SCC0173 – Mineração de Dados Biológicos

Classificação IV: Avaliação de Classificadores

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

Créditos

- O material a seguir consiste de adaptações e extensões dos originais:
  - gentilmente cedidos pelo Prof. André C. P. L. F. de Carvalho
  - de (Tan et al., 2006)

Aula de Hoje

- Avaliação de Classificadores
  - Procedimentos Básicos de Teste
    - Holdout e Cross-Validation
  - Medidas de Desempenho
- Problemas com Classes Difíceis
  - Técnicas Elementares para Classes Desbalanceadas
    - Balanceamento por Sub- e/ou Sobre-Amostragem

Desempenho de Classificação

- Espera-se de um classificador que ele apresente desempenho adequado para dados não vistos
  - Acurácia,
  - Pouca sensibilidade ao uso de diferentes amostras de dados, ...
- Desempenho do classificador deve ser avaliado
  - Para tanto utilizam-se conjuntos distintos de exemplos de **treinamento** e exemplos de **teste**
    - Permitem estimar a capacidade de generalização do classificador
    - Permitem avaliar a variância (estabilidade) do classificador

Avaliação de Desempenho

- Existem diferentes métodos para organização e utilização dos dados (exemplos) disponíveis em conjuntos de treinamento e teste
- Por exemplo:
  - **Holdout**
    - **Random Subsampling**
  - **Cross-Validation**

Holdout

- Também conhecido como *split-sample*
- Técnica mais simples
- Faz uma única partição da amostra em:
  - Conjunto de treinamento
    - geralmente 1/2 ou 2/3 dos dados
  - Conjunto de teste
    - dados restantes

Holdout

- **Problema**: dependência da composição dos conjuntos
- É mais crítico em "pequenas" quantidades de dados...
  - Quanto menor o conjunto de treinamento, maior a variância (sensibilidade / instabilidade) do classificador a ser obtido
  - Quanto menor o conjunto de teste, menos confiável a acurácia estimada do classificador para dados não vistos
  - Conjuntos de treinamento e teste podem não ser independentes
    - Classe sub-representada em um será super-representada no outro

Random Subsampling

- **Múltiplas execuções de Holdout**
  - Diferentes partições treinamento-teste são escolhidas de forma aleatória
    - Não pode haver interseção entre os dois conjuntos
    - Desempenho de classificação é avaliado para cada partição
    - Desempenho estimado para dados não vistos é o desempenho médio para as diferentes partições
- Permite uma estimativa de erro mais precisa
  - Porém, não controla número de vezes que cada exemplo é utilizado nos treinamentos e nos testes...

Random Subsampling

- Exemplo:
  - Supor que o conjunto de dados original seja formado pelos dados: x_1, x_2 , x_3 , x_4 , x_5 , x_6 , x_7 , x_8
    - Possíveis partições:

	Treinamento	Teste
Part. 1	x_2, x_4, x_6, x_7	x_5, x_8, x_1, x_3
Part. 2	x_3, x_4, x_5, x_8	x_1, x_7, x_2, x_6
Part. 3	x_3, x_4, x_5, x_7	x_2, x_8, x_1, x_6

Cross-Validation

- Validação cruzada
- Classe de métodos para estimativa da taxa de erro verdadeira
  - **k-fold cross-validation**
    - Cada objeto participa o mesmo número de vezes do treinamento (k – 1 vezes)
    - Cada objeto participa o mesmo número de vezes do teste (1 vez)

k-Fold Cross-Validation

- Divide conjunto de dados em k partições mutuamente exclusivas
  - A cada iteração, uma das k partições é usada para testar o modelo
    - As outras k – 1 são usadas para treinar o modelo
  - Taxa de erro é tomada como a média dos erros de teste das k partições
- Exemplo Típico
  - **10-fold cross-validation**

k-Fold Cross-Validation

- **k-fold cross-validation estratificada**
  - Mantém nas pastas as proporções de exemplos das classes presentes no conjunto total de dados

