Evaluación de Hipótesis

Aprendizaje Automático

Ingeniería Informática

Fernando Fernández Rebollo y Daniel Borrajo Millán

Grupo de Planificación y Aprendizaje (PLG)
Departamento de Informática
Escuela Politécnica Superior
Universidad Carlos III de Madrid

27 de febrero de 2009

En Esta Sección:

Introducción

- Evaluación de hipótesis: estimar la calidad de una hipótesis aprendida de la forma más precisa posible
- Motivación:
  - Decidir si utilizar la hipótesis o no:
    ¿Debemos utilizar una hipótesis o clasificador generado a partir de 40 instanias de entrenamiento??
  - La evaluación de hipótesis es un elemento fundamental en muchos algoritmos de aprendizaje:
    ¿Es conveniente realizar una poda en el árbol de decisión o no?
- Dificultades en la evaluación de hipótesis:
  - Bias en la estimación: estimaciones realizadas sobre el conjunto de entrenamiento son muy pobres para predecir el éxito en datos futuros
  - Varianza en la estimación: estimaciones sobre conjuntos de test también generan diferencias sobre los valores reales. Diferentes conjuntos producen valores distintos

Evaluación de Hipótesis

- Dados:
  - una hipótesis h,
  - un conjunto de datos que contiene n ejemplos generados aleatoriamente de acuerdo a una distribución $\mathcal{D}$,
- ¿Cuál es la mejor estimación del éxito de h sobre futuras instancias que también siguen la distribución $\mathcal{D}$?
- ¿Cuál es el error más probable en esta estimación?

Tipos de Error

- Error en Ejemplos de una hipótesis h con respecto a una función objetivo f y un conjunto de datos S:

$$error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x)) \tag{45}$$

donde n es el tamaño del conjunto S, y $\delta(f(x), h(x)) = 1$ si $h(x) = f(x)$, y 0 en cualquier otro caso

- Error Real de una hipótesis h con respecto a una función objetivo f y una distribución $\mathcal{D}$ es la probabilidad de que h clasifique de forma incorrecta una instancia que sigue la distribución $\mathcal{D}$:

$$error_D(h) = Prob_{x \in \mathcal{D}}[f(x) \neq h(x)] \tag{46}$$

Estimación del Error

- ¿Cuál es la desviación del error en ejemplos con respecto al error real, en función del tamaño del conjunto de entrenamiento?
- Generamos k tests con una hipótesis h. Cada test se realiza con un conjunto de datos S_i de tamaño n. Esto nos genera una variable aleatoria compuesta por los siguientes valores: $error_{S_1}(h)$, $error_{S_2}(h)$, ..., $error_{S_k}(h)$.
- Generamos un histograma que muestra la frecuencia con que aparece cada posible valor de error
- Ese histograma coincide con una distribución binomial.

