Evaluación

José Tomás Palma Méndez

Dept. de Ingeniería de la Información y las Comunicaciones. Universidad de Murcia
Contacting author: jtpalma@um.es

1. Introducción

En esta sesión prácticas vamos analizar cómo realizar comparaciones entre predictores en diferentes situaciones. Los paquetes necesarios para esta prácticas son: `PMCMR`, `nortest`, `car` y `multcomp`.

Los conjuntos de datos utilizados se pueden descargar del aula virtual.

2. Dos clasificadores en un dominio

2.1. Caso paramétrico

Para un determinado conjunto de datos se han probado dos clasificadores: C50 y redes neuronales. Cada clasificares se han evaluado utilizando una validación cruzada con 10 pliegues. En el fichero `ejemplo1.dat` puedes encontrar el índice **Kappa** para cada uno de los 10 pliegues (Tabla 1). Con estos datos ¿qué podemos decir sobre la eficacia de dichos clasificadores?.

	nnet	C50
1	0.8148148	0.6666667
2	0.5897436	0.7647059
3	0.6000000	0.8181818
4	0.1794872	0.2941176
5	0.8148148	1.0000000
6	0.0000000	-0.1111111
7	0.0000000	0.3333333
8	0.7647059	0.7647059
9	0.8181818	0.8181818
10	0.0000000	-0.1111111

Tabla 1. Índice Kappa en cada uno de los pliegues.

Como podemos observar estamos comparando dos clasificadores en un dominio, por lo tanto, primero deberíamos comprobar si se cumplen las condiciones para realizar un **test t de Student con medidas pareadas**: normalidad y homogeneidad de las varianzas (homocedasticidad). Para comprobar la normalidad podemos utilizar dos test: el **test de Shapiro-Wilk** y el **test de Kolmogorov-Smirnov con la corrección Lilliefors**. El test de Shapiro-Wilk se recomienda cuando el número de

muestras es menor a 50, mientras que el test de test de Kolmogorov-Smirnov con la corrección Lilliefors se recomienda cuando las muestras son mayores a 50. En este caso, aunque sólo tenemos 10 muestras vamos a realizar ambos test para aclarar cómo se aplican.

Primero cargamos los datos.

```
> ejemplo1 <- read.csv("ejemplo1.dat")
```

Ahora ya podemos proceder a realizar los test de normalidad. Recordad que la normalidad hay que comprobarla sobre las diferencias entre las medidas obtenidas en cada pliegue, ya que lo que queremos comprobar es si existen diferencias significativas entre las medias del rendimiento entre las dos clasificadores.

```
> shapiro.test(ejemplo1$nnet-ejemplo1$C50)

        Shapiro-Wilk normality test

data:  ejemplo1$nnet - ejemplo1$C50
W = 0.93528, p-value = 0.5018

> library(nortest)
> lillie.test(ejemplo1$nnet-ejemplo1$C50)

        Lilliefors (Kolmogorov-Smirnov) normality
        test

data:  ejemplo1$nnet - ejemplo1$C50
D = 0.1602, p-value = 0.6613
```

Como podemos ver, en este caso podemos afirmar que la muestra procede de una distribución normal con un 95 % de confianza ($W = 0{,}03$, p-value $= 0{,}5$ y $D = 0{,}16$, p-value $= 0{,}66$). En este caso sólo nos habría hecho falta aplicar el test de Shapiro-Wilk al ser la muestra menor que 50.

