Trabajo de Fin de Grado

Evaluación del estado del arte en técnicas estadísticas para el análisis comparativo de algoritmos de aprendizaje automático

Jacinto Carrasco Castillo

18 de septiembre de 2016

Planteamiento

Descripción del problema

La comparación de algoritmos debe realizarse de manera rigurosa para reducir la aleatoriedad asociada a los experimentos. Para realizar una correcta comparación es necesario:

- Selección de medida del rendimiento.
- Validación cruzada y remuestreo.
- Test estadísticos.

Descripción del problema

La comparación de algoritmos debe realizarse de manera rigurosa para reducir la aleatoriedad asociada a los experimentos. Para realizar una correcta comparación es necesario:

- Selección de medida del rendimiento.
- Validación cruzada y remuestreo.
- Test estadísticos.

Selección de medida del rendimiento. Diferentes medidas según el problema a resolver. Podemos interesarnos por maximizar la precisión o minimizar una función de coste o pérdida.

Validación cruzada y remuestreo. Para la validación de los resultados es necesario repetir los experimentos. Debido a que el número de datos disponibles es limitado es necesario reutilizar los datos.

Test estadísticos. Responden a las preguntas sobre si se pueden justificar los resultados estadísticamente o se han obtenido por azar. Los test se aplica a comparar dos algoritmos en un dominio, dos algoritmos sobre varios problemas o comparar múltiples algoritmos en varios dominios.

Objetivos

- Estudio de los test estadísticos disponibles.
- Integración de herramientas informáticas existentes para la aplicación de estos test.
- Comparación de las propiedades de los test.

Contenido matemático

Test paramétricos

Suponen que la muestra pertenece a una distribución conocida.

Test binomial para una muestra Comprobación de que el rendimiento para un problema sea igual a un valor θ_0.

***t*-test para muestras apareadas** Comparación de la media de dos muestras. Comparación de dos algoritmos para un conjunto de datos.

ANOVA Comparación de la media de múltiples algoritmos en múltiples problemas. Trata la varianza en un grupo, entre grupos y la combinación de ellas.

Suponen que la muestra pertenece a una distribución conocida.

Test binomial para una muestra Comprobación de que el rendimiento para un problema sea igual a un valor θ_0.

t-test para muestras apareadas Comparación de la media de dos muestras. Comparación de dos algoritmos para un conjunto de datos.

ANOVA Comparación de la media de múltiples algoritmos en múltiples problemas. Trata la varianza en un grupo, entre grupos y la combinación de ellas.

t-**test**: Para el t-test surgen modificaciones que ofrecen mejores resultados usando validación cruzada 5×2, esto es, se se divide en dos particiones, usando cada una de ellas como entrenamiento y validación, y se repite cinco veces. Para cada test se incluye en la memoria la obtención del estadístico y la distribución que sigue para determinar si el estadístico se encuentra en la región crítica, esto es, la región donde rechazaríamos la hipótesis nula.

Test no paramétricos

Comparación con test paramétricos

- Los test no paramétricos no suponen la pertenencia de la distribución a ninguna familia de distribuciones.
- Las hipótesis para aplicar estos test son más generales.
- Si se dan las hipótesis necesarias para los test paramétricos, tienen una menor potencia.
- Si se disponen de pocos datos las hipótesis de los test paramétricos no suelen darse y los test no paramétricos suplen la falta de potencia con mayor precisión.

- Los test no paramétricos no suponen la pertenencia de la distribución a ninguna familia de distribuciones.
- Las hipótesis para aplicar estos test son más generales.
- Si se dan las hipótesis necesarias para los test paramétricos, tienen una menor potencia.
- Si se disponen de pocos datos las hipótesis de los test paramétricos no suelen darse y los test no paramétricos suplen la falta de potencia con mayor precisión.

Las hipótesis de los test paramétricos son cuestiones más generales como la simetría de la población o la continuidad, mientras que en los test paramétricos se suelen exigir la normalidad de la población o la igualdad de las varianzas.