Medidas de Desempenho

- Principal objetivo de um modelo é predizer com sucesso o valor de saída para novos exemplos
  - *Errar* o mínimo possível
- Existem várias medidas de "erro" e "acerto"
  - diferentes medidas podem capturar diferentes aspectos do desempenho de classificadores

Taxa de Classificação Incorreta

- A medida mais básica para estimar a taxa de erro de um classificador é denominada de **taxa de classificação incorreta** (misclassification rate):
  - É simplesmente a proporção dos exemplos de teste que são classificados incorretamente pelo classificador
  - Usualmente é mensurada indiretamente através do seu complemento, a **taxa de classificação correta**
    - Denominada de **Acurácia**
    - Acurácia = 1 – taxa de classificação incorreta

Acurácia

- Também chamada de **accuracy** (do inglês)
  - Trata as classes igualmente...
  - Pode não ser adequada para classes desbalanceadas
    - Classe rara é normalmente mais interessante que a majoritária
    - No entanto, a medida tende a privilegiar a classe majoritária

Limitação da Acurácia

- Considere um problema de 2 classes
  - No. de exemplos da classe 0 = 9990
  - No. de exemplos da classe 1 = 10
- Se o modelo predizer qualquer exemplo como da classe 0, acurácia será 9990/10000 = 99.9 %
  - Acurácia pode ser enganadora...

Tipos de Erros

- Em classificação binária, em geral se adota a convenção de rotular os exemplos da classe de maior interesse como **positivos** (**+**)
  - Normalmente a classe rara ou minoritária
  - Demais exemplos são rotulados como **negativos** (**–**)
- Em alguns casos, os erros têm igual importância
- Em muitos casos, no entanto, esse não é o caso
  - Ex. diagnóstico negativo para indivíduo doente...

Tipos de Erros

- Dois tipos de erro em classificação binária:
  - Classificação de um exemplo **N** como **P**
    - Falso Positivo (**FP** – alarme falso)
      - Ex.: Diagnosticado como doente, mas está saudável
  - Classificação de um exemplo **P** como **N**
    - Falso Negativo (**FN**)
      - Ex.: Diagnosticado como saudável, mas está doente

Matriz de Confusão

- **Matriz de Confusão** (**Tabela de Contingência**)
  - Pode ser utilizada para distinguir os tipos de erros
  - Base de várias medidas de desempenho alternativas à accuracy
  - Pode ser utilizada com 2 ou mais classes

	Classe Verdadeira		
Classe Prevista	**1**	**2**	**3**
1	25	10	0
2	0	40	0
3	5	0	20

Avaliação de Desempenho

- Matriz de confusão para 2 classes

	Classe Verdadeira	
Classe Prevista	**P**	**N**
P	**70**	**40**
N	**30**	**60**

Classe Prevista \ Classe Verdadeira	P	N
P	**VP**	**FP**
N	**FN**	**VN**

Avaliação de Desempenho

- Medidas de erro

Taxa de FP = $\frac{FP}{FP+VN}$
(alarmes falsos)

Taxa de FN = $\frac{FN}{VP+FN}$

Erro do tipo I

Classe Prevista \ Classe Verdadeira	P	N
P	VP	FP
N	FN	VN

Erro do tipo II

Classe Prevista \ Classe Verdadeira	P	N
P	VP	FP
N	FN	VN

Exemplo

- Avaliação de 3 classificadores

Classe Prevista \ Classe Verdadeira	P	N
P	20	15
N	30	35

Classificador 1
TFN =
TFP =

Classe Prevista \ Classe Verdadeira	P	N
P	70	50
N	30	50

Classificador 2
TFN =
TFP =

Classe Prevista \ Classe Verdadeira	P	N
P	60	20
N	40	80

Classificador 3
TFN =
TFP =

Exemplo

- Avaliação de 3 classificadores

Classe Prevista \ Classe Verdadeira	P	N
P	20	15
N	30	35

Classificador 1
TFN = 0.6
TFP = 0.3

Classe Prevista \ Classe Verdadeira	P	N
P	70	50
N	30	50

Classificador 2
TFN = 0.3
TFP = 0.5

Classe Prevista \ Classe Verdadeira	P	N
P	60	20
N	40	80

Classificador 3
TFN = 0.4
TFP = 0.2

Exercício

- Avaliar os 3 classificadores abaixo:

Classe Prevista \ Classe Verdadeira	P	N
P	25	10
N	45	60

Classificador 1
TFN =
TFP =

Classe Prevista \ Classe Verdadeira	P	N
P	70	20
N	15	30

Classificador2
TFN =
TFP =

Classe Prevista \ Classe Verdadeira	P	N
P	70	95
N	30	5

Classificador 3
TFN =
TFP =

Avaliação de Desempenho

- Medidas freqüentemente utilizadas

Taxa de FP = (Erro tipo I) $\frac{FP}{FP+VN}$

Precisão = $\frac{VP}{VP+FP}$

Acurácia = $\frac{VP+VN}{VP+VN+FP+FN}$

Especificidade = $\frac{VN}{VN+FP}$ = 1–TFP

Taxa de VP = (Sensibilidade) $\frac{VP}{VP+FN}$

Revocação = (Recall) $\frac{VP}{VP+FN}$

Medida-F = $\frac{2}{1/prec+1/rev}$

Taxa de FN = (Erro tipo II) $\frac{FN}{VP+FN}$ = 1–TVP

Revocação *vs* Precisão

- **Revocação** (recall, sensibilidade, taxa de VP)
  - Taxa com que classifica como positivos todos os exemplos que são de fato positivos
    - Só considera os exemplos positivos
      - Normalmente classe de maior interesse
- **Precisão** (precision)
  - Taxa com que todos os exemplos classificados como positivos são realmente positivos
    - Só considera os exemplos classificados como positivos

Especificidade

- **Especificidade** (Especificity)
  - Taxa com que classifica como negativos todos os exemplos que são de fato negativos
    - Só considera os exemplos negativos

F-Measure

- **Medida F** (F-Measure)
  - Média harmônica ponderada da precisão e da revocação

$$\frac{(1+\alpha)\times(prec\times rev)}{\alpha\times prec+rev}$$

  - **Medida F_1**
    - Média harmônica simples (precision e recall com mesmo peso)

$$\frac{2\times(prec\times rev)}{prec+rev} = \frac{2}{\frac{1}{prec}+\frac{1}{rev}}$$

Exemplo

- Seja um classificador com a seguinte matriz de confusão. Calcular:
  - Acurácia
  - Precisão
  - Revocação (sensibilidade)
  - Especificidade

Classe Prevista \ Classe Verdadeira	P	N
P	70	40
N	30	60

Exemplo

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}$$

$$\text{Precisão} = \frac{VP}{VP+FP}$$

$$\text{Revocação} = \frac{VP}{VP+FN}$$

$$\text{Especificidade} = \frac{VN}{VN+FP}$$

Previsto \ Verdadeiro	P	N
P	VP	FP
N	FN	VN

Previsto \ Verdadeiro	P	N
p	70	40
n	30	60

Exemplo

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN} = (70 + 60) / (70 + 30 + 40 + 60) = 0.65$$

$$\text{Precisão} = \frac{VP}{VP+FP} = 70/(70+40) = 0.64$$

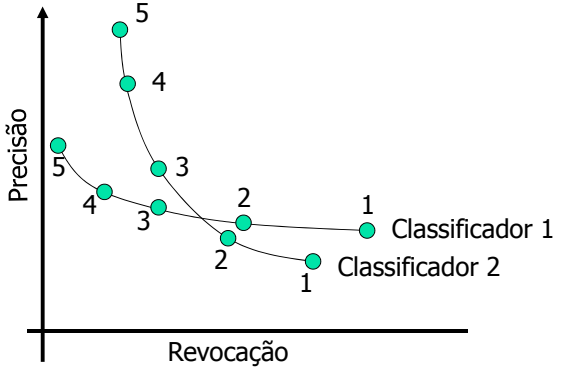
$$\text{Revocação} = \frac{VP}{VP+FN} = 70/(70+30) = 0.70$$

$$\text{Especificidade} = \frac{VN}{VN+FP} = 60/(40+60) = 0.60$$

Previsto \ Verdadeiro	P	N
P	VP	FP
N	FN	VN

Previsto \ Verdadeiro	P	N
P	70	40
N	30	60

Observação



Exercício

- Avaliar os 3 classificadores abaixo a partir de todas as medidas de desempenho de classificadores vistas :