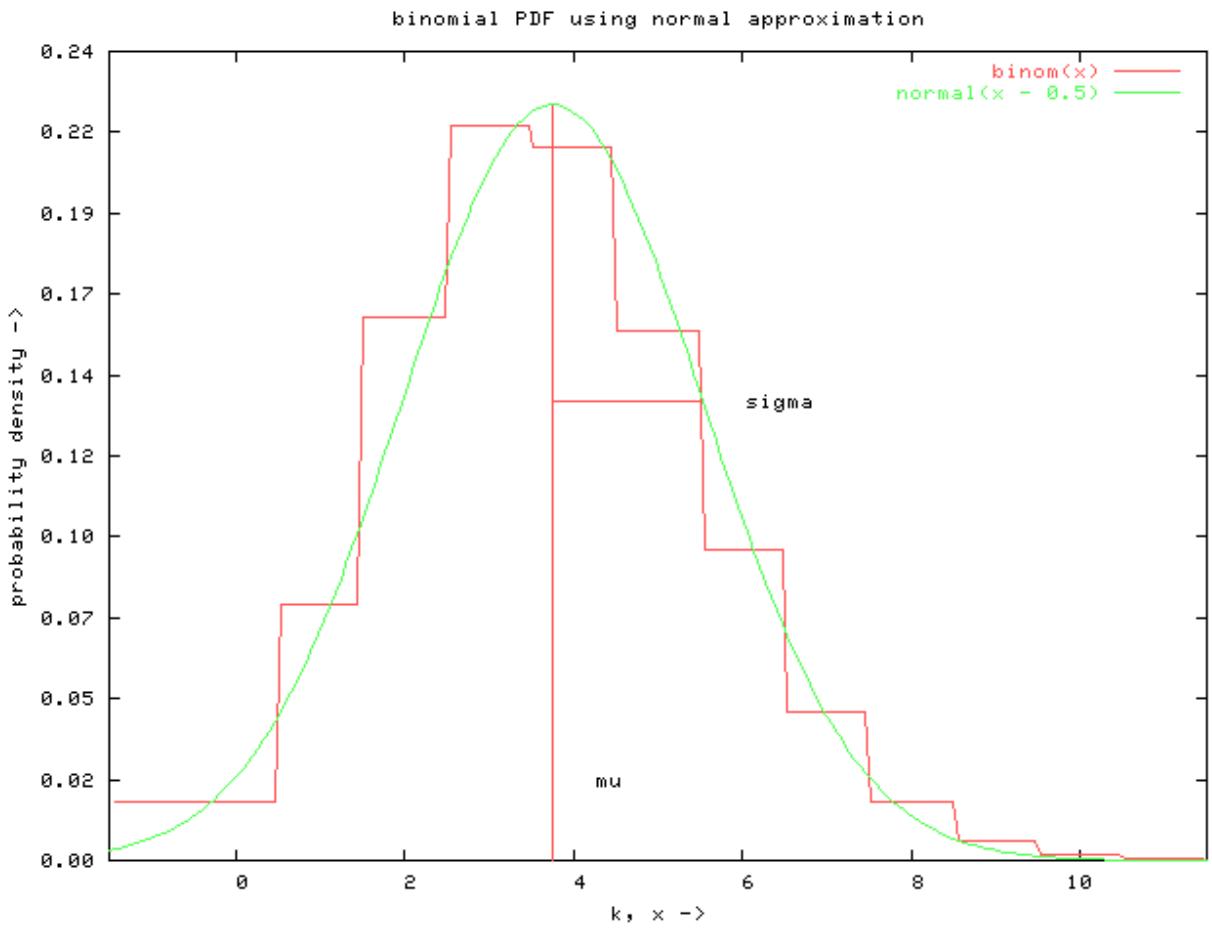
Distribución binomial

- Proporciona la probabilidad de observar r caras en una población de n lanzamientos de moneda independientes, cuando la probabilidad de obtener cara es p.
- Es decir, proporciona la probabilidad de obtener r fallos cuando se realiza un test sobre n ejemplos, y cuando la probabilidad real de fallos es p.

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \tag{47}$$

- Esperanza: $E[X] = np$
- Varianza: $Var(X) = np(1-p)$
- Desviación standar: $\sigma_X = \sqrt{np(1-p)}$

Binomial ($n = 25, p = 0{,}15$)



Error y Distribución Binomial

- La variable aleatoria $error_S(h)$ obedece una distribución binomial:
  - $error_S(h) = \frac{r}{n}$
  - $error_{\mathcal{D}}(h) = p$
- $error_S(h)$ tiende a $error_{\mathcal{D}}(h)$, dado que la esperanza de r es np
- La desviación estándar de $error_S(h)$ es:

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \tag{48}$$

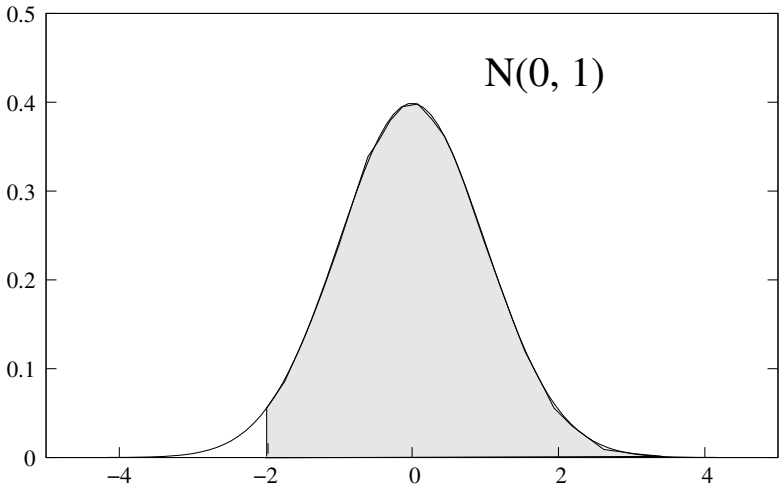
- Cuanto mayor es n, menor es la desviación estándar

Intervalos de Confianza

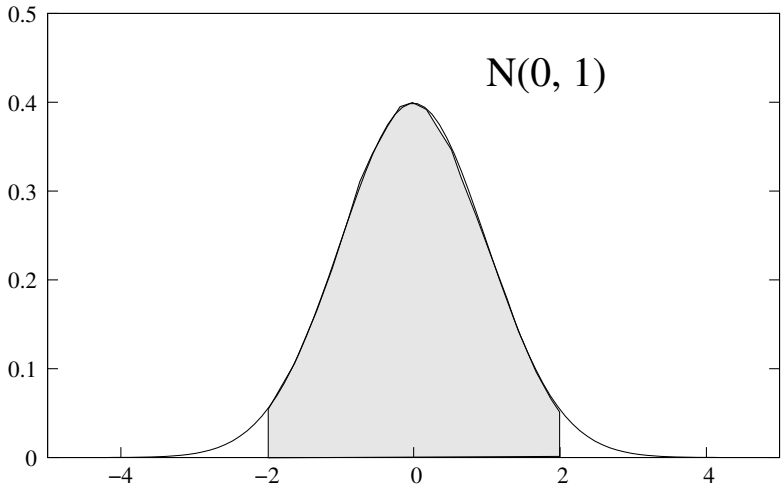
- Un intervalo de confianza del $N\,\%$ de un parámetro p es un intervalo que se espera que contenga p con una probabilidad del $N\,\%$
- Cálculo del intervalo de confianza de $error_D(h)$:
  - $error_S(h)$ sigue una distribución binomial
  - La media de esta distribución es $error_D(h)$
  - La desviación estándar es $\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$
  - Encontrar el intervalo centrado en la media que es suficientemente grande como para contener el $N\,\%$ de la probabilidad de la distribución.
- Cálculo complejo para una distribución binomial
- Pero bastante simple para una normal, que aproxima bastante bien a una binomial de igual media y varianza, cuando n es suficientemente grande

Intervalos de Confianza de una o dos Colas

- Área limitada por la distribución de probabilidad:
  - Una cola: $\int_{-\infty}^{b} N(\mu, \sigma)$, $\int_{a}^{\infty} N(\mu, \sigma)$
  - Dos colas: $\int_{a}^{b} N(\mu, \sigma)$



(a) Una cola



(b) Dos colas

Estimación de un Intervalo de Confianza

- Para resumir:
  - Si una variable aleatoria Y obedece una distribución normal con media μ y desviación estándar σ, entonces la variable aleatoria observada y de Y, caerá en un $N\,\%$ de las veces en el intervalo $\mu \pm z_N \sigma$
- En el caso de estimar el error sabemos que:
  - $error_S(h)$ sigue una distribución binomial con media $error_{\mathcal{D}}(h)$ y desviación estándar $\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$
  - Cuando n crece, la distribución binomial tiende a una distribución normal ($n \geq 30$ o $np(1-p) \geq 5$)
  - Podemos estimar el intervalo de confianza de una distribución normal

Estimación de un Intervalo de Confianza

- Por tanto:
  - Con un $N\,\%$ de probabilidad, $error_{\mathcal{D}}(h)$ está en el intervalo:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

