APRENDIZADO DE MÁQUINA - AM

Classificação: Problemas com Classes Difíceis e Avaliação de Desempenho de Classificadores

Tópicos

- Avaliação de Classificadores
  - Procedimentos de Teste e Validação
    - Holdout, Random Subsampling, Cross-Validation e Bootstrap
  - Matriz de Confusão, Curvas ROC e Métricas de Avaliação
- Problemas de Classificação com Classes Difíceis
  - Misturas de Classificadores (ensembles)
    - Bagging, Boosting e Outros
  - Técnicas Alternativas para Classes Desbalanceadas
    - Balanceamento por Sub- e/ou Sobre-Amostragem
    - Aprendizado com Custos Distintos Associados às Classes

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
  - Holdout
  - Random Subsampling
  - Cross-Validation
    - Leave-One-Out
  - Bootstrap

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

- Indicado para grandes quantidades de dados
- Se aplicado em pequena quantidade de dados...
  - Poucos exemplos são usados no treinamento
  - Modelo pode depender da composição dos conjuntos 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

Métodos de Re-Amostragem

- Utilizam várias partições do conjunto original de dados para constituir os conjuntos de treinamento e teste
  - Random subsampling
  - Cross-validation
    - Leave-one-out
  - Bootstrap

Random Subsampling

- Múltiplas execuções de Holdout
  - Diferentes partições treinamento-teste são escolhidas de forma aleatória
    - Não pode haver interseção entre os dois conjuntos
    - Taxa de erro de classificação é calculada para cada partição
    - Erro de classificação estimado para dados não vistos é a média dos erros 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: x1, x2, x3, x4, x5, x6, x7, x8
    - 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 validação 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
- **Leave-one-out** (LOO)
  - Caso particular com k = N
    - onde N = no. de exemplos

Leave-One-Out

- N pastas são utilizadas para uma amostra de tamanho N
  - **N-fold cross-validation**
  - A cada iteração, um dos exemplos é utilizado para testar o modelo
    - Os outros N – 1 exemplos são utilizados para o treinamento
  - Taxa de erro é obtida dividindo por N o número total de erros de validação observados

Leave-One-Out

- Sua estimativa de erro é não tendenciosa
  - Média das estimativas tende à taxa verdadeira
- Porém, é computacionalmente caro
- Recomendado quando se dispõe de uma quantidade relativamente pequena de exemplos
  - Custo computacional pode ser viável
  - Usa-se quase todos os exemplos no treinamento
- 10-fold cross validation aproxima leave-one-out

5 x 2 Cross-Validation

- Conjuntos de treinamento e teste com mesmo tamanho
- Dietterich, 1998

Seja um conjunto de N exemplos
Para i = 1 até 5
    Dividir N aleatoriamente em duas metades
    Usar metade 1 para treinamento e metade 2 para teste
    Usar metade 2 para treinamento e metade 1 para teste

5 x 2 Cross-Validation

- Mais que 5 folds
  - Sobreposição dos conjuntos se torna tão grande que não adiciona nova informação
- Menos que 5 folds
  - Não haverá exemplos suficientes para ajustar uma distribuição e testar hipóteses
    - Menos que 10 partições

Bootstrap

- Funciona melhor que cross-validation para conjuntos muito pequenos
- Forma mais simples de bootstrap:
  - Ao invés de usar sub-conjuntos dos dados, usa sub-amostras
    - Cada sub-amostra é retirada com reposição até a substituição do conjunto total de exemplos
      - Cada sub-amostra tem o mesmo no. de exemplos do conjunto original e é utilizada para treinamento
    - Os exemplos que restam são utilizados para teste

Bootstrap

- Se conjunto original tem N exemplos
  - Amostra de tamanho N tende a ter ~=63,2% dos exemplos originais (demais ~= 36,8% são réplicas)
- Processo é repetido b vezes
  - Resultado final = média dos b experimentos
- Existem diversas variações
  - Por exemplo, para estimar o erro do classificador
    - **.632 bootstrap**
    - ...

