Selección de modelos

Marta Rodríguez Sampayo
Minería de datos
Universidad Nacional de Educación a Distancia
mrodrigue8255@alumno.uned.es

22 de mayo de 2020

1. Marco teórico

A la hora de desarrollar una solución de *machine learning* para un problema concreto son varios los factores a tener en cuenta. En primer lugar, el tipo de problema (clasificación o regresión) restringe el número de modelos entre los que seleccionar uno, pero determinar el que mejor se adapta a partir de este conjunto no es en absoluto trivial. A mayores, una vez escogido el modelo que se considera mejor para el problema dado, este puede depender de uno o varios parámetros que alteren su comportamiento y, por tanto, el rendimiento del sistema.

Este problema es conocido como **selección de modelo** y puede aplicarse a diferentes niveles de abstracción. Por un lado, la selección de los hiperparámetros[1] del método escogido resulta imprescindible para garantizar un buen rendimiento de este sobre el conjunto de datos de entrada. Por otra parte, es deseable seleccionar el algoritmo con mejores resultados, en cuanto a eficacia y eficiencia, en la tarea a realizar. Además se debe tener en cuenta la diferencia entre selección y evaluación de modelos. Este último enfoque trata de estimar el error de generalización o predicción del modelo escogido sobre nuevos datos (no empleados durante el entrenamiento).

La necesidad de utilizar datos para la evaluación que no han sido vistos anteriormente por el modelo radica en el fenómeno de *overfitting* o sobreajuste. Este efecto se produce cuando un modelo es entrenado sobre un conjunto de datos y se tiene en cuenta el error de entrenamiento producido únicamente sobre esos datos. Este error no es una buena estimación del error de generalización, ya que obtiene un resultado demasiado optimista, puesto que el modelo se ha ajustado a los datos de entrenamiento. Por tanto, se puede dividir el proceso de selección de modelo en tres etapas diferenciadas: ajuste del modelo, estimación del error de predicción(selección) y evaluación del error de generalización.

[1] Parámetros del método de aprendizaje que se deben especificar con anterioridad. Por ejemplo, antes de comenzar el entrenamiento. No se deben confundir con los parámetros del modelo que resultan del ajuste del mismo.

Dos conceptos con un papel fundamental en la selección de modelos son el **sesgo** y la **varianza**. Por un lado, se puede definir el sesgo como la cantidad de desviación entre la estimación realizada por el modelo y el valor real. Es decir, un modelo con un sesgo alto no se ajustará correctamente a la estructura de los datos. Por otro, la varianza mide cuánto varía el modelo a medida que se modifica el conjunto de datos de entrenamiento. De forma que un modelo con una varianza alta será muy sensible a pequeñas variaciones de este conjunto, pudiendo provocar un sobreajuste del modelo. Por tanto, a medida que aumenta la complejidad del modelo, disminuye el sesgo y aumenta la varianza. Es necesario encontrar un equilibrio entre sesgo y varianza.

En general, la estrategia a seguir para llevar a cabo la selección de un método de resolución para un problema dado dependerá de la cantidad de datos de los que se disponga. En caso de que el conjunto de datos sea de gran tamaño, es recomendable dividirlo en tres partes: entrenamiento, validación y pruebas. De esta forma, el primer subconjunto se utilizará para ajustar todos los modelos candidatos (con sus correspondientes combinaciones de hiperparámetros de ser el caso) y el segundo para evaluarlos y seleccionar aquel que haya obtenido mejor rendimiento. Finalmente, se estima el error de generalización de este modelo (usualmente habiéndolo vuelto a entrenar sobre el conjunto de la unión de entrenamiento y validación) empleando el conjunto de pruebas. Una división habitual es 50 %/25 %/25 %.

Sin embargo, la situación habitual es no disponer de un conjunto de datos suficientemente amplio para poder realizar una división como la que se explica en el párrafo anterior. En este caso, el enfoque más común consiste en formar dos particiones (entrenamiento y pruebas) y aproximar el paso de validación. Esta aproximación se puede realizar de forma analítica, con métodos como AIC[1], BIC[2], MDL[3] o SRM[4], o bien reutilizando de forma eficiente las muestras, empleando validación cruzada o *bootstraping* [5].

