Ricardo Aler Mur

EVALUACIÓN DE TÉCNICAS DE APRENDIZAJE-2

COMPARACIÓN DE MODELOS

En esta clase se desarrolla de manera técnica una cuestión introducida en la clase anterior: la comparación de dos modelos.

- Además de estimar las prestaciones de un modelo, es interesante poder comparar de manera estadística las prestaciones de dos o más modelos. Se recalcará aquí, que aunque es importante que la prestación media de un modelo sea superior a la del otro, es también importante que la varianza no sea grande. De otra manera, no se podría afirmar rigurosamente que la diferencia en prestaciones no sea debida al azar.
- Se utiliza para ello el método descrito en el capítulo 5.5 del libro: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)
- Se utiliza una validación cruzada repetida múltiples veces y corregida para evitar infravalorar la varianza.
- Se explica la idea de comparación estadística de hipótesis, p-value e intérvalo de confianza.

Comparación de dos modelos

- Sean x_1, …, x_{10} los resultados de 10-fold crossvalidation del algoritmo A
- Sean y_1, …, y_{10} los resultados de 10-fold crossvalidation del algoritmo B

$$x_i = \frac{1}{10}\sum_{j=1}^{10} x_{ij} \qquad y_i = \frac{1}{10}\sum_{j=1}^{10} y_{ij}$$

- Es importante que las particiones usadas para A sean las mismas que las de B: siempre entrenamos y hacemos el test con el mismo fold. Disminuye la variabilidad y podemos hacer un test “pareado”

Comparación de dos modelos

- Según el teorema central del límite, la suma de muchas variables aleatorias sigue una distribución Normal (Gaussiana). Por tanto las dos siguientes medias siguen una Normal:

$$\bar{x}_T = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}\sum_{i=1}^{10}\frac{1}{10}\sum_{j=1}^{10} x_{ij} \qquad \bar{y}_T = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}\sum_{i=1}^{10}\frac{1}{10}\sum_{j=1}^{10} y_{ij}$$

- Si conocieramos la varianza de dichas Normales podríamos comprobar si las distribuciones solapan mucho (o si los intérvalos de confianza solapan mucho)

Comparación de dos modelos

- Desgraciadamente, aunque sabemos que $\bar{x}_T$ $\bar{y}_T$ siguen distribuciones normales, no conocemos la varianza, pero podemos estimarla así:

$$\sqrt{\sigma_x^2 / k}$$

- Donde σ_x^2 es la varianza de las k=10*10=100 muestras x_{ij}
- Tiene sentido que la varianza de la distribución disminuya a medida que aumentan las muestras (cuantas mas muestras, más precisa es la media)

Comparación de dos modelos

- Como hemos tenido que estimar la varianza, $\bar{x}_T \quad \bar{y}_T$ ya no siguen una normal, sino una t-student (parecida a la normal), con 10*10-1 grados de libertad
  - Usar un estimador para la varianza hace que haya más incertidumbre
- En lugar de usar x e y, usemos directamente la diferencia. Recordemos que los folds están pareados

$$d_{ij} = x_{ij} - y_{ij} \qquad \bar{d}_T = \frac{1}{10}\sum_{i=1}^{10}\frac{1}{10}\sum_{j=1}^{10} d_{ij}$$

Comparación de dos modelos

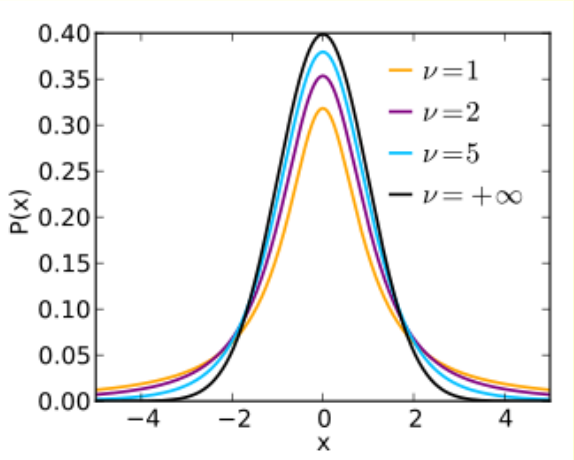
- Podemos decir que el siguiente estadístico t se distribuye según una t-student con 10*10-1 grados de libertad:

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$$

Comparación de dos modelos

- Para nuestros experimentos, d (y t) tendrá un valor concreto.
- Test de contraste de hipótesis: si suponemos que el d real es cero, ¿cuál es la probabilidad de que d tenga el valor que tiene? Si es baja, entonces lo mas probable es que no sea cero