.632 Bootstrap

- Leva em conta que há interseção entre as b amostras de teste, que envolvem apenas ~=36,8% dos exemplos cada
- Para estimar o erro do classificador, combina:
  - Acurácia de cada uma das b amostras (acc_i) com
  - Acurácia para o conjunto de treinamento que contém todos os dados originais (acc_t)

$$Acuracia = \frac{1}{b}\sum_{i=1}^{b}(0.632 \times acc_i + 0.368 \times acc_t)$$

Estimativa de Erro de Classificação

- Principal objetivo de um modelo é predizer com sucesso o valor de saída para novos exemplos
  - Errar o mínimo possível
- Geralmente não é possível medir com exatidão o erro do modelo para qualquer entrada
  - Sua taxa de erro deve ser estimada em um conjunto de exemplos não vistos durante o treinamento

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

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

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

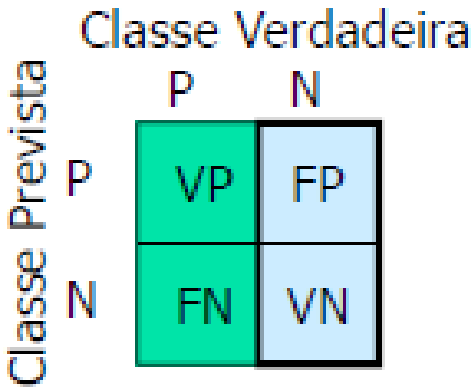
Avaliação de Desempenho

- Medidas de erro

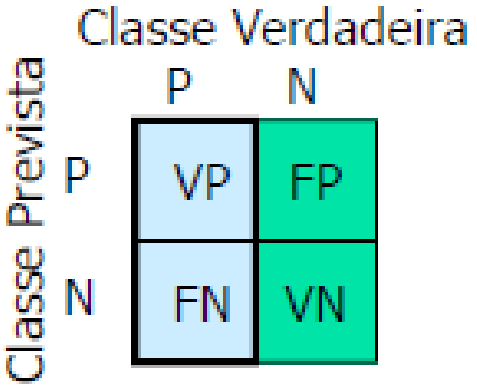
$$\textbf{Taxa de FP} = \frac{FP}{FP+VN}$$
(alarmes falsos)

