Laboratório de Inteligência Computacional (LABIC)

http://labic.icmc.sc.usp.br/

Avaliação da Precisão de Hipóteses

Maria Carolina Monard & Gustavo Batista

Universidade de São Paulo
Campus de São Carlos - SP, Brasil
email: {mcmonard, gbatista}@icmc.sc.usp.br

Apresentação baseada na dissertação de mestrado de Paulo S. Horst

Considerações Iniciais

Existem vários métodos estatísticos que podem ser aplicados para estimar a precisão de hipóteses. Três questões importantes a serem respondidas são [Mitchell, 1997]:

1. Quando se conhece a precisão de uma hipótese sobre uma amostra limitada de dados, deseja-se saber *o quanto ela pode servir para estimar a precisão sobre novos exemplos?*
2. Quando uma hipótese tem melhor desempenho que outra sobre uma determinada amostra de exemplos, *qual a probabilidade que esta hipótese seja em geral mais precisa?*
3. Quando os dados disponíveis são limitados, *qual a melhor maneira de usá-los tanto para aprender uma hipótese quanto para estimar sua precisão?*

- Estimar a precisão de uma hipótese quando se trabalha com uma amostra grande de dados não apresenta maiores problemas.
- No entanto, quando se trabalha com uma amostra limitada de dados, uma das dificuldades é que estas podem não representar bem a distribuição geral dos dados.
- Se este for o caso, a estimativa da precisão de uma hipótese utilizando essa amostra pode conduzir a falsos resultados.
- O uso de métodos estatísticos junto com suposições referentes a distribuição dos dados permitem limitar a diferença que existe entre a precisão que é observada sobre a amostra dos dados disponíveis e a precisão verdadeira sobre a distribuição total dos dados.

Existem diversas dificuldades básicas para aprender uma hipótese e estimar sua precisão futura quando se trabalha com uma amostra limitada de dados. Essas dificuldades estão relacionadas ao *bias na estimativa* e a *variância na estimativa*.

- Quanto ao *bias*, a precisão observada de uma hipótese gerada com exemplos de treinamento é, frequentemente, o pior estimador da precisão sobre novos exemplos.
- Uma estimativa geralmente será *unbiased* quando as hipóteses são testadas sobre conjunto de exemplos de teste escolhidos independentemente dos exemplos de treinamento e da hipótese.
- A *variância* na estimativa pode aparecer mesmo quando a precisão de uma hipótese é medida sobre um conjunto de teste *unbiased*. Esta precisão pode ainda variar em relação à precisão verdadeira em função do tamanho desse conjunto de teste. Quanto menor for o conjunto de exemplos de treinamento, maior será a variância.

Estimativa da Precisão de Hipóteses

- Ao analisar uma hipótese aprendida h sobre uma amostra de dados S que contém n exemplos extraídos aleatoriamente de acordo com a distribuição D, surgem duas questões de interesse:
  1. *Qual é a melhor estimativa da precisão de h sobre futuras instâncias extraídas da mesma distribuição D?*
  2. *Qual é o erro provável na estimativa dessa precisão?*
- Estas questões podem ser respondidas e para isso é necessário distinguir entre duas noções de precisão ou, equivalentemente, de erro. Elas são o *erro amostral* e o *erro verdadeiro*.





- O *erro amostral* é a taxa de erro de uma hipótese h sobre uma amostra de dados disponível S de instâncias extraídas da população X, ou seja, é a fração de S que h classifica erroneamente.

  Definição 2.1 o *erro amostral* da hipótese h, denotado por $erro_S(h)$, em relação a função meta f e amostra de dados S é

$$erro_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x)) \qquad (1)$$

  onde n = número de exemplos em S e

$$\delta(f(x), h(x)) = \begin{cases} 1 & \text{se } f(x) \neq h(x) \\ 0 & \text{caso contrário} \end{cases}$$

- Por outro lado, o *erro verdadeiro* é a taxa de erro de uma hipótese h sobre toda a distribuição desconhecida D de exemplos, isto é, a probabilidade que h classificará erroneamente uma única instância extraída aleatoriamente da distribuição D.

  Definição 2.2 o *erro verdadeiro* da hipótese h, denotado por $erro_D(h)$ em relação a função meta f e distribuição D é

$$erro_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)] \tag{2}$$

  onde $\Pr_{x \in D}$ denota justamente que a probabilidade é considerada sobre a distribuição D.

- Usualmente, procura-se pelo erro verdadeiro $erro_D(h)$ da hipótese, pois este é o erro esperado quando se aplica a hipótese a novos exemplos.
- No entanto, o que pode ser medido é apenas o erro amostral $erro_S(h)$. Em virtude disso, é preciso saber

  Quão boa uma estimativa do $erro_D(h)$ é fornecida pelo $erro_S(h)$?

No caso de classificação, isto é, a hipótese h assume somente valores discretos, considerando que

1. a amostra S contém n instâncias retiradas uma a uma de acordo com a distribuição D, independente de h;
2. $n \geq 30$;
3. a hipótese h classifica erroneamente r desses n exemplos (ou instâncias), ou seja, $erro_S(h) = \frac{r}{n}$;

então, se não existir outra informação, com base na teoria estatística é possível afirmar que o valor mais provável de $erro_D(h)$ é $erro_S(h)$ e, com aproximadamente 95% de probabilidade, o erro verdadeiro $erro_D(h)$ está no intervalo

$$erro_S(h) \pm 1.96\sqrt{\frac{erro_S(h)(1 - erro_S(h))}{n}} \qquad (3)$$

- Por exemplo, se a amostra S contém $n = 50$ exemplos e a hipótese h classifica erroneamente $r = 10$ desses exemplos, então $erro_S(h) = \frac{10}{50} = 0.2$.
- Se o experimento fosse repetido várias vezes retirando outras amostras S_1, S_2, etc., de 50 exemplos, espera-se que os erros dessas amostras $erro_{S_1}$, $erro_{S_2}$, etc., apresentem valores ligeiramente diferentes que $erro_S(h)$.
- Mas será encontrado que para 95% desses experimentos o intervalo calculado contém o erro verdadeiro.
- Por isso, esse intervalo é denominado intervalo de confiança de 95% da estimativa para $erro_D(h)$.

