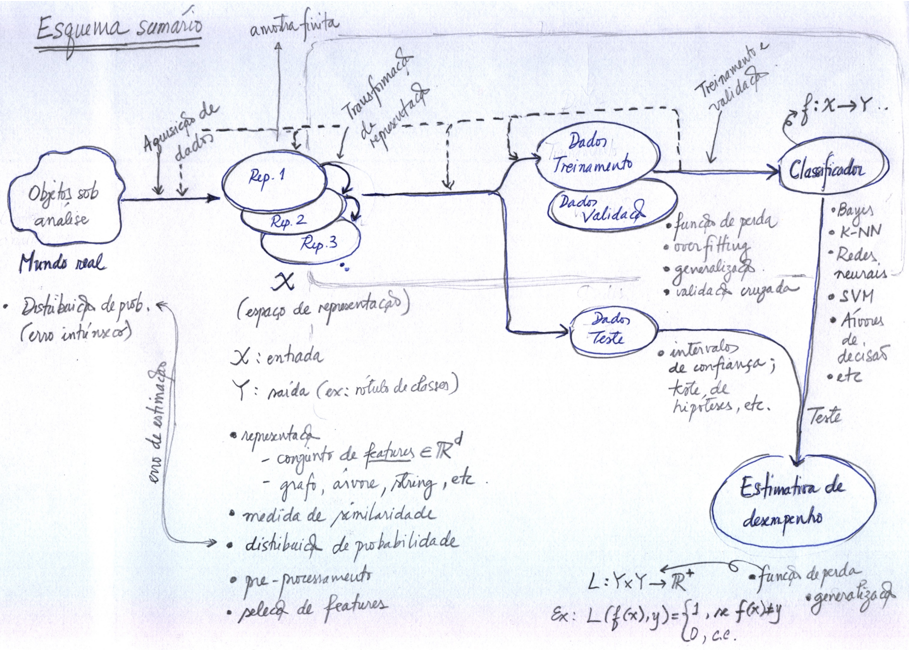
MAC 0460 / 5832

Aprendizagem Computacional
Modelos, algoritmos e aplicações

Nina Hirata (nina@ime.usp.br)
Sala 6 - bloco C

Monitor: Igor

Aula 9 (2012)


Esquema sumário
amostra finita
Aquisição de dados
Transformação de representação
Treinamento e validação
Objetos sob análise
Mundo real
Rep. 1
Rep. 2
Rep. 3
X
(espaço de representação)
Dados Treinamento
Dados Validação
Classificador
f : X → Y ..
• Bayes
• K-NN
• Redes neurais
• SVM
• Árvores de decisão
• etc
• função de perda
• overfitting
• generalização
• validação cruzada
Dados Teste
• intervalos de confiança; teste de hipóteses, etc.
Teste
Estimativa de desempenho
• Distribuição de prob. (erro intrínseco)
erro de estimação
X : entrada
Y : saída (ex: rótulo de classes)
• representação
- conjunto de features ∈ ℝ^d
- grafo, árvore, string, etc.
• medida de similaridade
• distribuição de probabilidade
• pre-processamento
• seleção de features
L : Y×Y → ℝ⁺
Ex: L(f(x),y) = { 1, se f(x)≠y ; 0, c.c.
• função de perda
• generalização

Erro de um classificador

Dada uma distribuição de probabilidade conjunta p sobre o espaço $X \times Y$, o erro de um classificador $f : X \to Y$, segundo uma função de perda L, é dado pela esperança:

$$Erro(f) = E\left[L(y, f(\mathbf{x}))\right]$$

No caso da função de perda zero-um, temos

$$Erro(f) = P_{\mathbf{x} \in X}\left(y \neq f(\mathbf{x})\right)$$

Erro estimado

Erro estimado de g (com respeito a um conjunto de teste T com n elementos):
(proporção de exemplos classificados incorretamente por g)

$$Erro_T(g) = \frac{1}{n} \sum_{\mathbf{x}_i \in T} \delta(y_i, g(\mathbf{x}_i))$$

na qual $\delta(a, b) = 1$ se $a \neq b$ e $\delta(a, b) = 0$ se $a = b$.