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

Erro do tipo I

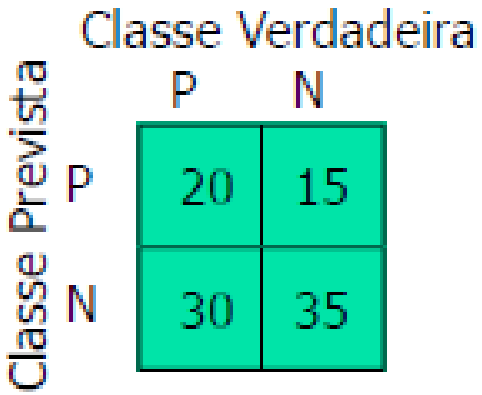


Erro do tipo II

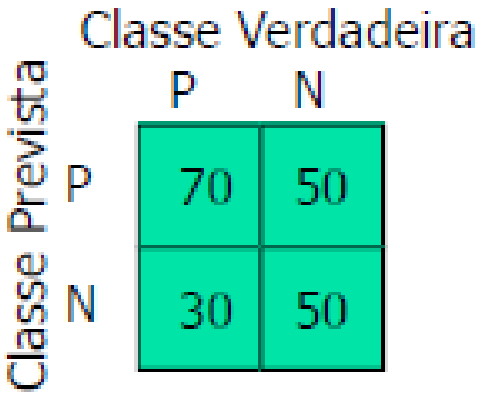


Exemplo

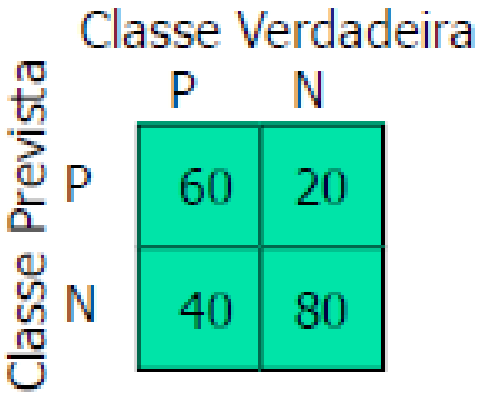
- Avaliação de 3 classificadores



Classificador 1
TFN =
TFP =



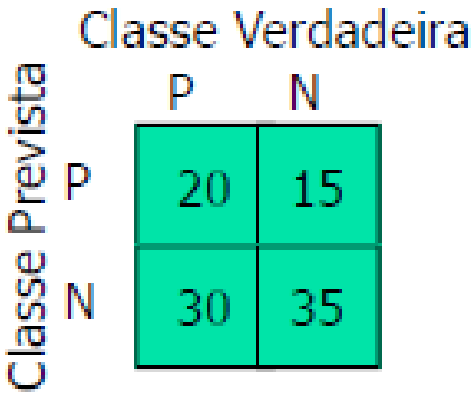
Classificador 2
TFN =
TFP =



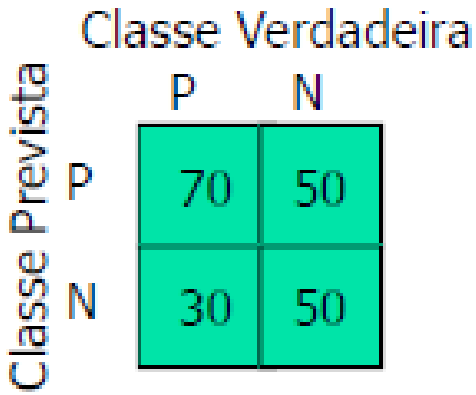
Classificador 3
TFN =
TFP =

Exemplo

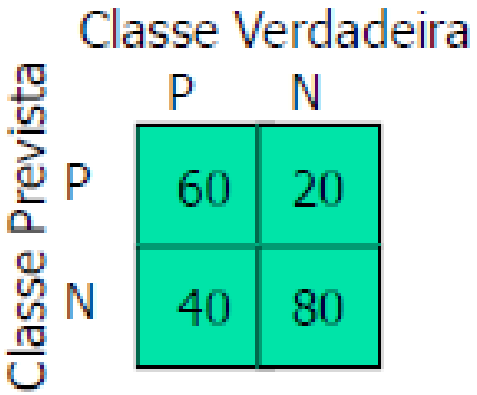
- Avaliação de 3 classificadores



Classificador 1
TFN = 0.6
TFP = 0.3



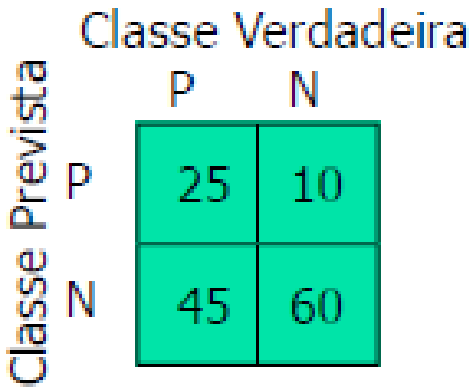
Classificador 2
TFN = 0.3
TFP = 0.5



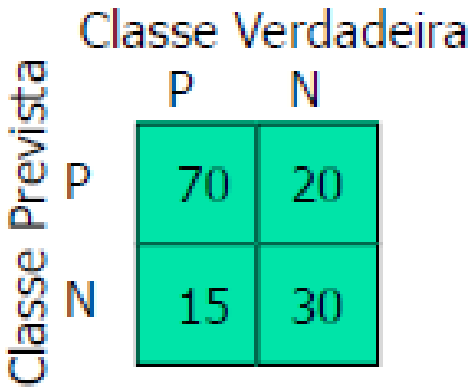
Classificador 3
TFN = 0.4
TFP = 0.2

Exercício

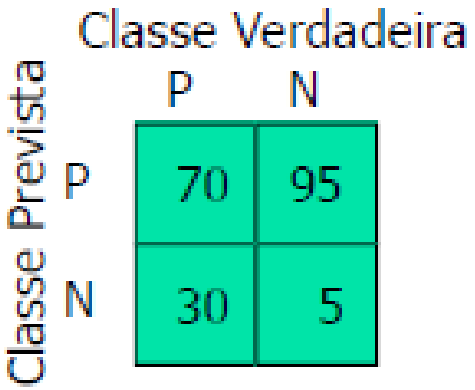
- Avaliar os 3 classificadores abaixo:



Classificador 1
TFN =
TFP =



Classificador2
TFN =
TFP =



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

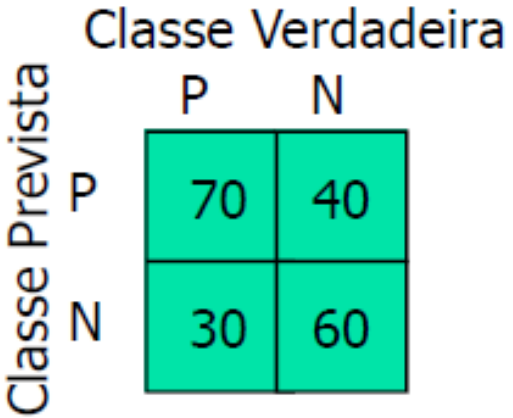
$$\frac{(1+\alpha)\times(prec\times rev)}{\alpha\times prec+rev}$$

- **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, definir:
  - Acurácia
  - Precisão
  - Revocação (sensibilidade)
  - Especificidade



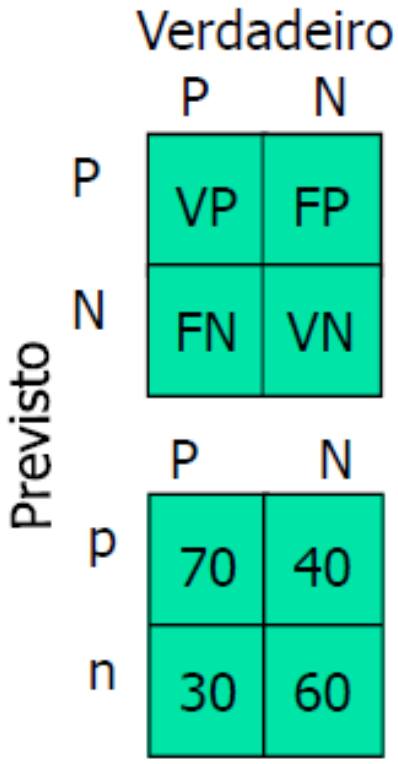
Exemplo

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

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

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

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



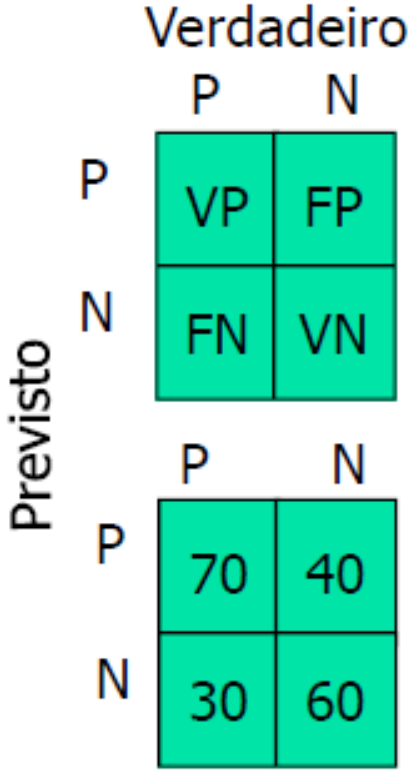
Exemplo

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

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

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

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



Precisão x Revocação

- A precisão é uma medida de fidelidade
- A revocação (também conhecida como cobertura ou sensibilidade) é uma medida de completude.
- No contexto de recuperação de informação:
  - a precisão é o número de elementos relevantes recuperados divididos pelo número total de elementos recuperados

$$Precisão = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ recuperados}$$

  - a revocação é definida como o número de elementos relevantes recuperados dividido pelo número total de elementos relevantes existentes (que deveriam ter sido recuperados)

$$Revocação = \frac{Número\ de\ elementos\ relevantes\ recuperados}{Número\ total\ de\ elementos\ relevantes}$$