Una vez comprobada la normalidad, hay que proceder con la comprobación de la homocedasticidad (homogeneidad de las varianzas). Para ello tenemos dos test: el **test de Levene** y el **test de Bartlett**. Este último es más potente en el caso de que se pueda asumir la normalidad, como es nuestro caso.

```
> datos <- stack(list(nnet=ejemplo1$nnet,C50=ejemplo1$C50))
> library(car)
> leveneTest(values ~ ind, data = datos)

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0017 0.9672
      18
```

```
> bartlett.test(values ~ ind, data = datos)

        Bartlett test of homogeneity of variances

data:  values by ind
Bartlett's K-squared = 0.05302, df = 1,
p-value = 0.8179
```

Como podemos ver, los dos test confirman al 95 % de confianza que las varianzas en las medidas obtenidas en cada clasificador son homogéneas ($K^2 = 0{,}053$, p-value $= 0{,}82 > 0{,}05$).

Al cumplirse las condiciones de aplicabilidad del test t de Student con medidas pareadas, podemos comprobar si las diferencias detectadas en los clasificadores es significativa o no:

```
> t.test(ejemplo1$nnet,ejemplo1$C50, paired = TRUE)

        Paired t-test

data:  ejemplo1$nnet and ejemplo1$C50
t = -1.2692, df = 9, p-value = 0.2362
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.18250096  0.05131642
sample estimates:
mean of the differences
            -0.06559227
```

Atendiendo a los resultados del test, podemos concluir que con el 95 % de confianza las diferencias entre las medias obtenidas por cada clasificador no son significativas (no podemos rechazar la hipótesis de que la diferencias entre las medias sea 0). Es decir, no existen diferencias significativas entre los dos clasificadores con el conjunto de datos utilizado ($|T| = 1{,}27 < T_{9,0,975}$, p-value $= 0.236$). Hay que tener en cuenta que hubiera dado el mismo resultado si utilizamos un test t de Student para una muestra, que es lo que se hace el paquete `caret` con la función `compare_models()`. En este caso la hipótesis nula consiste en que la media de la población es 0. Para ello, hay que llamar a la función `t.test()` con la diferencia obtenida en cada pliegue:

```
> t.test(ejemplo1$nnet-ejemplo1$C50)

        One Sample t-test

data:  ejemplo1$nnet - ejemplo1$C50
t = -1.2692, df = 9, p-value = 0.2362
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```
 -0.18250096  0.05131642
sample estimates:
  mean of x
-0.06559227
```

2.2. Caso no paramétrico

Supongamos ahora que queremos comprobar como se comportan un clasificador KNN y un análisis lineal discriminante sobre determinado conjunto de datos. Una vez entrenados los clasificadores se han evaluado mediante una validación cruzada con 10 pliegues. En el fichero `ejemplo2.dat` se pueden encontrar los valores para la precisión para cada uno de los pliegues (Tabla 2).

	SVM	LDA
1	0.9333333	0.9333333
2	0.8750000	0.8750000
3	0.8125000	0.8750000
4	0.8125000	0.8750000
5	0.9333333	0.9333333
6	0.8000000	0.8000000
7	0.6000000	0.5333333
8	0.9375000	0.9375000
9	0.7500000	0.9375000
10	0.8000000	0.7333333

Tabla 2. Precisión en cada uno de los pliegues.

Primero cargamos los datos.

```
> ejemplo2 <- read.csv("ejemplo2.dat")
```

Ahora volvemos a comprobar las condiciones de aplicabilidad del test t de Student con medidas pareadas. Primero comprobamos la normalidad de la muestra:

```
> shapiro.test(ejemplo2$SVM-ejemplo2$LDA)

        Shapiro-Wilk normality test

data:  ejemplo2$SVM - ejemplo2$LDA
W = 0.83634, p-value = 0.03989
```

Como podemos afirmar que las muestras **no** proceden de una distribución normal con el 95 % de confianza ($W = 0{,}86$, p-value $= 0{,}040$), no podemos aplicar el test t

de Student para muestras pareadas (o su versión para una muestra). En este caso tenemos que aplicar los test no paramétricos como el **test de McNemar** o el **test de la suma de rangos de Wilcoxon para muestras pareadas**.

