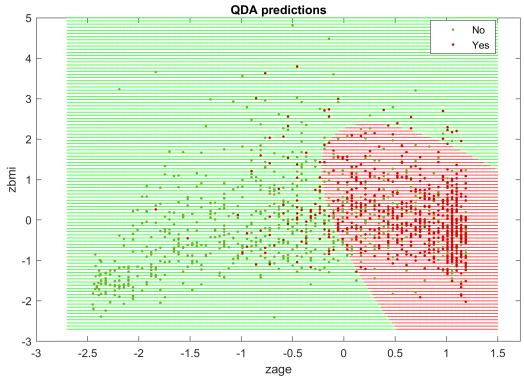
que al utilizar todo los modelos la misma muestra de validación, las proporciones de aciertos están relacionadas, no son independientes (Dietterich, 1998; Raschka, 2018). El procedmiento completo se repite con 100 muestras distintas. El objetivo es minimizar la posibilidad de sesgos de selección en la fase de "equilibrado", dado que se utiliza tan solo una pequeña fracción de la muestra total de sujetos de la clase *No* (concretamente se seleccionan 783 de 42.617).

```
% stroke -> dataset
% standarize predictors
predictors=["age","avg_glucose_level","bmi"];
stroke(:,predictors)=normalize(stroke(:,predictors),"zscore");

pvalslda=zeros([100 2]); % p-values for the different model comparisons
accs12=zeros([100 2]); % accuracies for model 1 and model 2
accs13=zeros([100 2]); % accuracies for model 1 and model 3
rng(13) % random seed

for i = 1:100

    no_stroke=stroke(stroke{:,'stroke'}==0,:); % filter class 'No'
    ids=randsample(no_stroke{:,'Var1'},783); % extract 783 random ids

    samp=[stroke(ids,:) ; stroke(stroke{:,'stroke'}==1,:)]; % join the 783+783 cases

    zage=samp{:,'age'};
    zbmi=samp{:,'bmi'};
    zglucose=samp{:,'avg_glucose_level'};
    group=samp{:,'stroke'}; % target

    % split dataset in train set and test set
    c=cvpartition(length(group),'HoldOut',.3);
    ids_train=training(c);
    ids_test=test(c);

    % age as predictor LDA
    lda1=fitcdiscr(zage(ids_train),group(ids_train), ...
        "DiscrimType","linear");

    % age+glucose as predictors LDA
    lda2=fitcdiscr([zage(ids_train) zglucose(ids_train)],group(ids_train), ...
        "DiscrimType","linear");

    % age+bmi as predictors LDA
    lda3=fitcdiscr([zage(ids_train) zbmi(ids_train)],group(ids_train), ...
        "DiscrimType","linear");

    % predictions LDA..............
    preds1_test=lda1.predict(zage(ids_test));
    preds2_test=lda2.predict([zage(ids_test) zglucose(ids_test)]);
    preds3_test=lda3.predict([zage(ids_test) zbmi(ids_test)]);

    % comparisons.....................
    % lda1 correct and lda2 incorrect
    b=nnz(preds1_test==group(ids_test) & preds2_test~=group(ids_test));
    % lda1 incorrect and lda2 correct
```

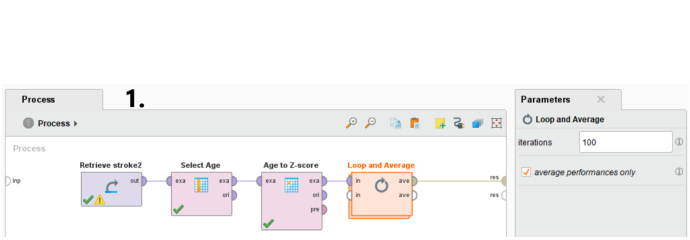
```
c=nnz(preds1_test~=group(ids_test) & preds2_test==group(ids_test));

mcnemar= (abs(b-c)-1)^2 / (b+c);
pval=chi2cdf(mcnemar,1,'upper');
pvalslda(i,1)=pval; % update matrix of LDA p-values for model 1 vs model 2
accs12(i,:)=[mean(preds1_test==group(ids_test)) mean(preds2_test==group(ids_test))];

% lda1 correct and lda3 incorrect
b=nnz(preds1_test==group(ids_test) & preds3_test~=group(ids_test));
% lda1 incorrect and lda3 correct
c=nnz(preds1_test~=group(ids_test) & preds3_test==group(ids_test));

mcnemar= (abs(b-c)-1)^2 / (b+c);
pval=chi2cdf(mcnemar,1,'upper');
pvalslda(i,2)=pval; % update matrix of LDA p-values for model 1 vs model 3
accs13(i,:)=[mean(preds1_test==group(ids_test)) mean(preds3_test==group(ids_test))];
```