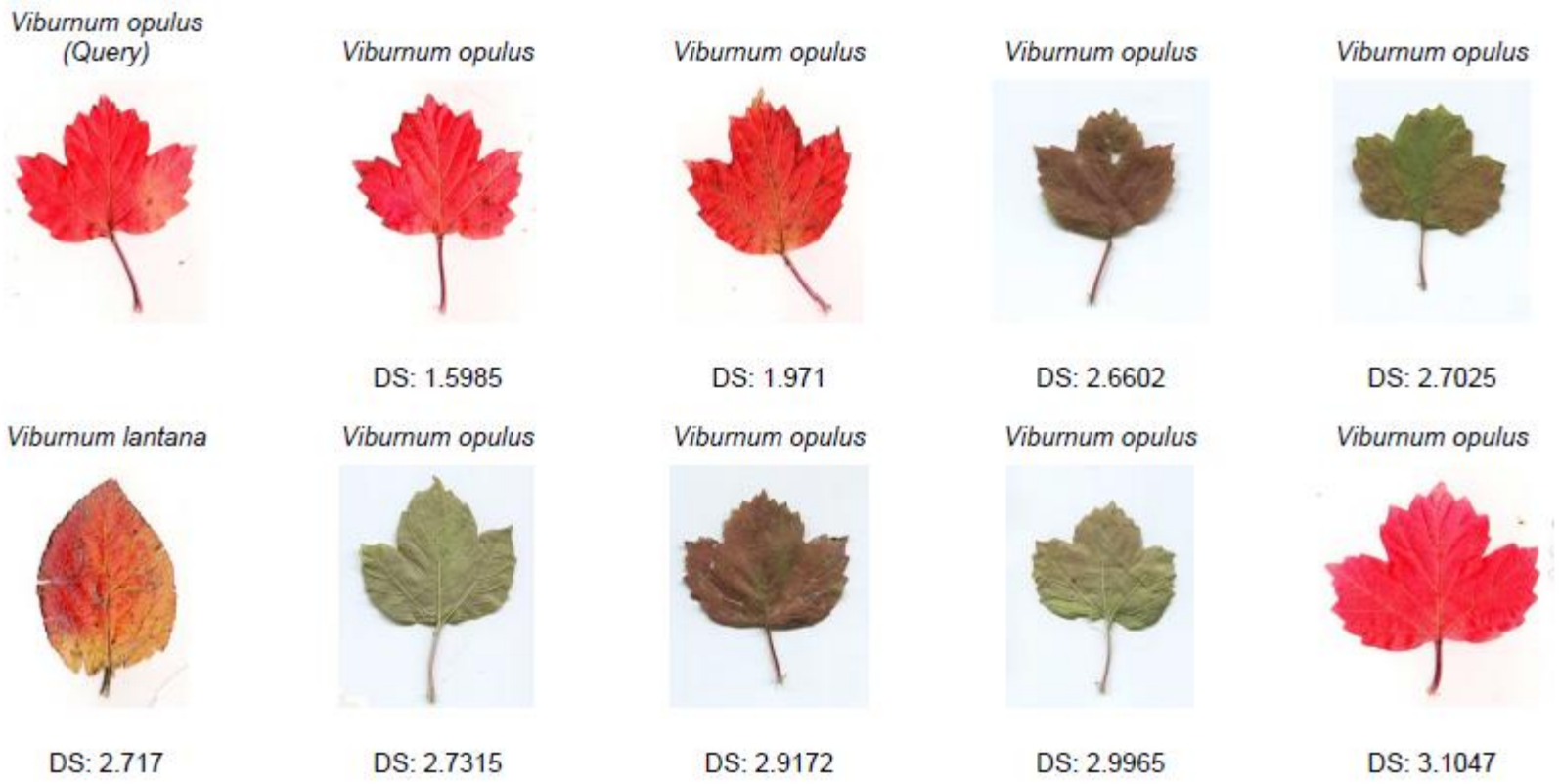
Precisão x Revocação

- O conceito de precisão x revocação é bastante utilizado no contexto de recuperação de imagens baseados no conteúdo
  - CBIR (do inglês Content Based Image Retrieval)