Empecemos por el test de McNemar, que necesita como datos la matriz de confusión (ver Tabla 3) para los dos métodos considerados:

	FALLECE	**VIVE**
FALLECE	31	11
VIVE	17	136

Tabla 3. Matriz de confusión para el ejemplo 2.

```
> matriz.conf <- read.csv("matrizConfEjemplo2.dat",
                          header = TRUE, row.names =1)
```

y ahora ya podemos realizar los tests:

```
> mcnemar.test(as.matrix(matriz.conf))

        McNemar's Chi-squared test with continuity
        correction

data:  as.matrix(matriz.conf)
McNemar's chi-squared = 0.89286, df = 1,
p-value = 0.3447

> wilcox.test(ejemplo2$SVM,ejemplo2$LDA, paired = TRUE)

        Wilcoxon signed rank test with continuity
        correction

data:  ejemplo2$SVM and ejemplo2$LDA
V = 7, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```

En este caso, el test de Wilcoxon no resulta muy fiable, al aparecer pliegues con empates y ceros. Por lo tanto, debemos fiarnos del resultado del test de McNemar que en este caso nos indica que, con una confianza del 95 %, no existen diferencias significativas ($\tilde{\chi}^2_{Mc} = 0{,}89 < \tilde{\chi}^2_{1,0,05}$, p-value $= 0.34 > 0.05$).

3. Dos predictores en múltiples dominios

Para comparar dos predictores se han utilizado 10 conjuntos de datos. Concretamente, se ha calculado la precisión de una máquina de soporte de vectores (SVM) con

kernel lineal y un árbol C50 en cada uno de los 10 conjuntos de datos. En la tabla `ejemplo3` se recoge dicha información (ver Tabla 4).

	X	SVMLinear	C50
1	hepatitis	0.8441667	0.8629167
2	iris	0.9600000	0.9466667
3	cox2	0.8050416	0.8352914
4	oil	0.9691667	0.9597980
5	dhfr	0.9261364	0.9107008
6	German	0.7470000	0.7560000
7	Seg	0.8033200	0.8459288
8	Breast	0.9642437	0.9614067
9	Pima	0.7734450	0.7578606
10	Sonar	0.7456061	0.8320779

Tabla 4. Precisión en cada uno de los conjuntos de datos.

Primero tenemos que cargar los datos:

```
> ejemplo3 <- read.csv("Ejemplo3.dat")
```

En este caso estamos comparando dos clasificadores en múltiples dominios y, como hemos comentado en clase de teoría, la recomendación es utilizar el **test de la suma de rangos de Wilcoxon para muestras pareadas** (o su versión de una muestra, calculando la diferencia de la precisión obtenida en cada dataset).

```
> wilcox.test(ejemplo3$SVMLinear,ejemplo3$C50, paired = TRUE)

        Wilcoxon signed rank test

data:  ejemplo3$SVMLinear and ejemplo3$C50
V = 19, p-value = 0.4316
alternative hypothesis: true location shift is not equal to 0
```

Como podemos ver, podemos afirmar que, con una confianza del 95 %, no existen diferencias significativas entre la precisión de los dos clasificadores sobre el conjunto de datos seleccionados ($W = 19 > W^{10}_{0,05}$, p-value $= 0{,}4316 > 0{,}05$).

4. Múltiples predictores en múltiples dominios

4.1. Caso paramétrico

Para esta sección vamos a seguir uno de los ejemplos presentados en [1]. En este ejemplo dispones de los datos de precisión de tres clasificadores en 10 conjunto de datos

	classA	classB	classC
1	0.8583	0.7586	0.8419
2	0.8591	0.7318	0.8590
3	0.8612	0.6908	0.8383
4	0.8582	0.7405	0.8511
5	0.8628	0.7471	0.8638
6	0.8642	0.6590	0.8120
7	0.8591	0.7625	0.8638
8	0.8610	0.7510	0.8675
9	0.8595	0.7050	0.8803
10	0.8612	0.7395	0.8718

Tabla 5. Precisión de los clasificadores en los diferentes conjuntos de datos.

diferentes (ver Tabla 5). El objetivo es determinar si existen diferencias significativas entre los distintos clasificadores en dichos conjunto de datos.

En este caso, al comparar múltiples clasificadores en múltiples dominios, debemos aplicar el **test ANOVA de una vía para medidas pareadas** o su equivalente no paramétrico el **test de Friedman**. Para ello debemos de comprobar las condiciones de aplicabilidad para el test ANOVA de una vía para medidas pareadas: normalidad y esfericidad (el equivalente a la homocedasticidad para el ANOVA de medidas pareadas).

En primer lugar cargamos los datos.