- Donde (para intervalos de confianza de dos colas):

Intervalo Confianza $N\,\%$	50 %	68 %	80 %	90 %	95 %	98 %	99 %
Constante z_N:	0,67	1,00	1,28	1,64	1,96	2,33	2,58

Diferencias de Error entre dos Hipótesis

- Diferencia real:

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

- Diferencia dados unos datos:

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

- Distribución que gobierna $\hat{d}$:
  - Asumimos que $error_{S_1}(h_1)$ y $error_{S_2}(h_2)$ siguen distribuciones normales
  - Por tanto, $\hat{d}$ sigue una distribución normal
  - Media: $\mu_{\hat{d}} = d$
  - Varianza: $\sigma^2_{\hat{d}} \approx \frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}$

Intervalos de Confianza en la Diferencia de Error entre dos Hipótesis

- La diferencia entre los errores de dos hipótesis caen con un $N\,\%$ de probabilidad en el intervalo:

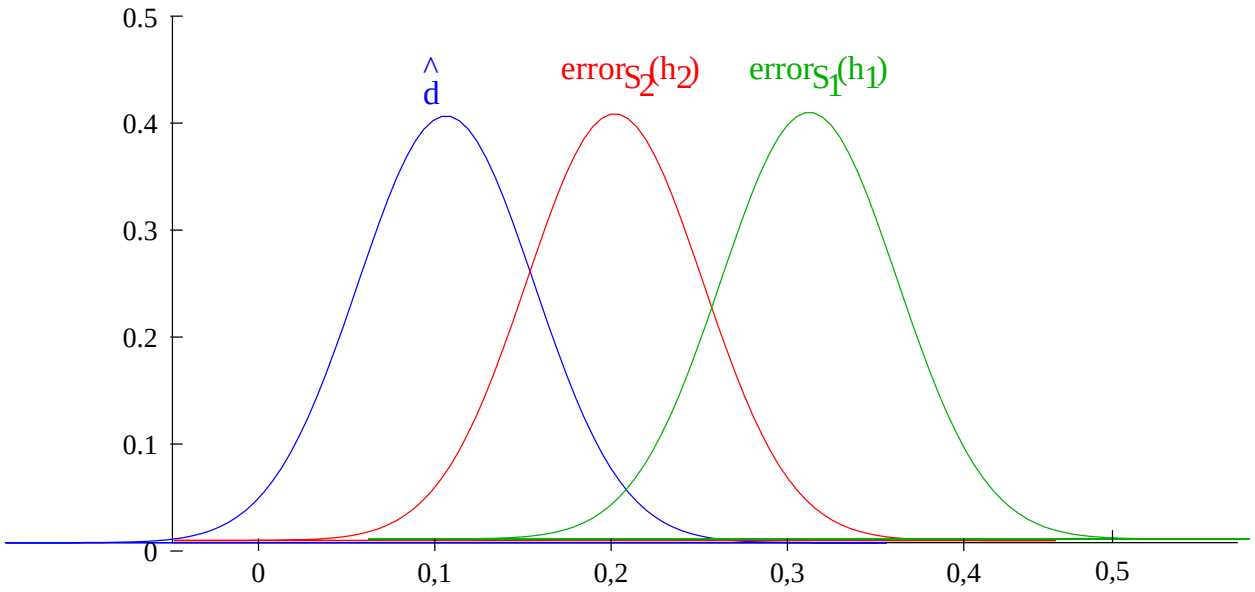
$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

- Tiene sentido utilizar para la estimación del error en ambas hipótesis el mismo conjunto de datos:

$$\hat{d} \equiv error_S(h_1) - error_S(h_2)$$

Comparación de Hipótesis

- ¿Cuál es la probabilidad de que $error_{\mathcal{D}}(h_1) \geq error_{\mathcal{D}}(h_2)$?
- Ejemplo:
  - $error_{S_1}(h_1) = 0{,}30$, $error_{S_2}(h_2) = 0{,}20$, $n_1 = n_2 = 100$
  - $\hat{d} = 0{,}10$