No exemplo considerado esse intervalo é dado por

$$0.2 \pm 1.96\sqrt{\frac{0.2 \cdot 0.8}{50}} = 0.2 \pm 0.110 \qquad (4)$$

- O intervalo de confiança de 95% definido pela Equação 3, Slide 8, pode ser generalizado para qualquer intervalo de confiança utilizando o valor da constante Z_N correspondente ao nível de confiança $N\%$ desejado, como mostra a Tabela 1

Nível de confiança $N\%$:	50%	68%	80%	90%	95%	98%	99%
Constante Z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Tabela 1: Valores de Z_N para Intervalos de Confiança $N\%$ *Two-Sided*

- Assim, a expressão geral para $N\%$ intervalo de confiança da estimativa para $erro_D(h)$ é dado por

$$erro_S(h) \pm Z_N \sqrt{\frac{erro_S(h)(1 - erro_S(h))}{n}} \qquad (5)$$

- Para o exemplo considerado, a Tabela 2 mostra os diferentes intervalos de confiança calculados com a Equação 5.
- Obviamente, intervalos com maior nível de confiança são maiores pois está se incrementando a probabilidade de $erro_D(h)$ estar nesse intervalo.

$N\%$	$Z_N\sqrt{\frac{erro_S(h)(1-erro_S(h))}{n}}$	Intervalo de Confiança
50%	0.037	[0.163 , 0.237]
68%	0.056	[0.144 , 0.256]
80%	0.072	[0.128 , 0.272]
90%	0.092	[0.108 , 0.292]
95%	0.110	[0.090 , 0.310]
98%	0.131	[0.069 , 0.331]
99%	0.145	[0.055 , 0.345]

Tabela 2: Intervalos de Confiança para $n = 50$ e $r = 10$

- Repetindo o experimento k vezes, isto é, retirando k amostras S_1,S_2,...,S_k de tamanho n, o erro dessas amostras $erro_{S_i}(h)$, $i = 1, ..., k$ é uma variável aleatória binomial.
- Assim, a probabilidade da hipótese h classificar erroneamente r desses n exemplos está dada por

$$\Pr[\#erro_S(h) = r] = \binom{n}{r} p^r (1-p)^{n-r}, \ r = 0, ..., n. \qquad (6)$$

- A probabilidade p, desconhecida, representa justamente o erro verdadeiro $erro_D(h)$ e a idéia é estimar p testando a hipótese h nas amostras S_i.
- Em outras palavras, sabendo que a variável aleatória $erro_S(h)$ obedece a distribuição Binomial, qual a provável diferença entre o erro amostral $erro_S(h) = \frac{r}{n}$ e o erro verdadeiro $erro_D(h) = p$?
- Em termos estatísticos $erro_S(h)$ é um estimador do erro verdadeiro $erro_D(h)$.

Estimadores, Bias e Variância

- Em geral, um estimador é qualquer variável aleatória usada para estimar algum parâmetro da população da qual a amostra é extraída.
- Uma questão óbvia referente a qualquer estimador é tentar saber se na média ele fornece uma estimativa correta.
- Define-se o *bias de estimação* como sendo a diferença entre o valor esperado do estimador e o valor verdadeiro do parâmetro.

Definição 3.1 o *bias de estimação* de um estimador Y para um parâmetro arbitrário p é $E[Y] - p$.

Se o *bias* de estimação é zero, diz-se que Y é um estimador $unbiased$ para p.

- Deve-se notar que isso ocorrerá se a média de muitos valores aleatórios de Y gerados por experimentos aleatórios repetidos (ou seja, $E[Y]$) convergem para p.
- Pode ser observado que $erro_S(h)$ é um estimador *unbiased* para $erro_D(h)$ tendo em vista que para uma distribuição Binomial o valor esperado de r é igual a np. Assim, dado que n é constante, o valor esperado de $\frac{r}{n}$ é p.

- Quando testa-se uma hipótese utilizando os exemplos de treinamento, tem-se uma estimativa otimistamente *biased* de erro da hipótese.
- Portanto, para obter uma estimativa *unbiased* de $erro_D(h)$, a hipótese h e a amostra S devem ser escolhidas independentemente.
- Outra propriedade importante de qualquer estimador é sua variância. Tendo alternativas de estimadores *unbiased* faz sentido escolher aquele que possui a menor variância, pois pela definição de variância, essa escolha produzirá o menor erro quadrado esperado entre o valor estimado e o valor verdadeiro do parâmetro.

- Em geral, dado r erros em uma amostra de n exemplos de teste extraídos independentemente, o desvio padrão para $erro_S(h)$ é dado por

$$\sigma_{erro_S(h)} = \frac{\sigma_r}{n} = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{7}$$

  o qual pode ser aproximado, substituindo p por $\frac{r}{n} = erro_S(h)$

$$\sigma_{erro_S(h)} \approx \sqrt{\frac{erro_S(h)(1-erro_S(h))}{n}} \tag{8}$$

Intervalos de Confiança

- Uma maneira comum de descrever a incerteza associada com uma estimativa é fornecer um intervalo dentro do qual é esperado que o valor verdadeiro esteja, juntamente com a sua probabilidade de estar nesse intervalo.
- Tais estimativas são chamadas de *estimativas de intervalo de confiança*.

Definição 4.1 um *intervalo de confiança* de $N\%$ para algum parâmetro p é o intervalo onde se espera conter p com probabilidade de $N\%$.

A primeira pergunta que surge é

Como derivar intervalos de confiança para $erro_D(h)$?