Además de las técnicas de evaluación de modelos comentadas, en [6] se describen una serie de enfoques de prueba de hipótesis estadísticas aplicados a la comparación de modelos y algoritmos[2] de *machine learning*.

1.1. Reutilización eficiente de muestras

Los métodos comentados en este apartado estiman directamente el error medio de generalización al aplicar el modelo a una muestra de prueba independiente a los datos de entrenamiento. Los enfoques principales se describen en los siguientes subapartados.

Validación cruzada de K iteraciones (*K-Fold Cross-Validation*) Consiste en: (1) dividir en K subconjuntos los datos de entrenamiento, (2) uno de estos subconjuntos se emplea como datos de prueba y los $K-1$ restantes como datos de entrenamiento, (3) se repite este proceso durante K iteraciones, con cada uno de los subconjuntos posibles y (4) se calcula la media aritmética de los resultados de cada iteración. Habitualmente el valor de K empleado es 5 o 10.

Validación cruzada dejando uno fuera (*Leave-one-out cross-validation* - LOOCV) La división de los datos se realiza de forma que en cada iteración se emplea una única muestra como datos de prueba y el resto como entrenamiento.

[2]Comparar conjuntos de modelos donde cada conjunto ha sido ajustado a diferentes conjuntos de datos de entrenamiento.

Bootstrap La idea básica consiste en crear conjuntos de datos aleatoriamente a partir del conjunto de entrenamiento, cada muestra del mismo tamaño que el conjunto original. Posteriormente se reajusta el modelo con cada conjunto *bootstrap* y se examina el comportamiento de los ajustes sobre las B réplicas.

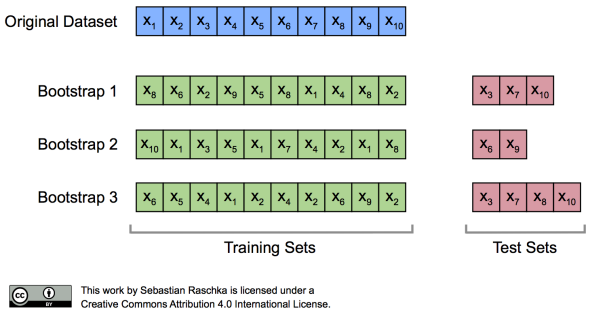

Figura 1: Bootstrap

1.2. Test estadísticos

En general, los tests estadísticos se basan en el planteamiento de una hipótesis nula (H_0) y el rechazo o aceptación de la misma en base al cálculo de un valor estadístico. A partir de la distribución de este estadístico se calcula un valor p correspondiente a la probabilidad de observar un valor de dicho estadístico igual o mayor al calculado empíricamente. Si este valor p es menor que el nivel de significancia α establecido previamente, se rechaza la hipótesis nula. A continuación se describen brevemente algunos de estos métodos estadísticos.

Test Q de Cochran Este test se puede aplicar para comparar el rendimiento de más de dos clasificadores. La definición de H_0 es que no existe diferencia entre las precisiones de clasificación de los distintos modelos ($\{C_1, ..., C_M\}$) probados sobre el mismo *dataset*. El estadístico empleado es $Q = (M-1)\frac{M\sum_{i=1}^{M} G_i^2 - T^2}{MT - \sum_{j=1}^{n} M_j^2}$ donde G_i es el número de muestras clasificadas correctamente por C_i, M_j es el n úmero de clasificadores que clasificaron correctamente la muestra j del conjunto de pruebas y T es el número total de votos entre los M clasificadores.