Usando o fato de que $Erro_T(g)$ pode ser encarado como tendo uma distribuição binomial (errou/acertou classificação) e o fato de uma distribuição binomial poder ser aproximada por uma distribuição normal, temos que

o $Erro(g)$, encontra-se com $N\%$ de probabilidade no intervalo

$$Erro_T(g) \pm z_N \sqrt{\frac{Erro_T(g)(1 - Erro_T(g))}{n}}$$

Valores de z_N para diferentes valores de N:

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

Não se pode tirar conclusões apenas a partir do erro estimado sobre um conjunto de testes T, especialmente quando $n = |T|$ é pequeno.

Opção: validação cruzada (com ou sem reposição)

Comparação de dois classificadores

Considerar a diferença

$$\hat{d} = Erro_{T_1}(g_1) - Erro_{T_2}(g_2)$$

como estimador para a diferença verdadeira $d = Erro(g_1) - Erro(g_2)$.

Aproximando a distribuição de cada um deles por uma normal, a distribuição de $\hat{d}$ pode ser aproximada por uma normal (pois a diferença de duas normais é uma normal).

De forma similar ao caso anterior, um **intervalo de confiança** de $N\%$ para d, é

$$\hat{d} \pm Z_N \sqrt{\frac{Erro_{T_1}(g_1)(1 - Erro_{T_1}(g_1))}{n_1} + \frac{Erro_{T_2}(g_2)(1 - Erro_{T_2}(g_2))}{n_2}}$$

Teste de hipóteses

Comparação baseada em teste de hipóteses

- Sejam g_1 e g_2 dois classificadores.
- $Erro(g_1)$ e $Erro(g_2)$ são seus erros verdadeiros.
- $d = Erro(g_1) - Erro(g_2)$. Logo $d > 0$ implica que erro de g_1 é maior que erro de g_2.
- Na prática só temos erros estimados $Erro_{T_1}(g_1)$ e $Erro_{T_2}(g_2)$.

Queremos saber:

Qual a probablidade de que "$Erro(g_1) > Erro(g_2)$", dado que foi observado $\hat{d} = Erro_{T_1}(g_1) - Erro_{T_2}(g_2)$?

EXEMPLO:

- **suponha** $Erro_{T_1}(g_1) = 0.3$ **e** $Erro_{T_2}(g_2) = 0.2$**; logo** $\hat{d} = 0.1$
- **Pergunta: qual é a probabilidade de que "**$Erro(g_1) > Erro(g_2)$ **?", dado que observamos** $\hat{d} = 0.1$**?**
- **equivalentemente, qual a probabilidade de que** $d > 0$**, dado que observamos** $\hat{d} = 0.1$**?**
- **temos** $d > 0 \iff d > \hat{d} - 0.1 \iff \hat{d} < d + 0.1$
- **qual é a distribuição de** $\hat{d}$ **? Pode ser aproximada por uma normal com média** $\mu_{\hat{d}}$
- **Assim, a probabilidade de** $\hat{d} < d + 0.1$ **é a mesma de** $\hat{d} < \mu_{\hat{d}} + 0.1$

- **Como sabemos a distribuição de $\hat{d}$, podemos calcular a massa de probabilidade no intervalo de interesse**
- $\sigma^2_{\hat{d}} \approx \frac{Erro_{T_1}(g_1)(1-Erro_{T_1}(g_1))}{n_1} + \frac{Erro_{T_2}(g_2)(1-Erro_{T_2}(g_2))}{n_2}$
- **Para** $Erro_{T_1}(g_1) = 0.3$ **e** $Erro_{T_2}(g_2) = 0.2$**,** $\sigma^2_{\hat{d}} \approx 0.61$
- **Logo** $\hat{d} < \mu_{\hat{d}} + 0.1$ **pode ser escrito como** $\hat{d} < \mu_{\hat{d}} \pm 1.64\sigma_{\hat{d}}$ **(pois** $1.64 \times 0.61 \approx 1$**)**
- **1.64 é o coeficiente associado ao intervalo de confiança de 90%**
- **Logo a probabilidade de** $\hat{d} < \mu_{\hat{d}} + 1.64\sigma_{\hat{d}}$ **é 90% + o extremo esquerdo (5%)**
- **Isto é, podemos dizer com 95% de confiança que** $\hat{d} < d + 0.1$

E nos casos de validação cruzada ??