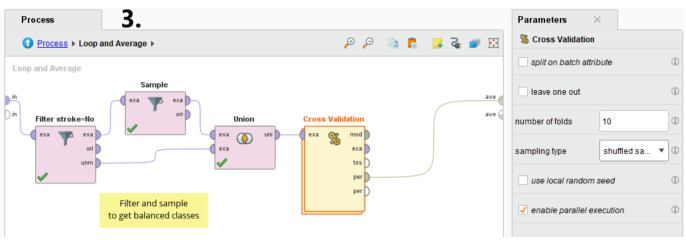
Esta fase se repite con modelos discriminantes cuadráticos, por razones teóricas y porque aparentemente parecen arrojar predicciones distintas, como puede apreciarse por la frontera de decisión que se genera en el caso del modelo con *age* y *bmi* como predictores.

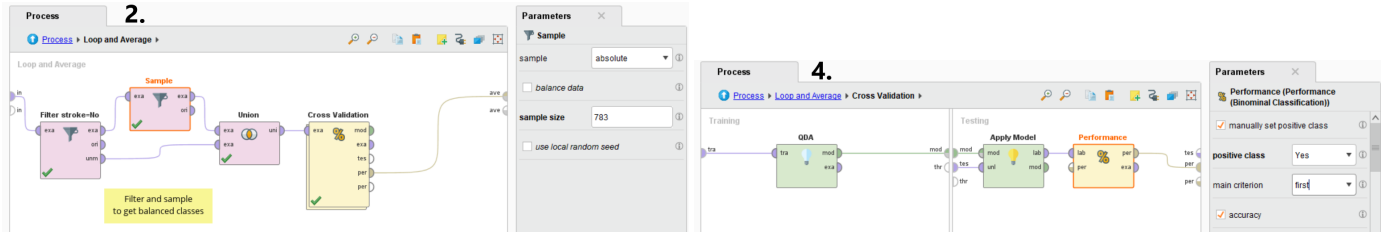




2. Una vez seleccionadas las variables y el tipo de modelo más adecuado, se vuelve a reestimar, pero en esta ocasión mediante validación cruzada de 10 iteraciones. Este procedimiento se repite 100 veces, cada vez con una submuestra de de casos *No* distinta, para evitar el problema mencionado anteriormente. Además, de manera adicional, y por el tipo de problema de predicción que se plantea, se introduce una penalización en el modelo para ponderar como más relevantes los falsos negativos. Las medidas del rendimiento finales son el promedio de las 100 iteraciones. Todo el procedimiento se lleva a cabo con Rapidminer y en Matlab.







3. Resultados y discusión

En el paso 1. se han comparado diversos modelos. Se han obtenido 100 niveles críticos de la prueba de McNemar. El porcentaje de valores p menores a 0.05 se ha empleado como criterio para juzgar la significación. Se encuentra que modelo con *age* como predictor no difiere de manera estadísticamente significativa de los modelos que incluyen además *avg_glucose_level*. Aunque en un 9% de los contrastes se obtiene $p < 0.05$, se decide tomar las diferencias como no significativas, ya que la prueba de McNemar, basada en la aproximación a χ^2, resulta algo más liberal cuando las casillas (n_{12} y n_{21}) de la tabla de contingencia contienen pocos casos (Raschka, 2018), como ocurre aquí (alrededor de 10 en todas las iteraciones). Sí se considera significativa la diferencia entre el modelo con *age* como predictor y el modelo con *age* y *bmi* como predictores. En un 50% de las iteraciones se hallan niveles críticos menores a 0.05. Se observa que la ventaja en proporción de aciertos es para el modelo más simple, por lo que se descarta *bmi* como predictor. Se produce la misma situación con los modelos discriminantes cuadráticos. Finalmente, tampoco existen diferencias significativas entre el modelo de un predictor basado en análisis discriminante lineal del basado en análisis discriminante cuadrático (solo en un 2% de las iteraciones se hallan diferencias estadísticamente significativas). Por ese motivo, se selecciona finalmente el modelo discriminante cuadrático con *age* como único predictor, por ser el más flexible.