Test F En este caso, la hipótesis nula es de nuevo que no existen diferencias entre la precisión de los clasificadores $\{C_1, ..., C_M\}$ probados sobre el mismo conjunto de datos. El valor de F se calcula siguiendo la fórmula $F = \frac{MSA}{MSAB}$ donde $MSA = \frac{SSA}{M-1}$ y $MSAB = \frac{SSAB}{(M-1)(n-1}$ siendo n el número de muestras en el conjunto de datos de pruebas. Estos dos valores son las medias de la suma de los cuadrados de los clasificadores $SSA = N\sum_{i=1}^{N}(M_j)^2$ y $SSAB = SST - SSA - SSB$ la suma de los cuadrados de la interacción clasificación-objeto, calculada a partir de la suma total de cuadrados $SST = M \cdot n \cdot ACC_{avg}^2(1 - ACC_{avg}^2$ y la suma de los cuadrados de los objetos $SSB = \frac{1}{M}\sum_{j=1}^{n}(M_j)^2 - M \cdot n \cdot ACC_{avg}^2$ (en ambas fórmulas $ACC_{avg} = \sum_{i=1}^{M} ACC_i$ es el promedio de las precisiones de los diferentes modelos).

Test t pareado remuestreado Consideramos dos clasificadores C_1 y C_2 y la hipótesis nula de que tienen el mismo rendimiento. Según este método, se divide el conjunto de datos etiquetado en dos partes (entrenamiento y pruebas) y se repite k veces, en cada iteración ambos modelos se ajustan al mismo conjunto de entrenamiento y se evalúan en el mismo conjunto de pruebas, obteniendo k evaluaciones. Después, se calcula el estadístico $t = \frac{ACC_{avg}\sqrt{k}}{\sqrt{\sum_{i=1}^{k}(ACC_i - ACC_{avg})/(k-1)}}$, de acuerdo al test t de Student, donde $ACC_i = ACC_{i,C_1} - ACC_{i,C_2}$

es la diferencia entre las precisiones de los modelos en la iteración i y $ACCavg = \frac{1}{k}\sum_{i=1}^{k} ACC_i$ representa la diferencia media entre rendimientos.

Test t pareado sobre k-fold cross-validation Este método es idéntico al anterior, reemplazando el muestreo por la validación cruzada de k iteraciones.

Test F 5x2CV combinado de Alpaydin El procedimiento de este enfoque es similar al de una de las técnicas objeto de esta memoria que se explica en mayor profundidad en el siguiente apartado. La diferencia fundamental es el estadístico empleado, en este caso $f = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}(ACC_{i,j})^2}{2\sum_{i=1}^{5} s_i^2}$.

2. Técnicas empleadas

En este informe se presentan los resultados obtenidos tras realizar pruebas experimentales sobre dos clasificadores empleando dos test estadísticos recomendados en [7]. En esta sección se describen brevemente ambos métodos.

Test de McNemar Este test estadístico no paramétrico se basa en una matriz de confusión 2x2(también denominada tabla de contingencia), cuya estructura se muestra en la Figura 2, para comparar dos modelos de clasificadores.

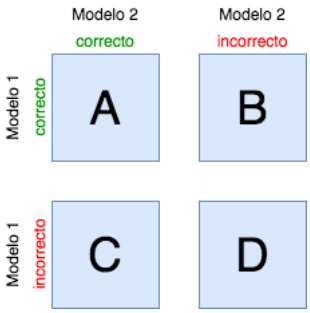

Figura 2: Matriz de confusión

Partiendo de la hipótesis nula de que ambos modelos poseen el mismo rendimiento, se emplea el estadístico $\chi^2 = \frac{(B-C)^2}{B+C}$, o su versión mejorada $\chi^2 = \frac{(|B-C|-1)^2}{B+C}$, para calcular un valor p (probabilidad de observar el valor χ^2 o mayor). Si este valor p es menor que el nivel de significancia α preestablecido, podemos rechazar la hipótesis nula.

Según las conclusiones de la comparación entre los distintos métodos de selección presentada en [6], el test de McNemar presenta una tasa baja de falsos positivos (errores de tipo I) y es más rápido, puesto que solo necesita ser ejecutado una única vez. Por ejemplo, el test t pareado sobre k-fold cross-validation precisa k veces más cómputos que el test de McNemar, ya que se reajustan los modelos con los conjuntos de entrenamiento, y posee una tasa de falsos positivos algo elevada. Por otro lado, el test Q de Cochran se puede emplear junto al test de McNemar en comparaciones de más de dos clasificadores.