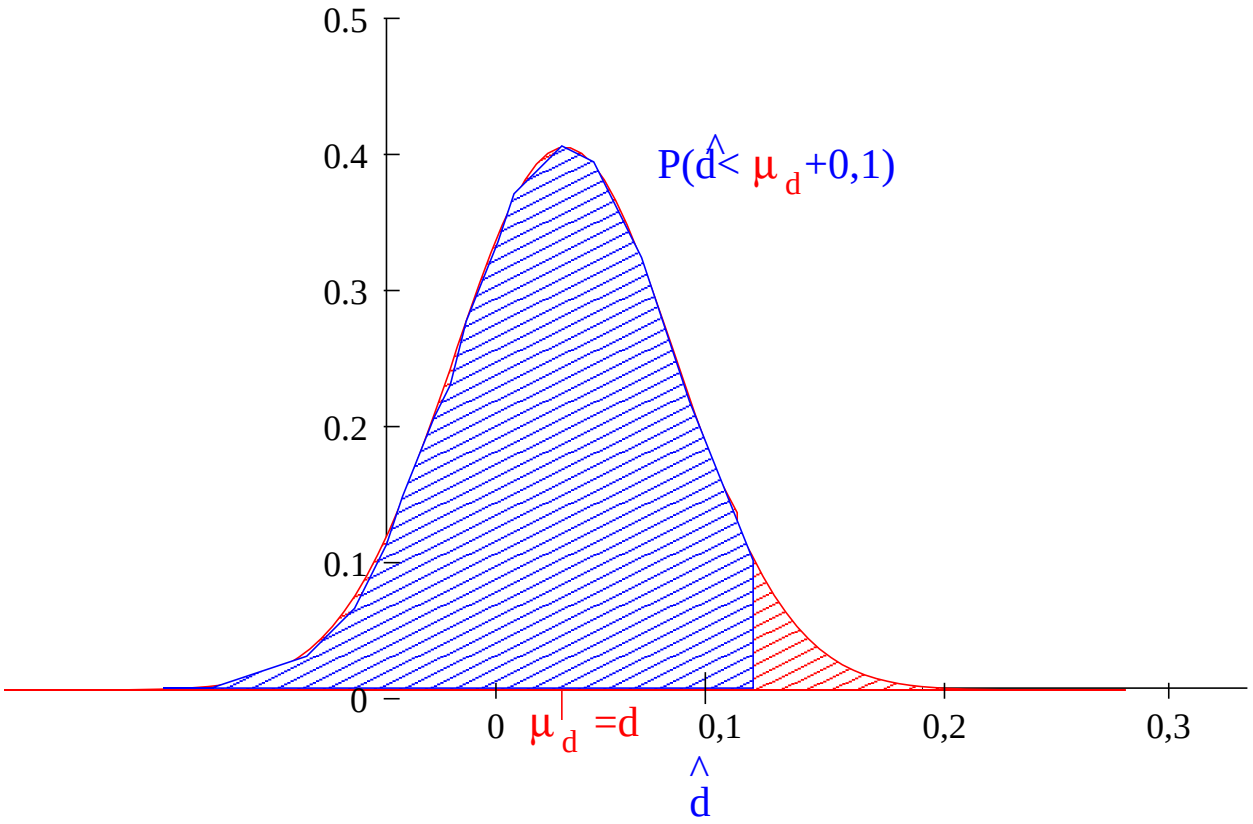


Comparación de Hipótesis

- ¿cuál es la probabilidad de que $error_{\mathcal{D}}(h_1) \geq error_{\mathcal{D}}(h_2)$ habiendo observado $\hat{d} = 0{,}10$?
- ¿cuál es la probabilidad de que $d > 0$ ($Prob(d > 0)$) habiendo observado $\hat{d} = 0{,}10$?
- ¿cuál es la probabilidad de que mi observación $\hat{d}$ no haya sobrestimado a d en más de 0.1? ¿$Prob(\hat{d} < d + 0{,}1)$?
- Dado que d es la media de la distribución que gobierna $\hat{d}$, ¿$Prob(\hat{d} < \mu_{\hat{d}} + 0{,}1)$?