Abordagens do tipo validação cruzada estimam o **erro médio** (e não o erro verdadeiro).

Tipicamente, em validação cruzada, o interesse é em comparar dois tipos de classificadores.

Em vez de considerar a probabilidade de erro, podemos considerar o **erro médio**

Dado um algoritmo de treinamento, para cada conjunto de treinamento S, temos um classificador $g(S)$ e um erro $Erro(g(S))$

O erro médio do algoritmo é a média dos erros calculados sobre cada possível conjunto de treinamento S de mesmo tamanho n.

Estimação do erro esperado

k-**fold Cross-validation paired** t-**Test**

Particiona-se S em k subconjuntos de mesmo tamanho T_1, T_2, ..., T_k, e obtém-se k classificadores usando os algoritmos L_A e L_B. Sejam então $g_{A\,i} = L_A(S \setminus T_i)$, $g_{B\,i} = L_B(S \setminus T_i)$ e $\hat{\hat{\delta}}_i = Erro_{T_i}(g_{A\,i}) - Erro_{T_i}(g_{B\,i})$. O **erro médio estimado** é então dado por

$$\hat{\hat{\delta}} = \frac{1}{k} \sum_{i=1}^{k} \hat{\hat{\delta}}_i$$

Qual é a qualidade de $\hat{\hat{\delta}}$?

- para valores grandes de n, podemos supor que $\hat{\delta}_i$ segue uma distribuição aproximadamente normal (teorema do limite central)
- Então $\hat{\delta}$ também segue uma distribuição normal com média conhecida, mas variância desconhecida.
- usando-se a variância amostral obtém-se uma distribuição t (em vez da normal)
- intervalo de confiança para $\hat{\delta}$ é calculado usando-se a tabela para a distribuição t: o intervalo de confiança de $N\%$ de δ é dado por

$$\hat{\delta} \pm t_{N,k-1} VAR[\hat{\delta}]$$

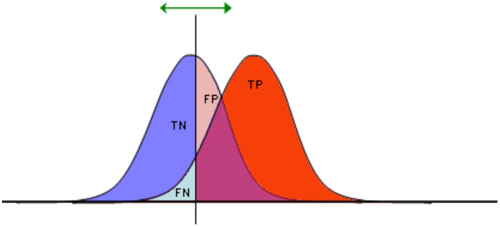
na qual

$$VAR[\hat{\delta}] = \sqrt{\frac{1}{k(k-1)\sum_{i=1}^{k}(\delta_i - \hat{\delta})^2}}$$

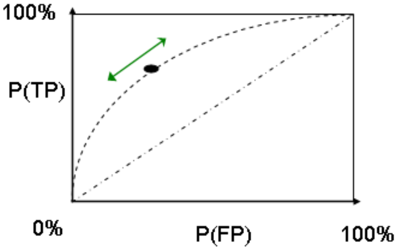
e $k-1$ representa o grau de liberdade (relacionado ao número de eventos independentes que produzem os valores de $\hat{\delta}$)

Por exemplo, para $N = 95$ e $k = 11$, a constante é $t_{95,9} = 2.23$.

Curva ROC (Receiving Operating Characteristic)



TP	FP
FN	TN
1	1



Curva ROC (Receiving Operating Characteristic)

Curvas ROC podem ser desenhadas quando se tem um classificador binário com algum parâmetro ordenado; por exemplo, quando aumentar o valor desse parâmetro implica em aumentar a quantidade de exemplos aceitos (isso aumenta taxa de TP, mas também aumenta taxa de FP).

Olhando a curva, pode-se escolher o parâmetro ótimo.

Ou então, no caso de dois classificadores, pode-se escolher aquele com a melhor curva.