```
> ejemplo4 <- read.csv("Ejemplo4.csv",row.names = 1)
```

Seguidamente vamos a transformar la tabla para que nos sea más cómodo trabajar con algunas de las funciones que vamos a utilizar. La idea es generar una tabla en que tenga tres columnas: una para la precisión, otra para los clasificadores y otra para los conjuntos de datos.

```
> df4.stack <- stack(ejemplo4)
> df4.stack$DataSet <- as.factor(rep(row.names(ejemplo4), times = 3))
> names(df4.stack) <- c("Accuracy","Method", "DataSet")
```

Para la condición de normalidad, debemos comprobar que la muestra de cada uno de los grupos procede de una población con distribución normal. Como hemos comentado anteriormente, esto se puede realizar mediante el test de Shapiro-Wilks:

```
> test <- tapply(df4.stack$Accuracy, df4.stack$Method,shapiro.test)
> test

$classA
```

```
        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.91888, p-value = 0.3477


$classB

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87406, p-value = 0.1114


$classC

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.92585, p-value = 0.4083
```

Como se puede deducir de los datos, las precisiones obtenidas por cada clasificador en cada uno de los conjuntos de datos proceden de una distribución normal con un 95 % de confianza ($W_{ClassA} = 0{,}92$, p-value$_{ClassA} = 0{,}35$, $W_{ClassB} = 0{,}87$, p-value$_{ClassA} = 0{,}11$, $W_{ClassA} = 0{,}92$, p-value$_{ClassA} = 0{,}41$). El siguiente paso es comprobar la homogeneidad de las varianzas, que en el caso de una ANOVA con medidas repetidas se traduce en comprobar la esfericidad de la matriz de covarianzas. Este proceso requiere utilizar la función `Anova` del paquete `car` que realiza el `test de Mauchly` para la comprobación de la esfericidad. En dicho test, la hipótesis nula es que la matriz de covarianzas cumple la propiedad de esfericidad.

```
> model.lm <- lm(as.matrix(ejemplo4) ~ 1)
> library(car)
> design <- factor(c("classA","classB","classC"))
> options(contrasts=c("contr.sum", "contr.poly"))
> aov <- Anova(model.lm, idata=data.frame(design), idesign=~design)
> summary(aov, multivariate=F)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

                SS num Df  Error SS den Df
(Intercept) 19.9103      1 0.0065593      9
design       0.1113      2 0.0070340     18
                  F    Pr(>F)
(Intercept) 27318.99 < 2.2e-16 ***
design        142.42 9.259e-12 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Mauchly Tests for Sphericity