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$$



Comparación de dos modelos

- Hasta ahora hemos supuesto que cada una de las 10 validaciones cruzadas se hace con un conjunto distinto (10 conjuntos independientes).
- Pero en la práctica, no se dispone de tantos datos y se reutiliza el mismo conjunto las 10 veces. Simplemente se desordena.
- Pero esto tiene el problema de que las 10 validaciones cruzadas ya no son independientes
- Estamos subestimando la varianza real. Por tanto muchas veces concluiremos que dos algoritmos son distintos cuando realmente no es así.
- En concreto, como el estimador de la varianza es inversamente proporcional a k, basta con hacer k enorme para que la varianza sea tan pequeña como queramos

Comparación de dos modelos

- Usaremos otro estimador corregido

$$k = 100,\ n_2/n_1 = 0.1/0.9,$$

$$t = \frac{\bar{d}}{\sqrt{\sigma_d^2/k}} \Rightarrow t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\sigma_d^2}}$$

Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

- Dos algoritmos A y B
- $\{x_{ij}, y_{ij} \mid i=1:r, j=1:k\}$ son el error de la i-esima validación cruzada en el fold j-esimo
- Las diferencias son: $d_{ij} = x_{ij} - y_{ij}$
  - Analogía: repetir r veces secuencias de k tiradas de una moneda (trucada o no) y contar las d_{ij} caras
- La media es: $\bar{d} = \frac{1}{r}\sum_{i=1}^{10}\frac{1}{k}\sum_{j=1}^{10} d_{ij}$
- Dos hipótesis:
  - H0: hipótesis nula. A=B o $\bar{d} = 0$
  - H1: A != B o $\bar{d} \neq 0$

Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

- Si H0 fuera cierta, los distintos valores $t = \frac{\bar{d}}{\sqrt{\sigma_d^2 / k}}$ para distintos conjuntos de datos disponibles T, se distribuirían según una t-student de media 0 con k-1 grados de libertad (en este caso k= r*k)
- Sabida la distribución, si efectuamos el experimento de extraer una muestra de datos y calcular r validacíones cruzadas de k folds, podemos calcular la probabilidad de que el resultado esté entre –t y +t (t es el estadístico que ha sido calculado con los datos que teníamos).
- Es decir, p = prob(-t <= x <= +t). Pvalue = 1 – p
- Si resulta que p es muy alta o que el pvalue es muy bajo, es que la probabilidad de observar t supuesta la hipótesis nula es muy baja, y por tanto hay que rechazar la hipótesis nula.
- Normalmente se considera que si pvalue <= 0.05, se rechaza H0

Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

- En R, probabilidad(x<t) = pt(t, df)
- Si t>0, entonces prob(x>t) = 1-pt(t,df) y como la t-student es simétrica, prob(-t<x) = 1-pt(t,df)
- Por tanto, supuesta H0, la probab
  - la probabilidad de que la observación sea mayor que |t| o menor que -|t| es:

    pvalue = 2(1-pt(t,df))
  - La probabilidad de que la observación esté en el intérvalo alrededor de 0, entre -|t| y +|t| es:

    1-pvalue = 1- 2(1-pt(t,df))

Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

- Ejemplo de comparación de C4.5 y knn en el dominio Iris con r==k==10
- KNN: error = 0.0470093 ± 0.02757903
- C4.5: error = 0.06272868 ± 0.02322911
- t = -1.250795