Test no paramétricos

Tipos de test no paramétricos:

- Test de aleatoriedad: basados en el número de rachas.
- Test de bondad del ajuste: Test chi cuadrado, Kolmogorov-Smirnov.
- Análisis del conteo de datos: Test de McNemar.
- Test basado en una muestra y muestras apareadas: Test de signo, test de rangos con signos de Wilcoxon.
- Análisis bidimensional de la varianza: Test de Friedman, modificación de Iman-Davenport, test de rangos alineados de Friedman, test de Quade.

Tipos de test no paramétricos:

- Test de aleatoriedad: basados en el número de rachas.
- Test de bondad del ajuste: Test chi cuadrado, Kolmogorov-Smirnov.
- Análisis del conteo de datos: Test de McNemar.
- Test basado en una muestra y muestras apareadas: Test de signo, test de rangos con signos de Wilcoxon.
- Análisis bidimensional de la varianza: Test de Friedman, modificación de Iman-Davenport, test de rangos alineados de Friedman, test de Quade.

Los test de bondad del ajuste podemos usarlos para aplicar test paramétricos con mayor certeza sobre la normalidad de la población. Usamos para la comparación de dos algoritmos en un problema el test de McNemar, los test de signo y de rangos con signos para varios problemas. Los restantes test para múltiples algoritmos en varios problemas. El test de Iman-Davenport tiene una mayor potencia. El test de Quade tiene en cuenta la dificultad de los test.

Test no paramétricos

Procedimientos *post-hoc*

Los test que comparan múltiples algoritmos indican la existencia de diferencias entre ellos. Necesitan un test adicional para indicar dónde están estas diferencias.

- *p*-valores ajustados: Al realizar múltiples comparaciones aumenta la probabilidad de cometer un error de tipo I. Hay distintos métodos para ajustar los *p*-valores obtenidos.
  - Procedimientos de un paso: Bonferroni-Dunn.
  - Procedimiento descendente: Holm, Holland, Finner.
  - Procedimiento ascendente: Hochberg, Hommel, Rom.
  - Procedimiento en dos pasos: Li.

Los test que comparan múltiples algoritmos indican la existencia de diferencias entre ellos. Necesitan un test adicional para indicar dónde están estas diferencias.

- *p*-valores ajustados: Al realizar múltiples comparaciones aumenta la probabilidad de cometer un error de tipo I. Hay distintos métodos para ajustar los *p*-valores obtenidos.
  - Procedimientos de un paso: Bonferroni-Dunn.
  - Procedimiento descendente: Holm, Holland, Finner.
  - Procedimiento ascendente: Hochberg, Hommel, Rom.
  - Procedimiento en dos pasos: Li.

El procedimiento de Hommel es más complejo. El procedimiento de Li ofrece resultados más inestables cuando los *p*-valores originales son mayores a 0.05.

Test no paramétricos

Test basados en permutaciones

Son test no paramétricos. La hipótesis nula es que una muestra **x** proviene de una misma población.
Suponiendo H_0 cierta, los individuos de cada subconjunto de la muestra se puede intercambiar por los de otro subconjunto.

Definición (Principio de los test basados en permutaciones)

Si dos experimentos toman valores en el mismo espacio muestral Ω con distribuciones P_1, P_2 dan el mismo conjunto de datos, las inferencias condicionales a los datos usando el mismo estadístico deben ser la misma.

Comparación con THN

La inferencia bayesiana ajusta un modelo de probabilidad a los datos y obtiene una distribución sobre los parámetros del modelo. Las diferencias con los test de hipótesis nula son:

- Se evitan decisiones dicotómicas marcadas por α.
- Los THN no estiman la probabilidad de la hipótesis.
- Con suficientes datos se rechaza casi toda hip. nula.
- No se tiene en cuenta magnitud de la diferencia ni incertidumbre.
- No se obtiene información si no se rechaza la hipótesis nula.

La inferencia bayesiana ajusta un modelo de probabilidad a los datos y obtiene una distribución sobre los parámetros del modelo. Las diferencias con los test de hipótesis nula son:

- Se evitan decisiones dicotómicas marcadas por α.
- Los THN no estiman la probabilidad de la hipótesis.
- Con suficientes datos se rechaza casi toda hip. nula.
- No se tiene en cuenta magnitud de la diferencia ni incertidumbre.
- No se obtiene información si no se rechaza la hipótesis nula.