       Test statistic p-value
design        0.67901 0.21257


Greenhouse-Geisser and Huynh-Feldt Corrections
 for Departure from Sphericity

        GG eps Pr(>F[GG])
design 0.75701  2.291e-09 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          HF eps   Pr(>F[HF])
design 0.8776464 1.480308e-10
```

Como se puede apreciar, en primer lugar se nos muestra el resultado del test ANOVA para medidas repetidas suponiendo que se cumple la esfericidad, que nos indica que existen diferencias entre los clasificadores altamente significativas ($F = 142{,}42$, p-value $< 0{,}001$). El segundo resultado se corresponde con el test de Mauchly para la esfericidad, que nos indica que no podemos rechazar la hipótesis de que la matriz de covarianzas cumpla esta propiedad (p-value = 0,21). Los otros resultados que se muestran son el resultado de aplicar un procedimiento que permite corregir la desviación producida por cumplir con la esfericidad a los resultados de test ANOVA anteriormente realizados.

Con estos resultados, tendríamos que proceder con el **test post hoc de Tukey** para determinar dónde están dichas diferencias. Sin embargo, y a modo de ilustración, vamos a ver cómo el la forma de realizar el test ANOVA de una vía para medidas pareadas con la función `aov()` sobre el conjunto de datos o la función `anova()` a un modelo lineal de los datos.

```
> Aov.mod <- aov(Accuracy ~ Method + Error(DataSet),data=df4.stack)
> summary(Aov.mod)

Error: DataSet
          Df   Sum Sq   Mean Sq F value Pr(>F)
Residuals  9 0.006559 0.0007288

Error: Within
          Df  Sum Sq Mean Sq F value   Pr(>F)
Method     2 0.11131 0.05565   142.4 9.26e-12 ***
Residuals 18 0.00703 0.00039
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> library(nlme)
> Lme.mod <- lme(Accuracy ~ Method, random = ~1 | DataSet, data=df4.stack)
> anova(Lme.mod)

            numDF denDF   F-value p-value
(Intercept)     1    18 27318.986  <.0001
Method          2    18   142.418  <.0001
```

Como ya hemos indicado, ahora debemos proceder con el test de Tukey para identificar donde están las diferencias. Sin embargo, la mayoría de las funciones que nos ofrece R para dicho test no se pueden aplicar sobre objetos generados mediante un test ANOVA para medidas pareadas. Una de las posibilidades de realizar este test, consiste en utilizar la función `glht()` del paquete `multcomp`. La función `glht()` nos permite comprobar hipótesis lineales generales y realizar múltiples comparaciones para modelos paramétricos. Los pasos para realizar el test de Tukey con dicha función son los siguientes:

```
> library(multcomp)
> tuk.lme <- glht(Lme.mod, linfct=mcp(Method="Tukey"))
> summary(tuk.lme)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lme.formula(fixed = Accuracy ~ Method, data = df4.stack, random = ~1 |
    DataSet)

Linear Hypotheses:
                      Estimate Std. Error z value
classB - classA == 0 -0.131880   0.008841 -14.918
classC - classA == 0 -0.005510   0.008841  -0.623
classC - classB == 0  0.126370   0.008841  14.294
                     Pr(>|z|)
classB - classA == 0   <1e-04 ***
classC - classA == 0    0.807
classC - classB == 0   <1e-04 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Primero calculamos un modelo lineal generalizado que se intentan compensar el hecho de que no puede existir variabilidad entre los grupos. Una vez tenemos el modelo lineal podemos usar la función `glht()` para realizar el test de Tukey. Como se pueden ver en los resultados, existen diferencias altamente significativas entre el clasificador B y el clasificador A ($Z = -14{,}92$, p-value $< 0{,}001$) y el entre el B y el C ($Z =$

14,29, p-value $<$ 0,001). En ambos casos, los datos estimados para la media de las diferencias indican que, teniendo en cuenta la precisión, el B es peor que el resto de los clasificadores. No podemos establecer diferencias significativas entre el A y el C ($Z = -0{,}63$, p-value $= -0{,}623$).

4.2. Caso no paramétrico

Supongamos que queremos comprobar la eficacia de una SVM con kernel lineal, el C5.0 y una red neuronal en diferentes conjuntos de datos. Para ello, hemos entrenado cada uno de los clasificadores en cada uno de los conjuntos de datos obteniendo el índice Kappa en cada evaluación. Los resultados se pueden apreciar en la Tabla 6.