Test t pareado sobre 5x2-*Fold Cross Validation* Según este enfoque, se realiza validación cruzada con $K = 2$ durante 5 iteraciones, ajustando los dos clasificadores a comparar con los datos de entrenamiento y evaluándolos sobre los de test, que se intercambian en cada iteración. Se calcula la precisión de cada clasificador (C_1 y C_2) de

forma que $ACC_A = ACC_{A,C_1} - ACC_{A,C_2}$ y $ACC_B = ACC_{B,C_1} - ACC_{B,C_2}$. Posteriormente, la media $ACC_{avg} = (ACC_A + ACC_B)/2$ se emplea para calcular la varianza $s^2 = (ACC_A - ACC_{avg})^2 + (ACC_B - ACC_{avg}^2)^2$ de los resultados. Esta varianza se utiliza en el cálculo del estadístico t, de forma que $t = \frac{ACC_{A,1}}{\sqrt{(1/5)\sum_{i=1}^{5} s_i^2}}$ donde $ACC_{A,1}$ es el ACC_A obtenido en la primera iteración. Finalmente, se calcula el valor p utilizando la distribución de t y se compara con α permitiendo rechazar o aceptar la hipótesis nula.

Este test presenta un ratio de falsos positivos similar al del test de McNemar y resulta ligeramente más potente que este, aunque es menos eficiente computacionalmente. Es una alternativa recomendable al test t pareado remuestreado, ya que este tiene una tasa alta de falsos positivos y requiere mayor coste computacional. Sin embargo, el test F 5x2CV combinado de Alpaydin es una alternativa similar y con mayor robustez.

3. Resultados

Los clasificadores empleados para llevar a cabo los experimentos son k vecinos más cercanos (KNN) y análisis discriminante lineal (LDA). El primero con un valor $k = 5$ y el segundo empleando descomposición en valores singulares (ambos valores por defecto). Se utiliza el conjunto de datos iris[3], formado por 3 clases de 150 muestras cada una.

Para realizar la comparación entre ambos modelos, en primer lugar se divide (de forma estratificada) el conjunto de datos en entrenamiento (70 % - 115 muestras) y pruebas (30 % - 45 muestras). Ambos modelos se ajustan utilizando el conjunto de entrenamiento y se evalúa su rendimiento a partir de los resultados de las predicciones sobre el conjunto de pruebas de forma individual. La precisión del clasificador KNN es de 95.56 % y la de LDA 97.78 %. ambas calculadas sobre este conjunto de datos. A priori se podría pensar, en base a estos resultados, que el modelo LDA tiene un mejor rendimiento, sin embargo, como se expone en esta memoria estos datos no son significativos por sí mismos, puesto que solo se ha efectuado una única prueba.

Como línea base para la discusión de resultados, se aplica validación cruzada de 10 iteraciones en cada uno de los modelos, obteniendo una precisión media de $0{,}9667 \pm 0{,}0447$ para KNN y $0{,}98 \pm 0{,}0305$ con LDA. Estos resultados indican que el rendimiento de ambos modelos es similar y que no existe evidencia de que uno resulte superior al otro.

En el caso del Test de McNemar la matriz de confusión obtenida en base a las predicciones de ambos modelos sobre el conjunto de test es $\begin{bmatrix} 42 & 0 \\ 2 & 1 \end{bmatrix}$, es decir, ambos modelos clasifican correctamente 42 muestras e incorrectamente 1, mientras que el clasificador KNN comete 2 errores que LDA no. A esta matriz se le aplica el Test de McNemar corregido, obteniendo $\chi^2 = 0{,}5$ y $p = 0{,}4795$. Analíticamente es inmediato comprobar que el valor del estadístico es correcto, ya que $\chi^2 = \frac{(|0-2|-1)^2}{0+2} = 0{,}5$.

El Test t de Student pareado sobre los resultados de 5x2CV se aplica sobre el conjunto de datos completo, sin realizar antes una división de los datos en entrenamiento y pruebas, puesto que esta división se realiza iterativamente durante la validación cruzada. Como resultado se obtiene un valor de $t = -1{,}3693$ y $p = 0{,}2292$.