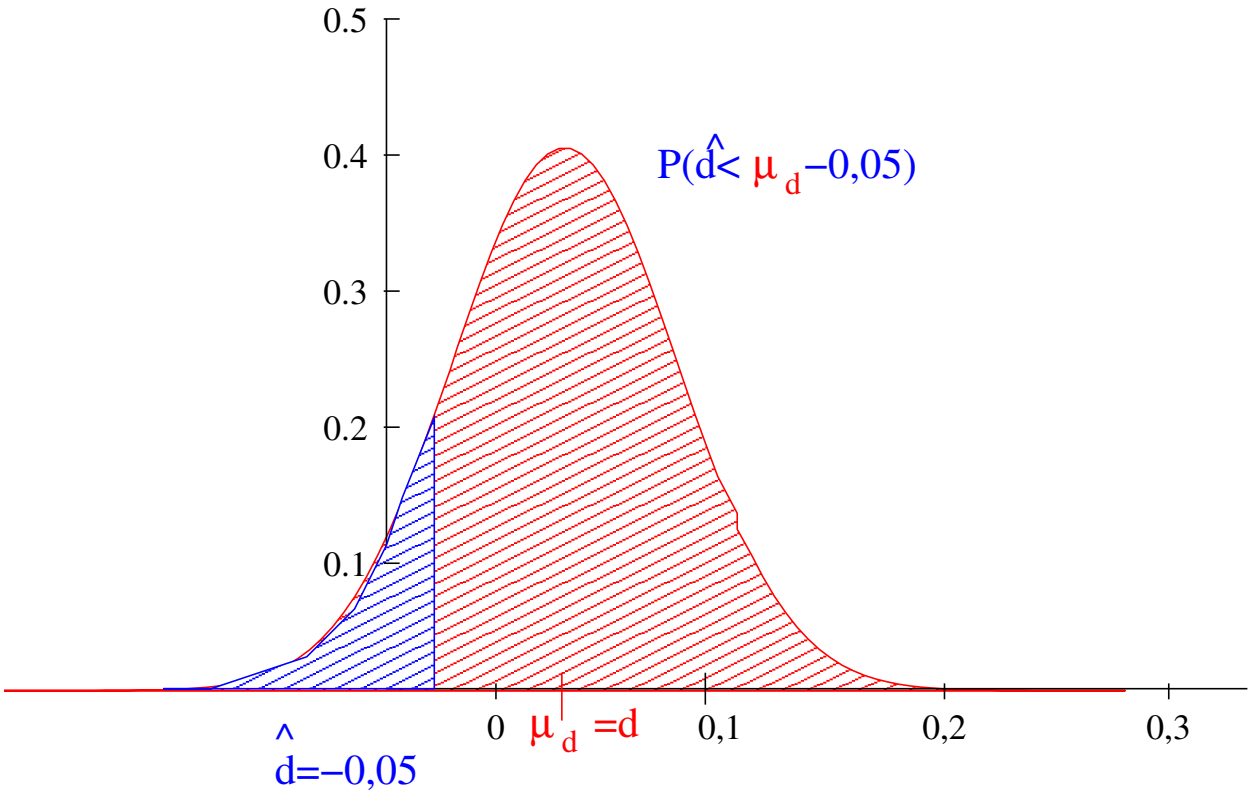
Ejemplo Comparación de Hipótesis

- $¿Prob(\hat{d} < \mu_{\hat{d}} + 0{,}1)?$



Ejemplo Comparación de Hipótesis

- ¿Y si mi observación de $\hat{d}$ fuese $-0{,}05$?
- ¿$Prob(\hat{d} < \mu_{\hat{d}} - 0{,}05)$?