- Derivar o intervalo de confiança para $erro_D(h)$ é simples, já que é conhecido que a distribuição de probabilidade Binomial governa o estimador $erro_S(h)$.
- A média dessa distribuição é $erro_D(h)$ e o desvio padrão é dado pela Equação 8, Slide 17.

- Portanto, para derivar um intevalo de confiança de 95%, é somente preciso encontrar o intervalo centrado em torno do valor médio $erro_D(h)$, que seja amplo o suficiente para conter 95% da probabilidade total dessa distribuição.
- Isso fornece um intervalo contendo $erro_D(h)$, dentro do qual $erro_S(h)$ deve cair 95% das vezes.
- Equivalentemente, fornece o tamanho do intervalo que contém $erro_S(h)$ dentro do qual $erro_D(h)$ deve cair 95% das vezes.

Outra pergunta que surge é

Para um dado N, como encontrar o tamanho do intervalo que contém $N\%$ da massa de probabilidades?

- No caso da distribuição Binomial, este cálculo é trabalhoso.
- Entretanto, na maioria dos casos a distribuição Binomial é bem aproximada pela distribuição Normal $N(np, \sqrt{np(1-p)})$
- A distribuição Normal padronizada, $N(0, 1)$, encontra-se tabelada especificando o tamanho do intervalo, à volta da média, que contém $N\%$ da distribuição acumulada — Tabela 1, Slide 10.

Esta é justamente a informação necessária para realizar o cálculo do $N\%$ intervalo de confiança como mostrado nos slides anteriores, pois dada uma variável aleatória X que obedece a distribuição Normal $N(\mu, \sigma)$ então:

- o valor aleatório medido y de Y estará $N\%$ das vezes no intervalo

$$\mu \pm Z_N \sigma \tag{9}$$

- a média μ estará $N\%$ das vezes no intervalo

$$y \pm Z_N \sigma \tag{10}$$

- Utilizando na Equação 10, o valor da média e desvio padrão — Equação 8, Slide 17 — de $erro_S(h)$ obtém-se a expressão geral para $N\%$ intervalo de confiança da estimativa para $erro_D(h)$,

$$erro_S(h) \pm Z_N \sqrt{\frac{erro_S(h)(1 - erro_S(h))}{n}}$$

que é a Equação 5, Slide 10.

- É muito importante observar que esta expressão é válida somente no caso de classificação, ou seja, exemplos com classe categórica. Além disso, duas aproximações foram realizadas para obtê-la:
  1. o erro verdadeiro $erro_D(h) = p$ na Equação 7, Slide 17, foi aproximado por $erro_S(h)$ para obter o valor aproximado do desvio padrão σ do $erro_S(h)$ na Equação 8, Slide 17.
  2. a distribuição Binomial foi aproximada pela distribuição Normal.

- Em termos estatísticos estas duas aproximações são consideradas muito boas sempre que $n \geq 30$ ou quando $np(1-p) \geq 5$.
- Para valores de n menores deve ser utilizada a própria distribuição Binomial.
- O intervalo de confiança visto anteriormente está limitado *two-sided*, ou seja, ele limita a quantidade estimada tanto no seu limite superior quanto inferior.
- Em alguns casos o interesse é somente em limites *one-sided*, por exemplo, quando é necessário responder questões tais como

  Qual a probabilidade que $erro_D(h)$ *seja no máximo* U *(limite superior do intervalo de confiança)?*

- Para encontrar tais limites de erro *one-sided* existe uma modificação simples do procedimento descrito.
- Isso vem do fato de que a distribuição Normal é simétrica em torno da sua média.
- Em função disto, qualquer intervalo de confiança *two-sided* baseado na distribuição Normal pode ser convertido para um intervalo de confiança *one-sided* com duas vezes a confiança.

Uma Metodologia para Derivar Intervalos de Confiança

- Foi descrito anteriormente como derivar estimativas de intervalo de confiança quando se deseja estimar $erro_D(h)$ para uma hipótese h, de valores discretos, obtida utilizando uma amostra de n instâncias de exemplos de treinamento extraídos independentemente de uma distribuição D.
- A descrição feita aqui ilustra uma metodologia geral aplicada a muitos dos problemas de estimativa, que pode ser visto como o problema de estimar a média (valor esperado) de uma população com base na média de uma amostra de tamanho n extraída aleatoriamente.

Este processo geral inclui os seguintes passos:

1. Identificar qual o parâmetro p da população a ser estimado, por exemplo, $erro_D(h)$.
2. Definir o estimador Y, por exemplo $erro_S(h)$. Como já comentado, é desejável escolher uma variância mínima e estimador *unbiased*.
3. Determinar a distribuição de probabilidade D_Y que governa o estimador Y, incluindo sua média e variância.
4. Determinar o intervalo de confiança de $N\%$ encontrando os limites inferior e superior denominados L e U respectivamente, tal que $N\%$ da massa na distribuição de probabilidades D_Y encontre-se entre L e U.

A justificativa desta metodologia encontra-se no teorema do limite central.

Comparação de Hipóteses

- Sejam duas hipóteses h_1 e h_2 para alguma função meta f discretamente valorada.
- Considerando que h_1 foi testada em uma amostra S_1 contendo n_1 exemplos extraídos aleatoriamente, e h_2 foi igualmente testada em uma amostra independente S_2 contendo n_2 exemplos extraídos da mesma distribuição D, então, a diferença d entre os erros verdadeiros dessas duas hipóteses, que é o parâmetro a ser estimado, é dado por

$$d \equiv erro_D(h_1) - erro_D(h_2) \tag{11}$$

- Utilizando a metodologia descrita, após identificar esse parâmetro (passo 1) o passo seguinte (passo 2) é definir um estimador.
- Neste caso o estimador é definido como a diferença entre os erros amostrais, denotado por $\hat{d}$
- É possível provar que $\hat{d}$ fornece uma estimativa *unbiased* de d; ou seja $E[\hat{d}] = d$.

$$\hat{d} \equiv erro_{S_1}(h_1) - erro_{S_2}(h_2) \tag{12}$$

- O próximo passo (passo 3) é determinar a distribuição de probabilidade que governa a variável aleatória $\hat{d}$.
- Sabe-se que para n_1 e n_2 grandes (≥ 30), tanto $erro_{S_1}(h_1)$ quanto $erro_{S_2}(h_2)$ têm distribuições que são aproximadamente Normal.
- Devido ao fato que a diferença de suas distribuições Normais é também uma distribuição Normal, então $\hat{d}$ tem uma distribuição aproximadamente Normal com média d e sua variância é a soma das variâncias de $erro_{S_1}(h_1)$ e $erro_{S_2}(h_2)$.