KNN VS. C4.5

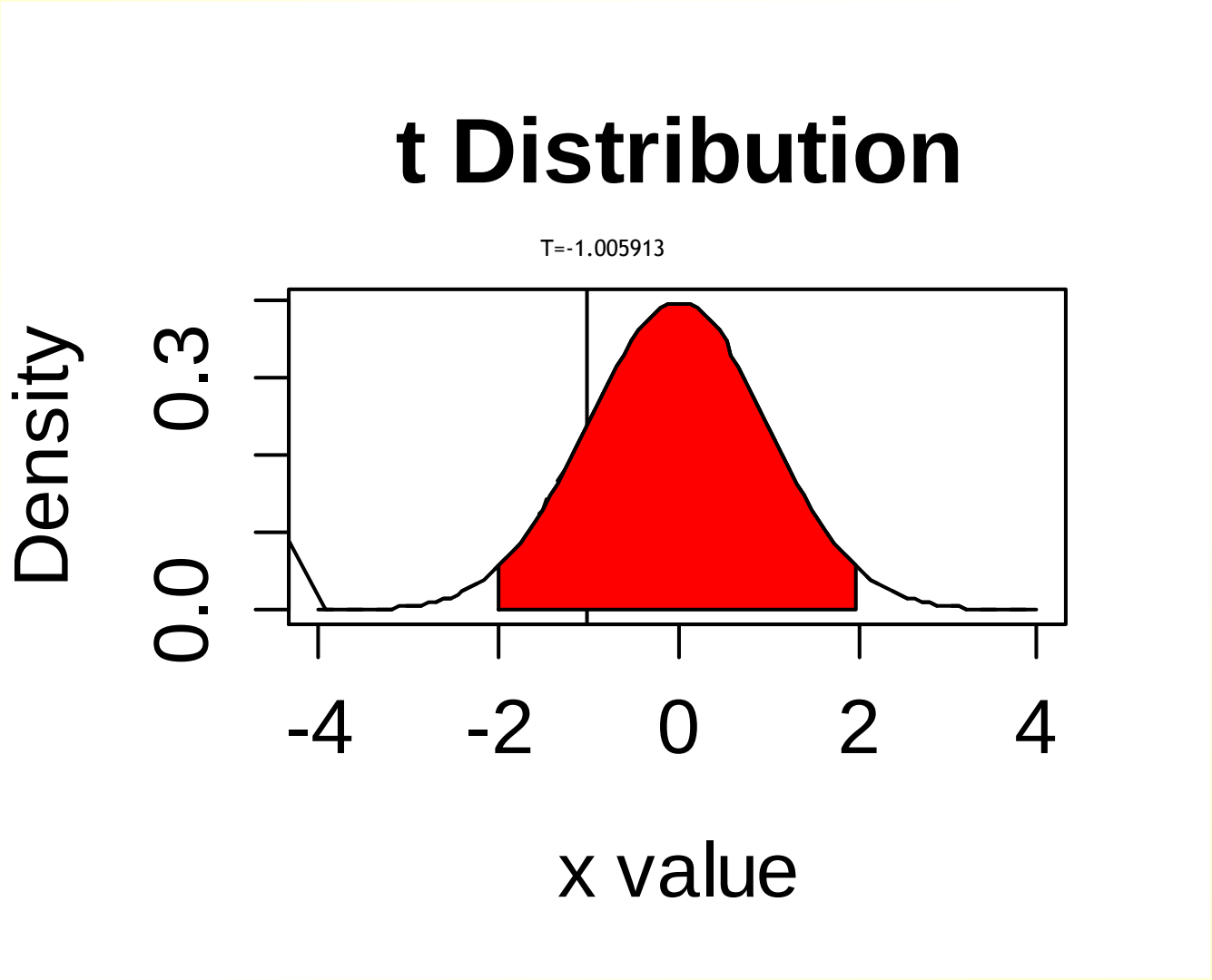
- t = -1.250795
- pvalue = 2*(1-pt(abs(t),df=k*r-1)) = 0.3169098
- El intérvalo que contiene el 95% de la probabilidad alrededor de 0 (es decir, deja fuera al 2.5%+2.5%=5% de la probabilidad) es:

  [qt(0.05/2,df), qt(1-0.05/2,df]

  [-1.984217 , 1.984217]

- Como t está dentro del intérvalo, no podemos rechazar la hipótesis nula

KNN VS. C4.5



Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

- Ejemplo de comparación de C4.5 y knn en el dominio Iris con r==k==10
- KNN: error = 0.04315691 ± 0.01321075
- ZeroR: error = 0.7831381 ± 0.01964745
- t = -24.06936
- Intérvalo 95%: [-1.984217, 1.984217]