Classe Prevista \ Classe Verdadeira	P	N
P	25	10
N	45	60

Classe Prevista \ Classe Verdadeira	P	N
P	70	20
N	15	30

Classe Prevista \ Classe Verdadeira	P	N
P	70	95
N	30	5

Gráficos ROC

- Do inglês, Receiver Operating Characteristics
- Medida de desempenho originária da área de processamento de sinais
  - Muito utilizada na área médica
  - Mostra relação entre custo (taxa de FP) e benefício (taxa de VP)
    - Taxa de FP = Erro do Tipo I (alarmes falsos)
    - Taxa de VP (Recall, Sensibilidade) = 1 – Erro do Tipo II

Exemplo

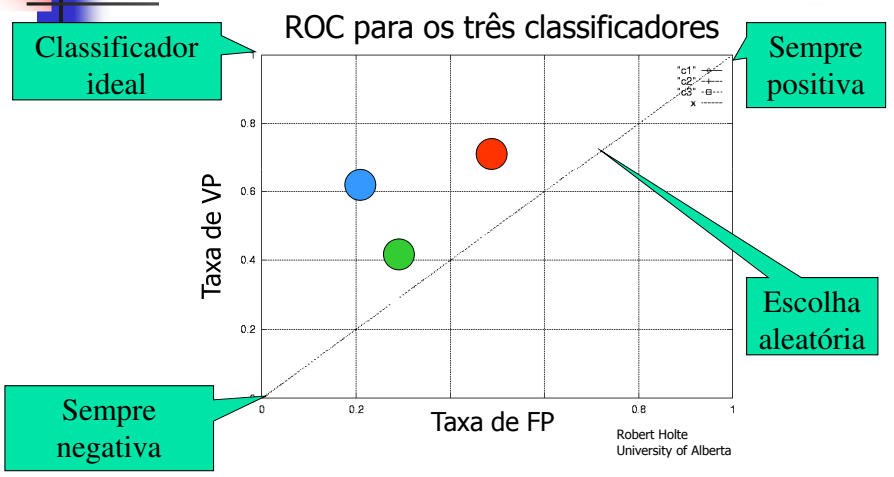
- Plotar no gráfico ROC os 3 classificadores do exemplo anterior

Classificador 1
TVP = 0.4
TFP = 0.3

Classificador2
TVP = 0.7
TFP = 0.5

Classificador 3
TVP = 0.6
TFP = 0.2

Gráficos ROC



Gráficos ROC

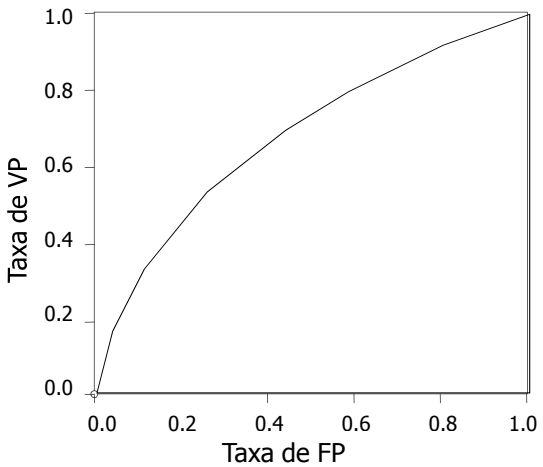
- Informalmente, melhor classificador é aquele cujo ponto está mais a noroeste
  - Classificadores próximos do canto inferior esquerdo são conservadores
    - Só fazem classificações positivas com forte evidência
      - Assim, cometem poucos erros de FP
  - Classificadores próximos ao canto superior direito são liberais (sob risco de alarme falso)