Utilizando a Equação 8, Slide 17, que aproxima a variância de cada uma dessas distribuições obtém-se a variância aproximada das distribuições dos dois erros

$$\sigma^2_{\hat{d}} \approx \frac{erro_{S_1}(h_1)(1 - erro_{S_1}(h_1))}{n_1} + \frac{erro_{S_2}(h_2)(1 - erro_{S_2}(h_2))}{n_2} \quad (13)$$

Determinada a distribuição de probabilidades que governa o estimador $\hat{d}$, é simples derivar intervalos de confiança que caracterizem o provável erro ao estimar d usando $\hat{d}$.

Uma estimativa de $N\%$ intervalo de confiança para d é aproximada por:

$$\hat{d} \pm Z_N \sqrt{\frac{erro_{S_1}(h_1)(1 - erro_{S_1}(h_1))}{n_1} + \frac{erro_{S_2}(h_2)(1 - erro_{S_2}(h_2))}{n_2}} \quad (14)$$

onde Z_N é a constante descrita na Tabela 1, Slide 10.

- Com a Equação 14, Slide 33, obtém-se o intervalo de confiança *two-sided* para estimar d.
- No entanto, para obter limites *one-sided* (limitando a maior ou a menor diferença possível em erros com algum nível de confiança) essa equação deve ser modificada.

- Apesar da análise anterior considerar o caso em que h_1 e h_2 são testadas em amostras de dados independentes, é geralmente aceitável usar o intervalo de confiança visto na Equação 14 no caso de h_1 e h_2 serem testadas em uma única amostra S (onde S é também independente de h_1 e h_2).
- Neste último caso $\hat{d}$ é redefinido como

$$\hat{d} \equiv erro_S(h_1) - erro_S(h_2) \tag{15}$$

- A variância deste novo $\hat{d}$, considerando S_1 e S_2 como S, tende a ser menor que a variância dada pela Equação 13. Isto deve-se ao fato que usando uma única amostra S elimina-se a variância devido as diferenças aleatórias nas composições de S_1 e S_2.
- Neste caso, o intervalo de confiança dado pela Equação 14 geralmente será um intervalo bastante conservador, no entanto, ainda correto.

Testando Hipóteses

Em alguns casos, além de determinar o intervalo de confiança para algum parâmetro há o interesse de encontrar a probabilidade de alguma suposição ser verdadeira. Assim, surgem questões como

Qual a probabilidade que $erro_D(h_1) > erro_D(h_2)$*?*

- Por exemplo, supondo que são medidos os erros amostrais para h_1 e h_2 usando duas amostras independentes S_1 e S_2 de tamanho 100 e é encontrado que $erro_{S_1}(h_1) = 0.30$ e $erro_{S_2}(h_2) = 0.20$, então a diferença observada é $\hat{d} = 0.10$.
- Entretanto, devido a variações aleatórias nas amostras, é possível observar essa mesma diferença entre os erros das amostras, mesmo no caso de ser $erro_D(h_1) \leq erro_D(h_2)$. Assim, surgem as seguintes questões:

Qual a probabilidade que $erro_D(h_1) > erro_D(h_2)$, *dado que a diferença nos erros da amostra é* $\hat{d} = 0.10$?

e equivalentemente

Qual a probabilidade que $d > 0$, *dado que* $\hat{d} = 0.10$?

- Nota-se que a probabilidade $\Pr(d > 0)$ é igual a probabilidade que $\hat{d}$ não tenha superestimado d para um valor maior que 0.10.
- Por outro lado, esta é a probabilidade que $\hat{d}$ caia dentro de um intervalo *one-sided* $\hat{d} < d + 0.10$. Desde que d seja a média da distribuição que governa $\hat{d}$ esse intervalo pode ser expresso como $\hat{d} < \mu_{\hat{d}} + 0.10$.

- Para resumir, a probabilidade $\Pr(d > 0)$ é igual a probabilidade de $\hat{d}$ cair dentro do intervalo *one-sided* $\hat{d} < \mu_{\hat{d}} + 0.10$. Uma vez que já tenha sido calculada a distribuição aproximada que governa $\hat{d}$, pode-se determinar a probabilidade que $\hat{d}$ caia dentro deste intervalo *one-sided* ao calcular a massa da probabilidade da distribuição $\hat{d}$ dentro desse intervalo.
- Para realizar esse cálculo, o intervalo pode ser re-expresso em termos do número de desvios padrões com que ele se afasta do meio. Usando a Equação 13, Slide 32, encontra-se que $\sigma_{\hat{d}} \approx 0.061$, então pode-se re-expressar o intervalo como, aproximadamente

$$\hat{d} < \mu_{\hat{d}} + 1.64\sigma_{\hat{d}} \tag{16}$$

- Quando deseja-se saber

  Qual o nível de confiança associado com esse intervalo one-sided para uma distribuição Normal?

  deve ser consultada a Tabela 1, Slide 10, onde encontra-se que 1.64 desvios padrões sobre a média corresponde a intervalo *two-sided* com nível de confiança de 90%, assim, o intervalo *one-sided* tem um nível de confiança associado de 95%.