KNN VS. ZeroR

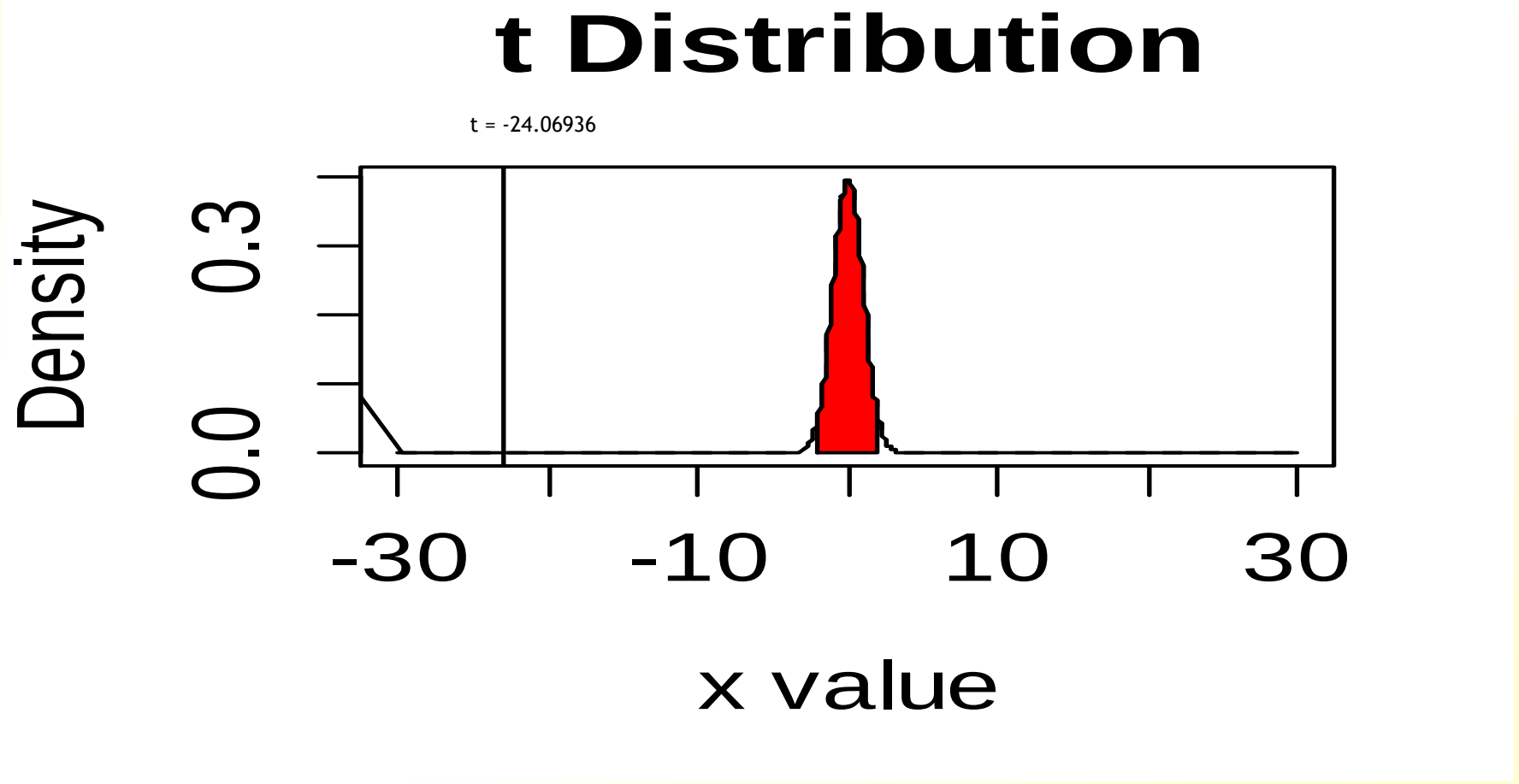
- t = -24.06936
- pvalue = 2*(1-pt(abs(t),df=k*r-1)) = 0
- El intérvalo que contiene el 95% es:

  [qt(0.05/2,df), qt(1-0.05/2,df]

  [-1.984217 , 1.984217]

❑ Como t está muy fuera del intérvalo, rechazamos la hipótesis nula

KNN VS. ZeroR



Pasos para hacer el contraste de hipótesis entre el algoritmo A y B

1. $\{x_{ij}, y_{ij} \mid i=1:r, j=1:k\}$ son el error de la i-esima validación cruzada en el fold j-esimo. $d_{ij} = x_{ij} - y_{ij}$
2. Calcular la media: $\bar{d}_T = \frac{1}{r}\sum_{i=1}^{10}\frac{1}{k}\sum_{j=1}^{10} d_{ij}$
3. Calcular el estadístico t: ($n_2=1/k$, $n_1=(k-1)/k$)

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{k}+\frac{n_2}{n_1}\right)\sigma_d^2}}$$

4. Pvalue = 1-prob(|x|<= t, t-student con df = r*k-1)

   pvalue = 1-2*(1-pt(abs(t),df=k*r-1))

5. Si pvalue <=0.05 entonces rechazar la hipótesis nula