Con los test bayesianos sí se responde a la pregunta más natural de "¿cuál es la probabilidad de que el rendimiento de dos clasificadores sea el mismo?".

Con los THN podríamos pensar que se obtienen datos hasta que se consigue rechazar la hipótesis nula.

El enfoque realizado es la estimación bayesiana de parámetros, pues en el análisis mediante el factor de Bayes se dan problemas similares a los mencionados.

Test bayesianos

t-test bayesiano correlado Comparación de dos dos algoritmos en un único conjunto de datos. Tiene en cuenta la correlación de los conjuntos de entrenamientos en CV. Se obtiene una distribución T de Student sobre la diferencia de las medias.

Test bayesiano de signo Versión bayesiana del test de signo. Se obtiene distribución sobre la probabilidad de que la diferencia entre algoritmos esté en la *rope*, a la izquierda o a la derecha.

Test bayesiano de rangos alineados Versión bayesiana del test de rangos alineados. No se obtiene una distribución clara de los parámetros, pero se puede obtener una muestra.

Se considera la *rope* la región de una equivalencia entre los algoritmos.Si la distribución se encuentra dentro de esta región, consideraremos equivalentes el rendimiento de los algoritmos para ese conjunto de datos. Si la distribución se encuentra a la derecha o a la izquierda, un algoritmo será mejor que otro.

Contenido informático

rNPBST

Paquete desarrollado en `R`. Incluye:

- Conjunto de datos sobre la aplicación de 5 algoritmos en 29 conjuntos de datos para realizar los test.
- Test no paramétricos de la biblioteca `JavaNPST` a los que se accede mediante `rJava`.
- Test bayesianos y representación gráfica de sus resultados.

rNPBST

Instalación

Paquete de `R` disponible en

`https://github.com/JacintoCC/TFG/tree/master/rNPBST`

Para la instalación, ejecutar donde se encuentre la carpeta con el software:

```
> R CMD build rNPBST
> R CMD INSTALL rNPBST
```

Aplicación de test

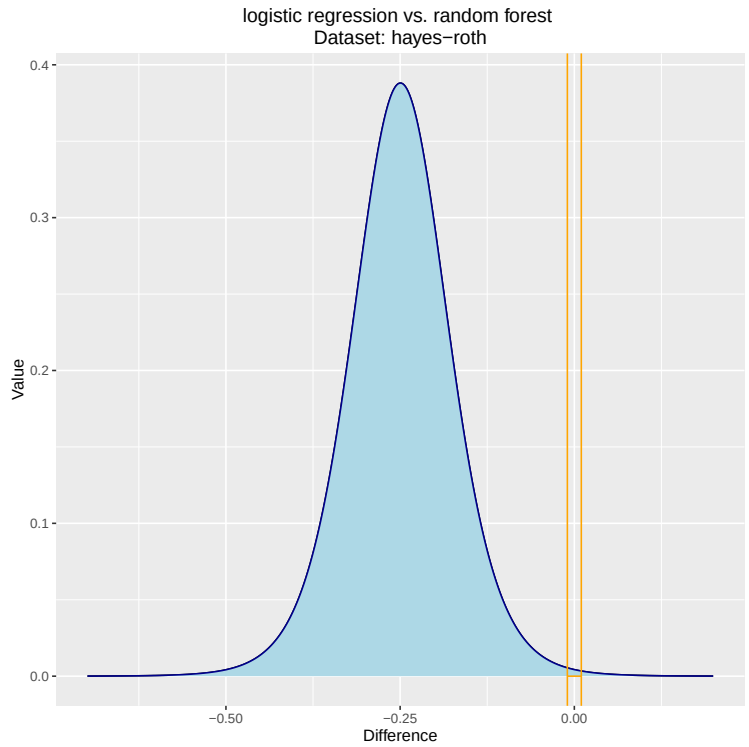
```
> data("results")
> ft <- friedman.test(results)
> ft$htest
    Friedman test
data: results
s = 2812.000, q = 39.056, p-value = 6.789e-08
```

Con estos resultados rechazaríamos la hipótesis nula de la equivalencia de los algoritmos. El test nos devuelve en el parámetro `report` también la suma del ránquin medio de cada algoritmo.