- Tendo em vista que $\hat{d} = 0.10$, a probabilidade que $erro_D(h_1) > erro_D(h_2)$ é aproximadamente 0.95, ou seja, em termos estatísticos diz-se que é aceita a hipótese que "$erro_D(h_1) > erro_D(h_2)$" com 0.95 de confiança ou, alternativamente, diz-se que é rejeitado a hipótese oposta (também chamada de hipótese nula) em um nível $(1 - 0.95) = 0.05$ de significância.

Comparação de Algoritmos de Aprendizado

No caso de se querer comparar o desempenho de dois algoritmos de aprendizado L_A e L_B em vez de duas hipóteses específicas, surge a seguinte questão

> *Qual o teste apropriado para comparar algoritmos de aprendizado e como determinar quando uma diferença entre os algoritmos é estatisticamente significante ou não?*

O primeiro passo é especificar o parâmetro a ser estimado para determinar qual entre os dois algoritmos, L_A e L_B, é melhor, na média, para aprender a função meta f.

- Uma maneira razoável de definir o que está "na média" é considerar o desempenho relativo desses dois algoritmos considerando a média do desempenho de ambos algoritmos sobre todos os conjuntos de treinamento de tamanho n que podem ser extraídos de acordo com a distribuição D.
- Ou seja, deseja-se estimar o valor esperado da diferença dos erros entre os dois algoritmos

$$E_{S \subset D}[erro_D(L_A(S)) - erro_D(L_B(S))] \quad (17)$$

onde $L(S)$ representa a hipótese encontrada pelo algoritmo L utilizando os exemplos de treinamento da amostra S. O subscrito $S \subset D$ indica que o valor esperado E é calculado sobre amostras S extraídas de acordo com a distribuição D.

- No entanto, na prática, quando se deseja comparar dois algoritmos de aprendizado, geralmente o que se tem é apenas uma amostra limitada D_0 dos dados.
- Nesse caso, uma maneira de estimar o valor da Equação 17 é dividir D_0 em um conjunto de treinamento S_0 e um conjunto disjunto de teste T_0.
- Assim, S_0 pode ser usado para treinar tanto L_A quanto L_B e T_0 pode ser usado para comparar a precisão das duas hipóteses aprendidas através da seguinte equação:

$$erro_{T_0}(L_A(S_0)) - erro_{T_0}(L_B(S_0)) \qquad (18)$$

Deve ser observado que existem duas diferenças fundamentais entre esse estimador e o estimador definido pela Equação 17, elas são:

1. $erro_{T_0}(h)$ é usado para aproximar $erro_D(h)$;
2. o valor medido é a diferença entre os erros apenas para um conjunto de treinamento S_0 ao invés de medir o valor esperado dessa diferença sobre todas as amostras S que podem ser extraídas de acordo com a distribuição D.

A estimativa obtida pela Equação 18 pode ser melhorada particionando repetidamente os dados D_0 em conjuntos disjuntos de treinamento e teste calculando a média dos erros sobre os conjuntos de teste para esses diferentes experimentos.

- Essa idéia é melhor descrita pelo procedimento na Tabela 3, Slide 45, o qual pode ser utilizado para estimar a diferença entre os erros de dois algoritmos de aprendizado baseado na amostra D_0 de dados disponíveis.
- Esse procedimento primeiro particiona os dados em k subconjuntos disjuntos de igual tamanho, sendo que esse tamanho deve ser pelo menos 30.
- Depois treina e testa os algoritmos de aprendizado k vezes utilizando cada vez um dos k subconjuntos como conjunto de teste e o subconjunto restante como conjunto de treinamento.
- Dessa forma, os algoritmos de aprendizado são testados sobre k conjuntos de testes independentes e a diferença média dos erros, denotada por $\overline{\delta}$, é retornada como uma estimativa do valor esperado da diferença dos erros entre os dois algoritmos de aprendizado.

Ou seja, $\overline{\delta}$ é uma estimativa de

$$E_{S \subset D_0}[erro_D(L_A(S)) - erro_D(L_B(S))] \quad (19)$$

onde S representa uma amostra aleatória de tamanho $\frac{k-1}{k}|D_0|$ extraída aleatoriamente de D_0.

1. Particionar os dados disponíveis D_0 em k subconjuntos disjuntos $T_1, T_2, \ldots, T_k$ de igual tamanho — pelo menos 30 exemplos em cada T_i
2. Para i de 1 até k,
   utilizar T_i como conjunto de teste e o conjunto restante como conjunto de treinamento S_i
   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow erro_{T_i}(h_A) - erro_{T_i}(h_B)$
3. Retornar o valor $\overline{\delta}$, onde

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

Tabela 3: Procedimento para Estimar a Diferença do Erro entre dois Algoritmos de Aprendizado L_A e L_B

- A única diferença entre a Equação original 17 e a Equação 19 é que, nesta última, o valor esperado é calculado sobre subconjuntos do conjunto de dados disponíveis D_0, e não sobre subconjuntos retirados da população total de acordo com a distribuição D.
- O $N\%$ intervalo de confiança para estimar a quantidade definida pela Equação 19 é dado por

$$\overline{\delta} \pm t_{N,k-1} s_{\overline{\delta}} \tag{20}$$

onde $t_{N,k-1}$ é uma constante análoga a Z_N mas para a distribuição t e $s_{\overline{\delta}}$ é definida por

$$s_{\overline{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \overline{\delta})^2} \tag{21}$$

é uma estimativa do desvio padrão da distribuição t que governa a variável aleatória δ.

- O primeiro subscrito N na constante $t_{N,k-1}$ da Equação 20 representa o nível de confiança. O segundo subscrito, usualmente denotado v, é o número de graus de liberdade.
- Esse número está relacionado com o número de testes independentes realizados para encontrar o valor da variável aleatória δ que, neste caso, é $v = k - 1$.

A Tabela 4 mostra alguns valores de $t_{N,v}$, que se aproximam a Z_N para $v \longrightarrow \infty$.