[3] `https://archive.ics.uci.edu/ml/datasets/iris`

Empleando un nivel de significación $\alpha = 0{,}05$, comúnmente utilizado (específicamente en la literatura consultada) y partiendo de la hipótesis nula de que ambos clasificadores poseen el mismo rendimiento. Se puede concluir que los valores de p obtenidos con los dos test estadísticos son menores que α, por lo que se acepta la hipótesis nula.

Al tratarse de un conjunto de datos de tamaño pequeño no hay una diferencia notable de coste computacional entre las ejecuciones de ambos test. Además, los resultados concuerdan con los obtenidos aplicando *10-fold cross-validation*. Es necesario resaltar la notoria simplicidad de este ejemplo, tanto por el tamaño del conjunto de datos como por los clasificadores empleados, con una configuración por defecto, sin reparar en el efecto de la variación de hiperparámetros o método de ajuste empleado. Sin embargo, se considera suficiente para comparar las técnicas de selección de modelos e ilustrar el proceso seguido.

4. Conclusiones

En este informe se recoge un breve análisis sobre el problema de la comparación y selección de modelos en el ámbito del aprendizaje automático. Son múltiples los métodos que se han propuesto y estudiado a lo largo de los años para llevar a cabo una comparación eficaz y eficiente entre diferentes sistemas, los que se han considerado más destacables se definen de forma concisa en las primeras secciones de este documento.

Uno de los métodos más empleados, si no el que más, es la validación cruzada, ya que permite la reutilización eficiente de muestras y la evaluación de modelos de forma no paramétrica. Siendo suficientemente eficaz en conjuntos de datos de menor tamaño. Sin embargo, al ser un método basado en varias iteraciones puede resultar en un alto coste de computación al combinarse con otras técnicas.

El uso de test estadísticos para realizar la comparación entre modelos o algoritmos es también una práctica bastante extendida. En particular, se han realizado experimentos con dos técnicas dentro de este ámbito: test de McNemar y test t pareado sobre 5x2CV. Se han elegido dos clasificadores, sin realizar ningún método de selección de complejidad y se ha comparado su rendimiento mediante ambos métodos en la tarea de clasificación del conjunto de datos iris. En ambos casos, los test estadísticos indican que no existe una diferencia significativa entre los clasificadores, también se realiza validación cruzada de 10 iteraciones para comprobar que la conclusión extraída del resultado es la misma. No obstante, el procedimiento seguido para cada una de las técnicas de comparación es ligeramente diferente. En el caso del Test de McNemar, se realiza una única vez la división en conjuntos de entrenamiento y pruebas y se calcula el estadístico χ^2 a partir de los resultados de las predicciones sobre el conjunto de pruebas. Por otra parte, al aplicar 5 veces validación cruzada de 2 iteraciones se obtiene la precisión media de los dos clasificadores sobre las diferentes combinaciones de datos; empleando la varianza de esos resultados y la diferencia de precisiones obtenida en la primera iteración con la primera combinación, se calcula el estadístico t.

Tanto la bibliografía consultada como los resultados obtenidos en los experimentos son una evidencia de que la evaluación individual[4] de varios modelos es insuficiente para decantarse de forma razonada por uno de ellos y, por tanto, resulta evidente la necesidad de un procedimiento

[4]Las tasas de error (o medida de rendimiento empleada) no resultan de utilidad por sí solas, la información no resulta suficiente de no ser acompañadas por su varianza.

de selección de modelo donde se realice el entrenamiento de los modelos a comparar, se estime el error de generalización y se evalúe de forma lógica si la diferencia de resultados es significativa.

Referencias

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[2] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 03 1978. [Online]. Available: https://doi.org/10.1214/aos/1176344136

[3] J. Rissanen, "Paper: Modeling by shortest data description," *Automatica*, vol. 14, no. 5, p. 465–471, Sep. 1978. [Online]. Available: https://doi.org/10.1016/0005-1098(78)90005-5

[4] V. Vapnik and A. Y. Chervonenkis, "Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya. (russian) [theory of pattern recognition: Statistical problems of learning]," 1974.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Math. Intell., 01 2017, vol. 27.

[6] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018.

[7] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.