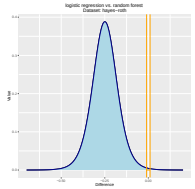
Aplicación de test

Test bayesianos - *t*-Test bayesiano correlado

```
> dataset.index <- 13
> correlatedBayesianT.test(results.lr[dataset.index, ],
                           results.rf[dataset.index, ])
```



```
> dataset.index <- 13
> correlatedBayesianT.test(results.lr[dataset.index, ],
                           results.rf[dataset.index, ])
```

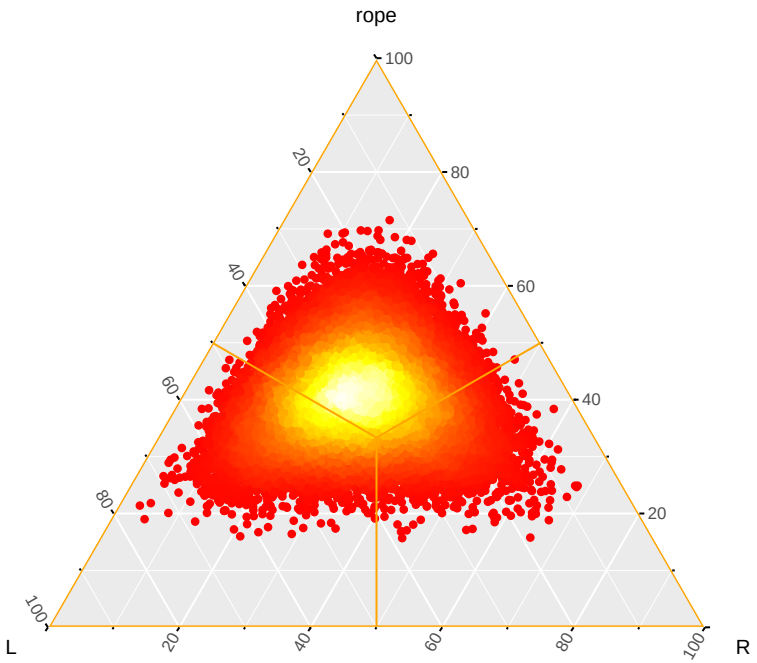


Al situarse la mayor parte de la distribución a la izquierda de la *rope*, tenemos que para este conjunto de datos el algoritmo *random forest* funciona mejor que la regresión logística.

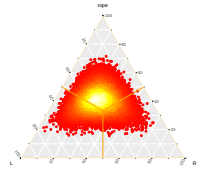
Aplicación de test

Test bayesianos - Test bayesiano de rangos con signo

```
> bayesianSignedRank.test(results$KNN,
                          results$neural.network)
```