	$N\%$ Intervalo de Confiança			
	90%	95%	98%	99%
$v = 2$	2.92	4.30	6.96	9.92
$v = 5$	2.02	2.57	3.36	4.03
$v = 10$	1.81	2.23	2.76	3.17
$v = 20$	1.72	2.09	2.53	2.84
$v = 30$	1.70	2.04	2.46	2.75
$v = 120$	1.66	1.98	2.36	2.62
$v = \infty$	1.64	1.96	2.33	2.58

Tabela 4: Valores de $t_{N,v}$ para Intervalos de Confiança *two-sided*

- Deve ser observado que o procedimento utilizado para estimar $\overline{\delta}$, descrito na Tabela 3, considera amostras idênticas — denominado de testes pareados — enquanto que o método descrito em Comparação de Hipóteses (a partir do Slide 29) utiliza amostras independentes.
- Em geral, testes pareados produzem intervalos de confiança menores já que qualquer diferença nos erros observados deve-se às diferenças entre as hipóteses.
- Por outro lado, quando as hipóteses são testadas em amostras diferentes, a diferença do erro entre duas amostras pode ser atribuído não somente às diferenças entre as hipóteses mas também devido as diferenças nessas duas amostras.

- Um outro teste estatístico simples está baseado na diferença entre a taxa de erro de cada um dos dois algoritmos L_A e L_B, ou seja, os erros não são pareados e as partições T_1, T_2,...,T_k na Tabela 3 não são, necessariamente, idênticas para ambos algoritmos.
- Neste caso, são conhecidos os valores $\overline{\delta}_A = \frac{1}{k}\sum_{i=1}^{k}\delta_{i_A}$ e $\overline{\delta}_B = \frac{1}{k}\sum_{i=1}^{k}\delta_{i_B}$ onde $\delta_{i_A} = erro_{T_{i_A}}(h_A)$ e $\delta_{i_B} = erro_{T_{i_B}}(h_B)$, bem como os desvios padrões $\sigma_{\overline{\delta}_A}$ e $\sigma_{\overline{\delta}_B}$.

- Após realizar diversas aproximações e no caso de se verificar que

$$\frac{|\overline{\delta}_A - \overline{\delta}_B|}{\sqrt{\frac{\sigma^2_{\overline{\delta}_A} + \sigma^2_{\overline{\delta}_B}}{2}}} > 2 \qquad (22)$$

  pode-se considerar que existe uma diferença, com 95% de nível de confiança, de um algoritmo superar o outro.

- Se $\overline{\delta}_A > \overline{\delta}_B$ então o algoritmo L_B supera o algoritmo L_A, caso contrário L_A supera L_B.

- A fim de exemplificar a comparação de dois algoritmos de aprendizado utilizando este último teste, foram considerados um subconjunto de algoritmos, dados e medidas realizadas e publicadas em [Baranauskas & Monard, 1999].
- Dentre os diversos algoritmos analisados nesse trabalho, foram escolhidos os algoritmos simbólicos $\mathcal{CN}2$ [Clark & Niblett, 1989] e $\mathcal{C}4.5$ [Quinlan, 1993], disponíveis na biblioteca $\mathcal{MLC}$++ [Félix et al., 1998, Kohavi et al., 1994], os quais são bastante conhecidos pela comunidade de Aprendizado de Máquina.

- A Tabela 5 mostra o subconjunto escolhido do conjunto de dados analisados nesse trabalho.
- Nessa tabela, para cada conjunto de dados é mostrado em cada coluna: o número de exemplos (#E), número de atributos (#A) conforme os tipos contínuo (c) e discreto (d), número de classes (#C), erro majoritário e se o conjunto de dados possui valores desconhecidos.
- Os conjuntos de dados são apresentados em ordem crescente do número de atributos.

Conjunto de Dados	#E	#A(c,d)	#C	Erro Majoritário	Valores Desconhecidos
bupa	345	6 (6,0)	2	42.03%	N
pima	769	8 (8,0)	2	34.98%	N
hungaria	294	13 (13,0)	2	36.05%	S
crx	690	15 (6,9)	2	44.49%	S
letter	15000	16 (16,0)	26	95.92%	S
hepatitis	155	19 (6,13)	2	20.65%	S
sonar	208	60 (60,0)	2	46.63%	N
dna	3186	180 (0,180)	3	48.09%	N

Tabela 5: Descrição dos Conjuntos de Dados

- A Figura 1 mostra a dimensionalidade dos conjuntos de dados, isto é, número de atributos e número de instâncias de cada conjunto de dados.
- Devido a grande variação, o número de instâncias na Figura 1 é representado como $\log_{10}(\#\text{Instâncias})$.

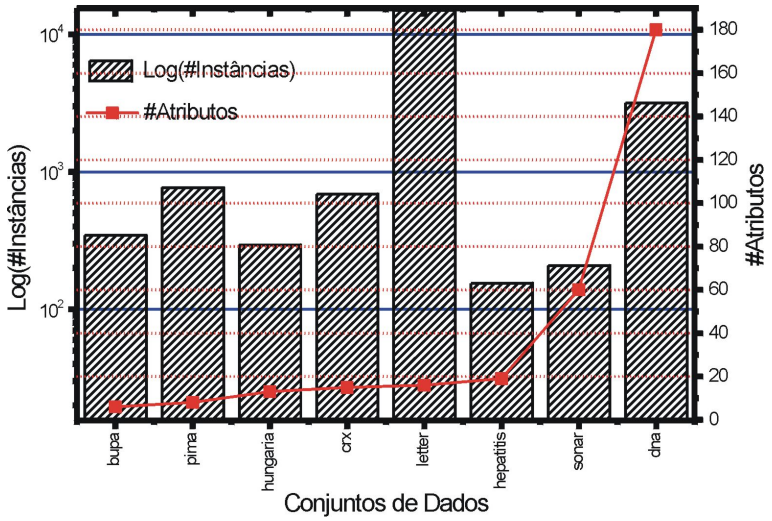


Figura 1: Dimensionalidade dos Conjuntos de Dados

Cada um dos conjuntos de dados está relacionado com algum domínio específico conforme a descrição resumida a seguir:

- bupa: para predizer quando um paciente masculino terá ou não desordens hepáticas baseado em vários testes de sangue e na quantia de consumo de álcool;
- pima: para predizer quando um paciente terá ou não teste positivo para diabetes de acordo com critérios da Organização Mundial de Saúde;
- hungaria: está relacionado com o diagnóstico de doenças do coração;

- crx: está relacionado com aplicações de cartões de crédito;
- letter: para reconhecer uma das 26 letras maiúsculas do alfabeto inglês identificadas como pontos em preto e branco dentro de um retângulo;
- hepatitis: está relacionado com a expectativa de vida de pacientes com hepatite;
- sonar: para predizer diferenças entre sinais de sonar transmitidos de metais e sinais transmitidos de rocha;
- dna: relacionado com a área de biologia molecular trata sobre características de DNA.

Os conjuntos de dados são do repositório UCI Irvine [Blake et al., 1998], com excessão do dna foi obtido do Projeto StatLog [Taylor et al., 1994].

- Para cada um dos conjuntos de dados foram aplicados os algoritmos de aprendizado $\mathcal{C}4.5$ e $\mathcal{CN}2$ executados utilizando a biblioteca $\mathcal{MLC}$++ com seus parâmetros $default$.
- Na Tabela 6 são mostradas as taxas de erro (média e desvio padrão das 10-$fold\ stratified\ cross\text{-}validation$) de $\mathcal{C}4.5$ e $\mathcal{CN}2$ para cada um dos conjuntos de dados [Baranauskas & Monard, 1999].
- Nessa tabela também é mostrado, na última coluna, o valor da diferença entre as duas taxas de erro calculado conforme a Equação 22.
- A diferença entre k-$fold$ e k-$fold\ stratified\ cross\text{-}validation$ é que neste último caso procura-se manter a distribuição das classes em cada uma das k partições.

Conjunto de Dados	$\mathcal{C}4.5$	$\mathcal{CN}2$	Diferença entre Taxas de Erro
bupa	32.29 ± 1.73	32.18 ± 2.11	0.06 (+)
pima	25.74 ± 1.13	25.38 ± 1.38	0.29 (+)
hungaria	22.48 ± 4.20	22.07 ± 3.06	0.11 (+)
crx	15.65 ± 1.18	16.80 ± 1.21	0.96 (−)
letter	13.41 ± 0.34	29.66 ± 0.30	50.68 (−)
hepatitis	20.62 ± 2.27	18.25 ± 3.83	0.75 (+)
sonar	30.26 ± 1.97	28.81 ± 3.30	0.53 (+)
dna	7.60 ± 0.46	11.85 ± 0.62	7.79 (−)

Tabela 6: Taxas de Erro Obtidas para $\mathcal{C}4.5$ e $\mathcal{CN}2$ Utilizando 10-*fold stratified cross-validation*

- Para determinar quando a diferença entre os dois algoritmos (L_A=$\mathcal{C}4.5$ e L_B=$\mathcal{CN}2$) é significativa ou não, é mostrado na Figura 8 um gráfico de barras, no qual cada barra representa a medida definida pela Equação 22.
- Interpretando a Figura 8, qualquer barra que tenha comprimento maior que 2 indica que esse resultado é significante com nível de confiança de 95%.
- Quando a barra está abaixo de zero indica que o primeiro algoritmo ($\mathcal{C}4.5$) supera o segundo algoritmo ($\mathcal{CN}2$); se a barra está acima do zero indica o contrário, isto é, que o segundo algoritmo supera o primeiro.
- Assim, quando o comprimento da barra está abaixo de 2 indica que o primeiro algoritmo supera significantemente o segundo algoritmo, no caso contrário, o segundo supera significantemente o primeiro.

- Pode-se observar na Figura 8 que apenas para os conjuntos de dados dna e letter o algoritmo $\mathcal{C}4.5$ supera significantemente o algoritmo $\mathcal{CN}2$ com nível de confiança de pelo menos 95%, nos demais conjuntos de dados a diferença entre os dois algoritmos não é significante.
- Outro fato, bem conhecido na comunidade de Aprendizado de Máquina, e também constatado na Figura 8, refere-se a impossibilidade de afirmar que um algoritmo de aprendizado será sempre melhor que outro em todas as situações, pois os conjuntos de dados influenciam bastante no desempenho dos algoritmos de aprendizado.

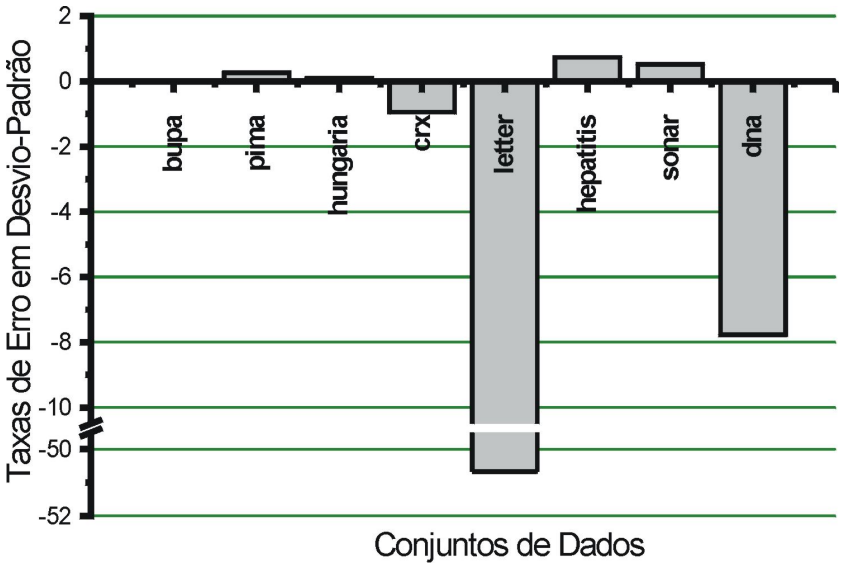


Figura 2: Diferença Absoluta em Desvios Padrões das Taxas de Erro para $\mathcal{C}4.5$ e $\mathcal{CN}2$

- Deve ser observado que existem vários outros métodos, propostos na literatura, para comparar algoritmos de aprendizado.
- Nos últimos anos, têm aparecido debates ativos na comunidade de Aprendizado de Máquina, relacionados à eficácia de vários desses métodos, especialmente quando aplicados a conjuntos de exemplos pequenos.