En el paso 2. se evalúa la capacidad predictiva del modelo. El código de matlab es esencialmente igual que el mostrado antes, pero con la diferencia de que los modelos a ajustar son:

```
% age as predictor with and without cost matrix
qda=fitcdiscr(zage,group,'DiscrimType','quadratic','CrossVal','on');
qda_cost=fitcdiscr(zage,group,'DiscrimType','quadratic','CrossVal','on','Cost',[0 1;2 0]);
```

La matriz de confusión (resultado de sumar los valores de 100 matrices de confusión, una por iteración) del modelo sin costes asociados es:

	Predicción=No	Predicción=Yes
Verdadero=No	51719	26581
Verdadero=Yes	10427	67873

La matriz de confusión del modelo con la matriz de costes $\begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$ asociada es:

	Predicción=No	Predicción=Yes
Verdadero=No	44983	33317
Verdadero=Yes	6007	72293

La precisión media alcanzada por el primer modelo es del 76.37%. Estos resultados son similares a los obtenidos en Rapidminer (76.7%), con una dispersión de ±0.58 desviaciones típicas. Este resultado parece consistente con lo que han hecho otros autores con la misma base de datos (Nwosu et. al, 2019), si bien ellos

no utilizaron funciones discriminantes.

Por otra parte, la precisión del modelo con costes es del 74.89%, lo que supone una reducción del 1.48%. Resulta difícil llevar a cabo un contraste de significación en estas circunstancias (violación del supuesto de independencia), pero se puede juzgar la discrepancia como poco relevante si tenemos en cuenta que la sensibilidad aumenta del 86.68% al 92.33%. En este contexto, en el que resulta primordial pronosticar adecuadamente aquellos sujetos que potencialmente sufrirán ictus, la reducción en la precisión global puede estar justificada si conlleva aumento de aciertos en la detección de verdaderos positivos. Además, si comparamos los valores F_1 de cada modelo, es decir, si tomamos en cuenta una medida que contempla tantos los falsos positivos como los falsos negativos, las diferencias desaparecen. Para el modelo sin costes la F_1 asociada es 0.7857, mientras que para el modelo con costes es 0.7861.

4. Conclusiones

El modelo propuesto se ha generado a partir de muestras equilibradas con la mitad de casos pertenecientes a la categoría *No* y la otra mitad perteneciente a *Yes*, con lo que podemos decir que el modelo mejora en aproximadamente un 26% las predicciones que se harían simplemente por azar. Está lejos de poder considerarse un resultado espectacular, además de que la variable predictora es la edad, lo cual no resulta soprendente ni especialmente útil, pues no es un factor de riesgo modificable. Ello pone de manifiesto la necesidad de incoporar nuevas variables cuantitativas en futuros modelos si se pretende emplear el análisis discriminante. El hecho de que ni el índice de masa corporal ni el nivel de glucosa consigan reducir los errores de clasificación más de lo que lo hace la edad, (introduciendo ruido, incluso, en el caso del índice masa corporal), no implica necesariamente que no sean factores de riesgo relevantes. Es probable que las variables deban ser transformadas mediante *feature engineering* para extraer información más útil. En el caso del nivel de glucosa, por ejemplo, podría resultar interesante dicotomizar o politomizar la variable, creando categorías referidas al riesgo de sufrir diabetes, por ejemplo.

Finalmente es necesario destacar que, al trabajar con clases altamente desequilibradas, se ha optado por llevar a cabo la estrategia de submuestreo (*undersampling*), pero puede no ser la mejor solución al problema. Se ignora la verdadera distribución de las clases en la realidad, pues el modelo se entrena y se valida en conjuntos de datos equilibrados. El modelo así estimado probablemente ofrezca peores estimaciones en una muestra con una proporción de casos por clase como la que se encuentra en el conjunto de datos original.

5. Referencias bibliográficas

Center of Disease Control and Prevention (2018). *Stroke facts*. Recuperado de https://www.cdc.gov/stroke/facts.htm

Consejo Interterritorial del Sistema Nacional de Salud (2009). *Estrategia en Ictus del Sistema Nacional de Salud*. Recuperado de https://www.mscbs.gob.es/organizacion/sns/planCalidadSNS/docs/EstrategiaIctusSNS.pdf

De la Fuente, S. (2011). Análisis discriminante. Recuperado de https://estadistica.net/Master-Econometria/Analisis_Discriminante.pdf

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7), 1895-1923.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). Classification. En G. James, D. Witten, T. Hastie y R. Tibshirani, *An introduction to statistical learning* (pp. 138-150). New York: springer.

Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., y John, D. (2019, July). Predicting stroke from electronic health records. En 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5704-5707). IEEE.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.