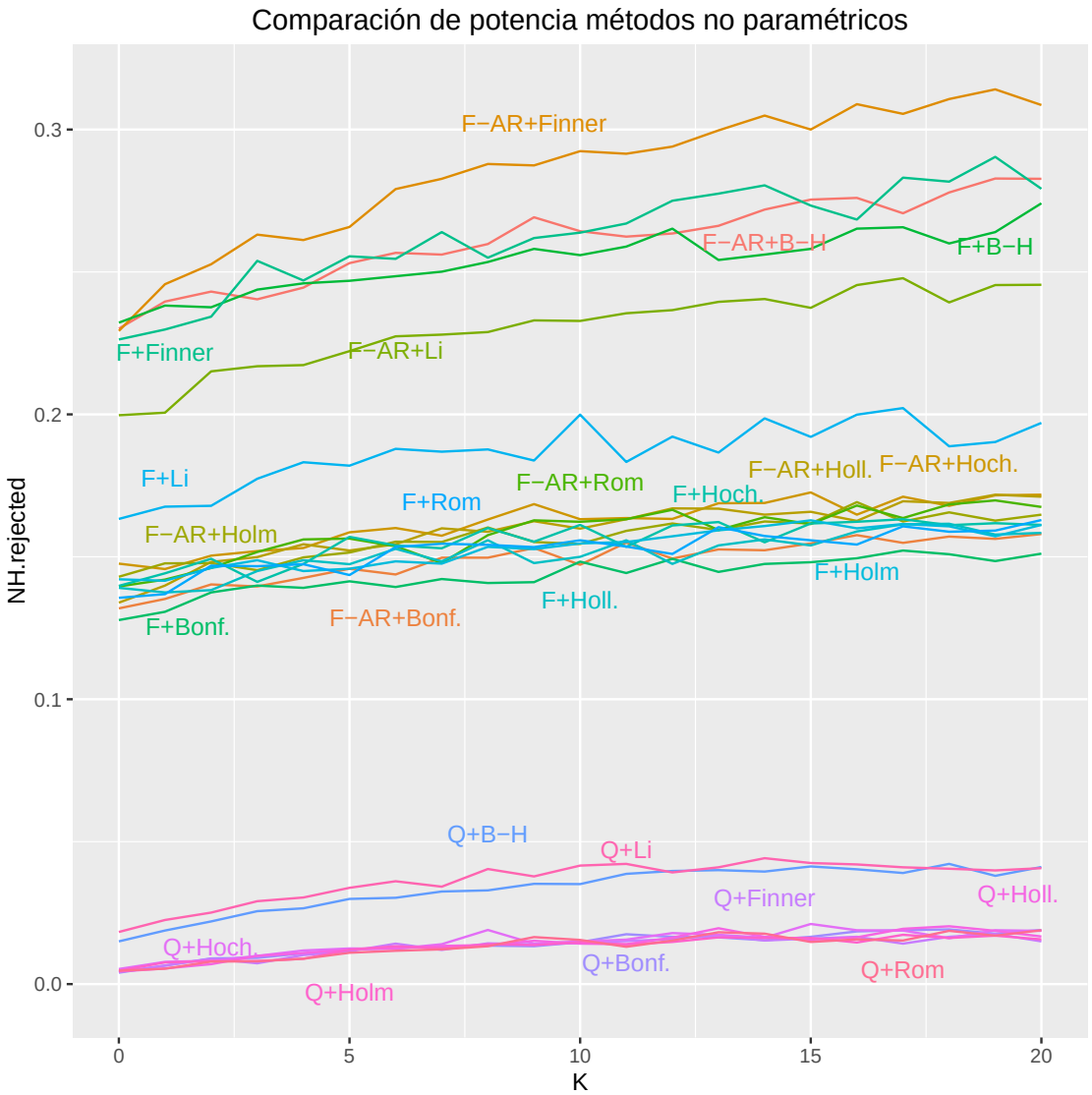


```
> bayesianSignedRank.test(results$KNN,
                          results$neural.network)
```

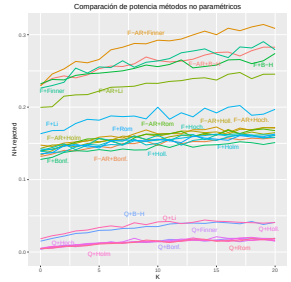


Se observa una mayor concentración en la región donde hay más probabilidad de que el parámetro caiga en la *rope*, con lo que no podríamos decir que hay diferencias entre los algoritmos comparados

Comparación de la potencia de test estadísticos



Comparación de la potencia de test estadísticos



Se han seleccionado aleatoriamente un subconjunto de conjuntos de datos y se ha comprobado el número de hipótesis nulas rechazadas. El parámetro K influye en escoger aquellos conjuntos de datos con una mayor diferencia. Para estos datos el test de Quade obtiene los peores resultados y los métodos para ajustar los p-valores con mayor potencia son el de Finner y el de Bergmann y Hommel. Otro posible experimento podría consistir en ver el error de tipo I de cada test, es decir, cuándo se rechaza la hipótesis nula siendo cierta.