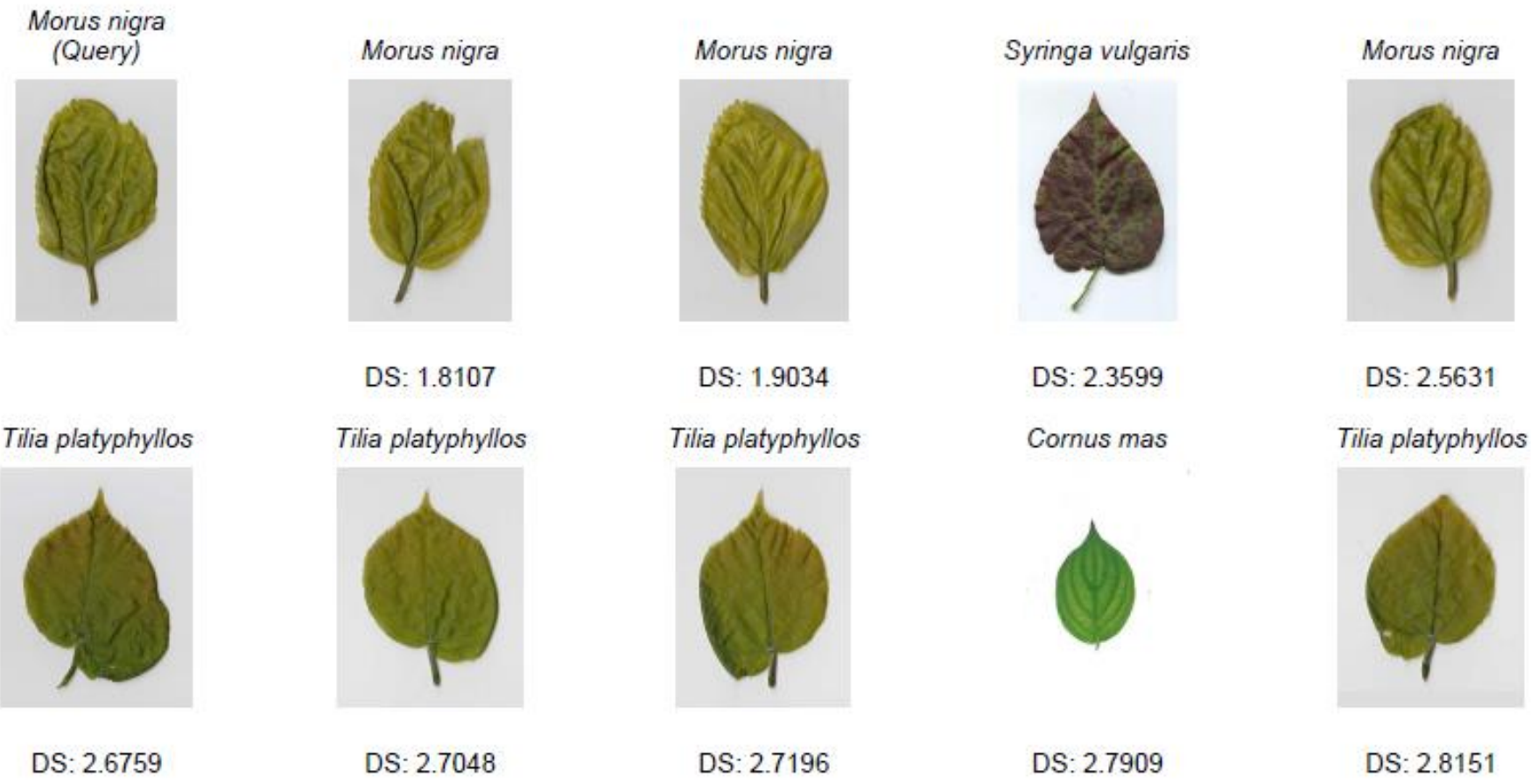
- A **precisão** mede a fração de folhas relevantes recuperadas (folhas da mesma espécie da consulta) pelo número de folhas recuperadas

- A **revocação** mede a fração de folhas relevantes pelo total de folhas relevantes existentes na base de dados.