Curvas ROC

- Classificadores que geram **escores**:
  - Diferentes valores de **limiar** para os scores associados à classe Positiva podem ser utilizados para gerar um classificador
    - Cada valor produz um classificador diferente
      - Corresponde a um ponto diferente no gráfico ROC
    - Ligação dos pontos gera uma **Curva ROC**

Curvas ROC

Exemplo:

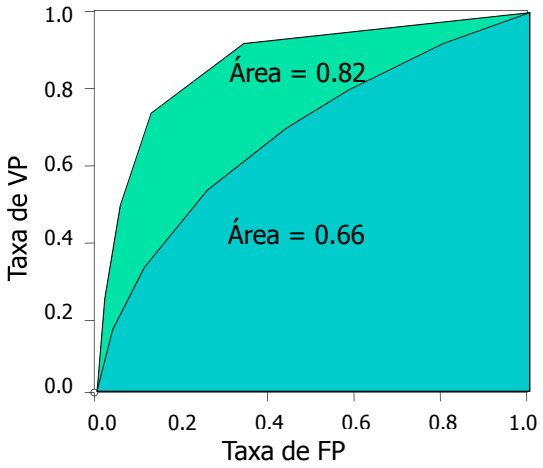


Área Sob a Curva ROC (AUC)

- Medida de desempenho de classificadores
- Gera um valor contínuo no intervalo [0,1]
  - Quanto maior melhor
  - Não deve ser vista como um critério absoluto
    - Deve ser vista como uma medida auxiliar às anteriores !
- Mais confiável: valor médio para cross-validation

Área Sob Curvas ROC

Exemplos:



Classes Difíceis

- Alguns problemas de classificação são caracterizados por possuírem classes difíceis de serem aprendidas por um classificador
- Duas das principais razões são:
  - Distribuição espacial complexa no espaço dos atributos
  - Classes desbalanceadas
    - Classes raras

Classes Desbalanceadas

- No. de exemplos varia para as diferentes classes
  - Natural ao domínio; ou
  - Problema com geração / coleta de dados
- Várias técnicas de DM não conseguem ou têm dificuldade para lidar com esse problema
  - Tendência a classificar na(s) classe(s) majoritária(s)

Classes Difíceis / Desbalanceadas

- Alternativa mais simples:
  - **Balanceamento Artificial**
    - sobre-amostragem
    - sub-amostragem
    - híbrido

Sobre-Amostragem

- Sobre-amostragem (**oversampling**) é uma técnica de balanceamento artificial dos dados
  - Consiste em aumentar artificialmente os exemplos da classe minoritária (classe positiva) até que os dados de treinamento estejam balanceados
  - Duas Abordagens:
    - Replicação
    - Repovoamento
  - Pode potencializar ruído e risco de overfitting

Sobre-Amostragem

- Sobre-amostragem (**oversampling**) é uma técnica de balanceamento artificial dos dados
  - **Replicação:**
    - Não insere informação nova, apenas aumenta a representatividade de padrões já existentes, fazendo com que esses sejam mais significativos para o algoritmo
  - **Repovoamento:**
    - Cria padrões novos intermediários aos padrões já existentes e seus k vizinhos mais próximos. Logo, insere informação nova, porém artificial ...

Sub-Amostragem

- Sub-amostragem (**undersampling**) é uma técnica de balanceamento artificial dos dados
  - Diminui artificialmente os exemplos da classe majoritária (negativa) até que dados de treinamento estejam balanceados
  - Pode descartar informação útil sobre a classe majoritária, especialmente se houver apenas um no. muito pequeno de exemplos da minoritária. Solução:
    - Repetir amostragem várias vezes; ou
    - Fazer amostragem informada
      - Desprivilegiar casos seguros; privilegiar exemplos de fronteira

Amostragem Híbrida

- Amostragem híbrida mescla oversampling e undersampling para amenizar os possíveis problemas de cada abordagem