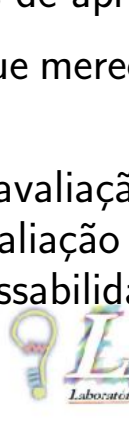
- Um dos problemas é que a maioria das aproximações estatísticas realizadas para encontrar uma estimativa do valor $E_{S \subset D}$, Equação 17, Slide 41, são válidas sempre que as medidas sejam realizadas sobre conjuntos independentes.
- Entretanto, se considerarmos a partição dos dados disponíveis, D_0, em k subconjuntos disjuntos $T_1, T_2, ..., T_k$, como proposto no procedimento descrito na Tabela 3, Slide 45, é possível observar que ainda que cada subconjunto de teste é independente do outro, isso não se verifica para os conjuntos de treinamento, os quais não são independentes.
- Na realidade, todo par de conjunto de treinamento compartilha 80% dos exemplos em D_0.

- Um outro problema é que o tamanho dos conjuntos de teste é $\frac{1}{k}|D_0|$. Assim, o número de exemplos de teste pode não ser suficientemente grande para aproximar a distribuição Binomial pela Normal.
- No excelente trabalho de [Dietterich, 1996] é apresentada uma discussão sobre testes estatísticos para determinar quando um algoritmo é significantemente melhor que outro numa determinada tarefa de aprendizado.
- Esses testes estatísticos são comparados experimentalmente a fim de medir a probabilidade de cometer um erro de Tipo I.
- Um erro de Tipo I acontece quando a hipótese nula é verdadeira, isto é, não existe diferença entre os algoritmos L_A e L_B, mas a hipótese nula é rejeitada.

- Também em [Salzberg, 1997] são discutidos alguns testes estatísticos para comparar algoritmos de aprendizado.
- Segundo [Dietterich, 1996] o teste pareado descrito na Tabela 3, Slide 45, apresenta geralmente um bom desempenho, enquanto que o teste da Equação 22, Slide 51, pode apresentar, em algumas situações, uma alta probabilidade de erro do Tipo I.

Considerações Finais

- Foram discutidos alguns dos métodos utilizados para estimar a precisão de hipóteses, comparar a precisão de hipóteses e de algoritmos de aprendizado quando existem poucos dados disponíveis.
- Parte do desenvolvimento deste capítulo bem como a notação utilizada estão baseados no Capítulo 5 de [Mitchell, 1997].
- Como mencionado anteriormente, existem debates ativos na comunidade relacionados aos estimadores estatísticos utilizados para comparar corretamente o desempenho de algoritmos de aprendizado.
- Consideramos esse um tema muito importante que merece a atenção dos pesquisadores.
- Deve ser observado que tão importante quanto a avaliação de hipóteses objetivando calcular e estimar sua precisão é a avaliação de hipóteses com objetivos de medir a sua qualidade e interessabilidade.



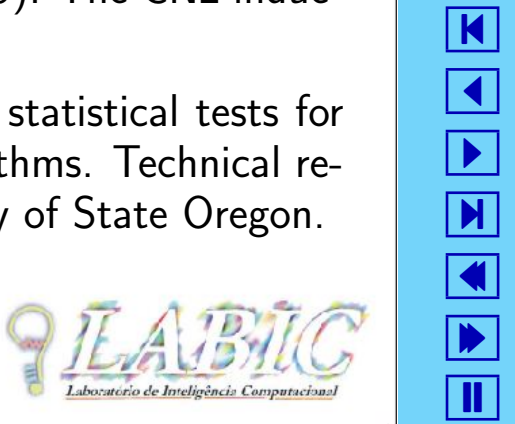
Referências

[Baranauskas & Monard, 1999] Baranauskas, J. A. & Monard, M. C. (1999). The MLC^{++} wrapper for feature subset selection using decision tree, production rule, instance based and statistical inducers: Some experimental results. Technical Report 87, ICMC-USP.

[Blake et al., 1998] Blake, C., Keogh, E., & Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/ mlearn/MLRepository.html.

[Clark & Niblett, 1989] Clark, P. & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, 3(4):261–283.

[Dietterich, 1996] Dietterich, T. G. (1996). Proper statistical tests for comparing supervised classification learning algorithms. Technical report, Department of Computer Science, University of State Oregon.



[Félix et al., 1998] Félix, L. C. M., Rezende, S. O., Doi, C. Y., de Paula, M. F., & Romanato, M. J. (1998). MLC^{++} biblioteca de aprendizado de máquina em C^{++}. Technical Report 72, ICMC-USP.

[Kohavi et al., 1994] Kohavi, R., Sommerfield, D., & Dougherty, J. (1994). MLC^{++}: *A Machine Learning Library in* C^{++}. IEEE Computer Society Press.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, Los Altos, California, USA.

[Salzberg, 1997] Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. http://www.cs.jhu.edu/∼salzberg/critique.ps.

[Taylor et al., 1994] Taylor, C., Mitchie, D., & Spiegelhater, D. (1994). *Machine Learning, Neural and Statistical Classification*. Paramount Publishing International.