Ejemplo Comparación de Hipótesis

- Volviendo al caso inicial: $\hat{d} = 0{,}1$, ¿$Prob(\hat{d} < \mu_{\hat{d}} + 0{,}1)$?
- Reescribimos la ecuación en función de cuánto me puedo desviar de la media: $\hat{d} < \mu_{\hat{d}} + z_N \sigma_{\hat{d}}$
- Calculamos $\sigma_{\hat{d}}$
  - $\sigma^2_{\hat{d}} \approx \frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2} = \frac{0{,}3(1-0{,}3)}{100} + \frac{0{,}1(1-0{,}1)}{100} = \frac{0{,}21}{100} + \frac{0{,}09}{100}$
  - $\sigma_{\hat{d}} = \sqrt{\frac{0{,}3}{100}} = 0{,}055$
- Por tanto: $z_N \sigma_{\hat{d}} = 0{,}1 \rightarrow z_N = \frac{0{,}1}{0{,}055} = 1{,}82$
- En la tabla de intervalos de confianza de dos colas, corresponde con almenos un 90 %, por lo que dado que en este caso sólo tenemos una cola, corresponde con más de un 95 %

Ejemplo Comparación de Hipótesis

- Conclusión:
  - Podemos afirmar que $error_{\mathcal{D}}(h_1) \geq error_{\mathcal{D}}(h_2)$ con una confianza de almenos el 95 %
  - Podemos rechazar la hipótesis de que $error_{\mathcal{D}}(h_1) \geq error_{\mathcal{D}}(h_2)$ con una probabilidad de algo menos que 0,05

Algoritmo de Validación Cruzada para t-test

1. Partir el conjunto de datos disponible, D_0 en K conjuntos disjuntos, $T_1, T_2, \ldots, T_k$ de igual tamaño, donde este tamaño debe ser mayor que 30.
2. Desde $i = 1$ hasta k hacer
   - Utilizar T_i como conjunto de test, y el resto de los datos como conjunto de entrenamiento, S_i:
     - $S_i \leftarrow \{D_0 - T_i\}$
     - $h_A \leftarrow L_A(S_i)$
     - $h_B \leftarrow L_B(S_i)$
     - $\delta_i = error_{T_i}(h_A) - error_{T_i}(h_B)$
3. Devolver $\bar{\delta}$ donde:

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$

t-test pareado

- Objetivo: Estimar $E_{S \in D_0}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$
- Con un intervalo de confianza del $N\%$, el valor anterior está en el rango:

$$\bar{\delta} \pm t_{N,k-1} S_{\bar{\delta}}$$

Donde:

- $t_{N,k-1}$ es una constante que juega un rol parecido a z_N para una distribución t
- $S_{\bar{\delta}}$ es una estimación de la desviación estándar de la distribución t que gobierna $\bar{\delta}$:

$$S_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

Valores de $t_{N,\upsilon}$

	Intervalo de Confianza N			
	90 %	95 %	98 %	99 %
$\upsilon = 2$	2,92	4,3	6,96	9,92
$\upsilon = 5$	2,02	2,57	3,36	4,03
$\upsilon = 10$	1,81	2,23	2,76	3,17
$\upsilon = 20$	1,72	2,09	2,53	2,84
$\upsilon = 30$	1,70	2,04	2,46	2,75
$\upsilon = 120$	1,66	1,98	2,36	2,62
$\upsilon = \infty$	1,66	1,96	2,33	2,58

Valores de $t_{N,\upsilon}$ para intervalos de confianza de dos colas.
Cuando $t_{N,\upsilon} \rightarrow \infty$, $t_{N,\upsilon}$ tiende a z_N.

Resumen

- La teoría estadística proporciona la base para estimar el error real de una hipótesis h, $error_{\mathcal{D}}(h)$, basado en su error observado $error_S(h)$
- Intervalos de confianza basados en las distribuciones que gobiernan $error_S(h)$, y por tanto de S
- Sesgos de estimación y sesgos por la varianza
- Comparación de algoritmos:
  - t-test
  - Basado en simplificaciones: distribución normal aproxima una binomial, distribución t aproxima una normal, etc.

Bibliografía

- Machine Learning, Tom Mitchell. Capítulo 5