	SVMLinear	C50	NNET
hepatitis	0.4581748	0.5237671	0.4410976
iris	0.9400000	0.9200000	0.9500000
cox2	0.3568394	0.3924460	0.4277848
oil	0.9572989	0.9466109	0.9829787
dhfr	0.8421899	0.8098204	0.8439817
German	0.3579698	0.3786082	0.3884126
Seg	0.5674660	0.6643880	0.5781698
Breast	0.9202480	0.9146436	0.9312434
Pima	0.4769739	0.4573509	0.4857494
Sonar	0.4899545	0.6611616	0.6985279

Tabla 6. Kappa en cada uno de los conjuntos de datos.

Como en el caso anterior, tenemos que elegir entre un **test ANOVA de una vía para medidas pareadas** o su equivalente no paramétrico el **test de Friedman**, dependiendo si se cumplen los supuestos para el test ANOVA. Primero cargamos los datos:

```
> ejemplo5 <- read.csv("Ejemplo3TablaKappa.csv",row.names = 1)
```

Como en el caso anterior, resulta más cómodo transformar la tabla, con la información sobre las agrupaciones en una columna.

```
> df5.stack <- stack(ejemplo5)
> df5.stack$DataSet <- as.factor(rep(row.names(ejemplo5), times = 3))
> names(df5.stack) <- c("Kappa","Method", "DataSet")
```

Una vez dispuesta la tabla en el nuevo formato, podemos comprobar la normalidad en cada grupo:

```
> test.norm <- tapply(df5.stack$Kappa, df5.stack$Method,shapiro.test)
> test.norm

$C50

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.89558, p-value = 0.1958


$NNET

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.87459, p-value = 0.113


$SVMLinear

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.83722, p-value = 0.04086
```

Como podemos ver la muestra obtenida para el clasificador SVM no cumple la normalidad ($W_{SVM} = 0{,}84$, p-value $= 0{,}04$) con lo que debemos proceder a aplicar el test de Friedman.

```
> test.friedman <- friedman.test(Kappa ~ Method | DataSet, data = df5.stack )
> test.friedman

        Friedman rank sum test

data:  Kappa and Method and DataSet
Friedman chi-squared = 7.4, df = 2, p-value
= 0.02472
```

Como podemos observar, podemos afirmar con una del 95 % que existen diferencias significativas en los índices Kappas obtenidos por los tres clasificadores en los conjuntos de datos utilizados ($F = 7{,}4$, p-value $= 0{,}025$). Una vez confirmado la existencia de diferencias entre los clasificadores, tenemos que aplicar el **test post hoc de Nemenyi** para localizar dónde se encuentran dichas diferencias.

```
> library(PMCMR)
> nemenyi.test <- posthoc.friedman.nemenyi.test(Kappa ~ Method | DataSet,
                                                data = df5.stack)
> nemenyi.test

        Pairwise comparisons using Nemenyi multiple comparison test
             with q approximation for unreplicated blocked data

data:  Kappa and Method and DataSet

          C50   NNET
NNET      0.065 -
SVMLinear 0.973 0.037

P value adjustment method: none
```

Examinado los resultados, podemos afirmar, con una confianza del 95 %, que no existen diferencias entre C50 y el resto de clasificadores, pero si entre la red neuronal y la SVM.

A la vista de los resultados podemos afirmar, con el 95 % de confianza, que existe una diferencia significativa entre los clasificadores, tal y como muestra en test de Friedman ($F = 7{,}4$, p-value $= 0{,}025$). Además, aplicando el test post hoc de Nemenyi hemos podido comprobar que los clasificadores que muestran un comportamiento significativamente diferente son la SVM y la red neuronal (p-value <0.05) indicando que la red neuronal tienen un índice Kappa medio ligeramente superior. Los tests realizados no permiten determinar si existen diferencias significativas entre C50 y el resto de clasificadores.

Referencias

1. N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.