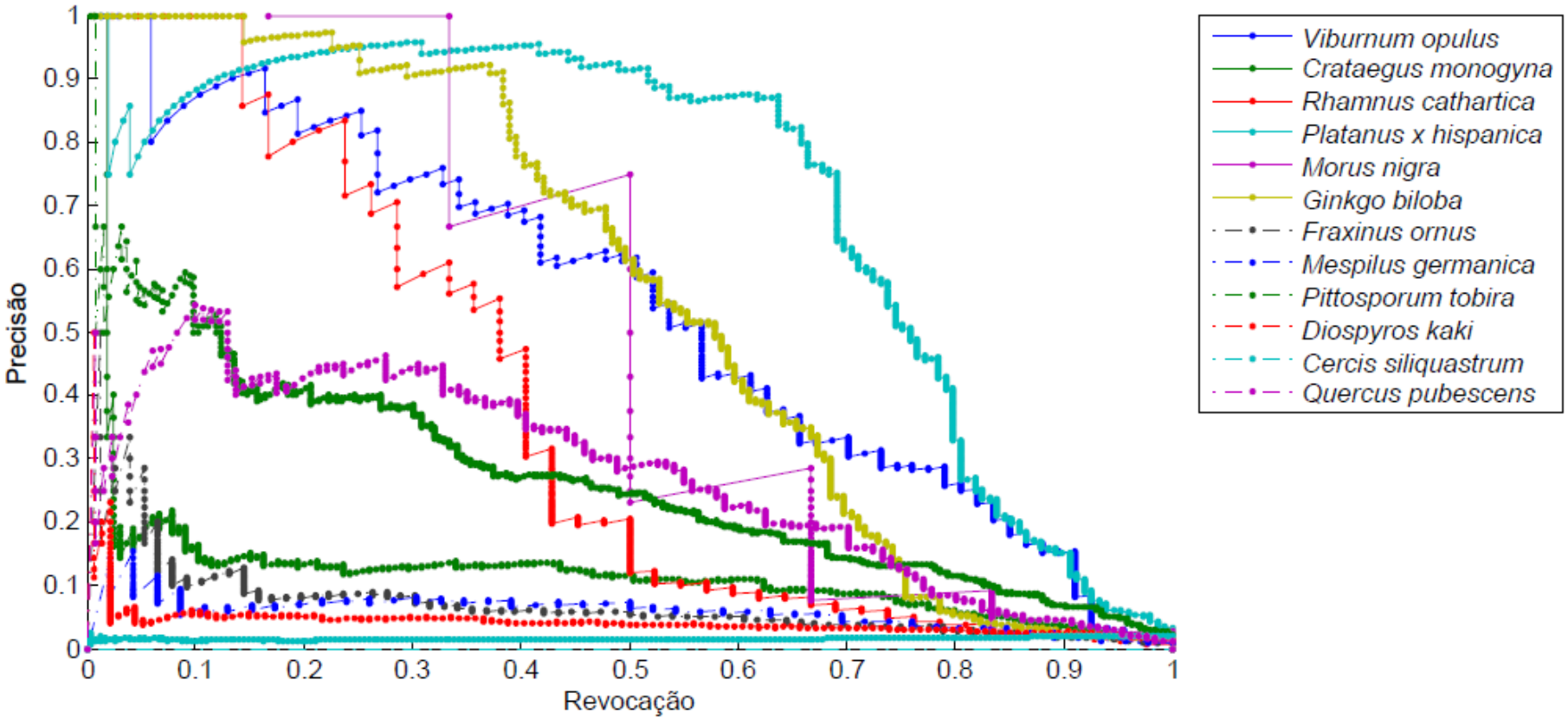
Precisão x Revocação



Precisão x Revocação



Precisão x Revocação

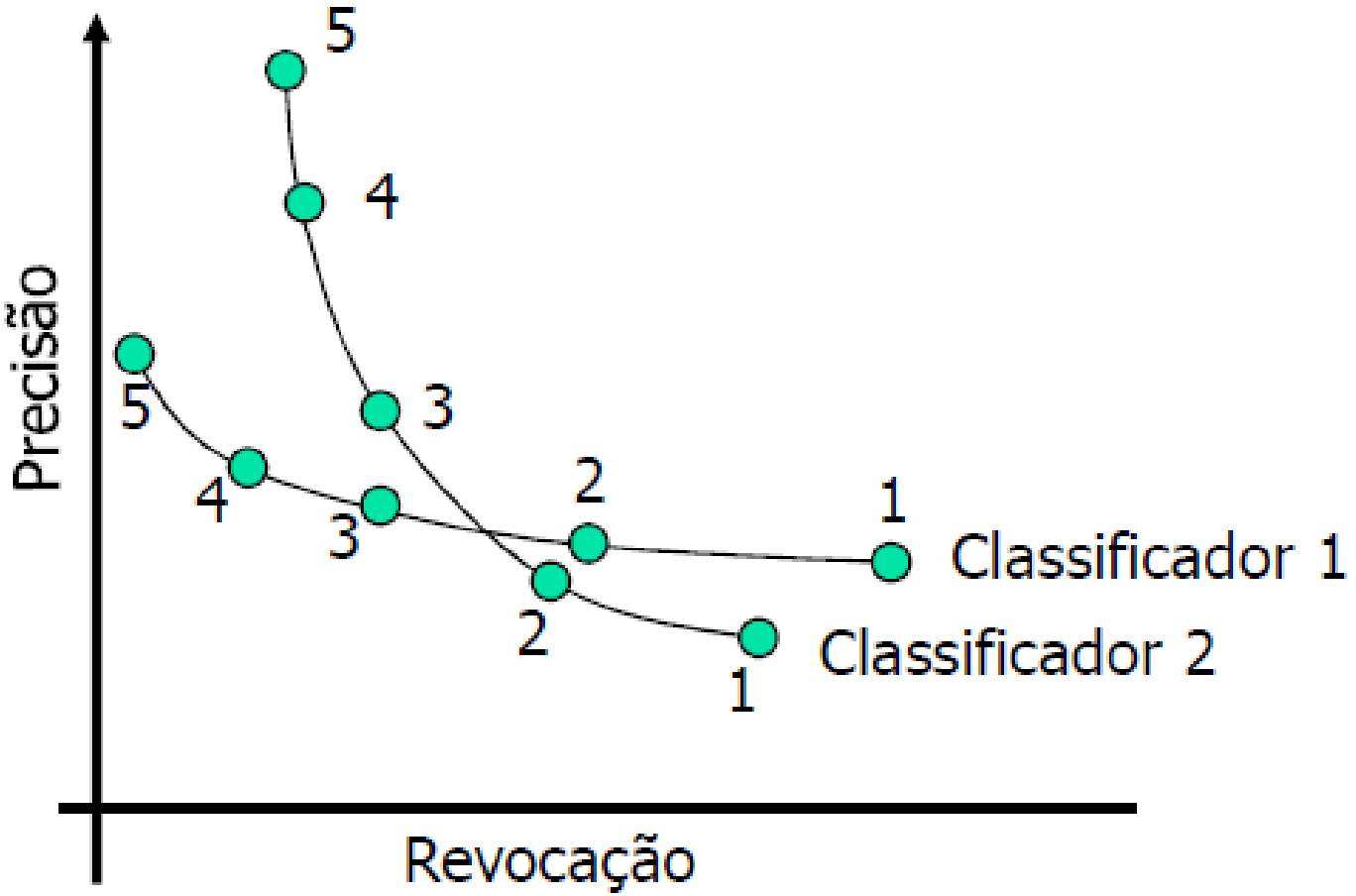


Precisão x Revocação

- Para a precisão o valor de 1 significa que cada resultado obtido por uma pesquisa foi relevante (mas não diz nada sobre se todos os elementos relevantes foram recuperados), enquanto
- O valor 1 para revocação significa que todos os elementos relevantes foram recuperados pela pesquisa (mas nada diz sobre quantos elementos irrelevantes também foram recuperados).

- Por vezes pode existir uma relação inversa entre precisão e revocação, onde é possível aumentar uma ao custo de reduzir outra.
  - Pode-se, por exemplo, aumentar a revocação recuperando mais elementos, ao custo de um número crescente de elementos irrelevantes recuperados (diminuindo a precisão).

Observação

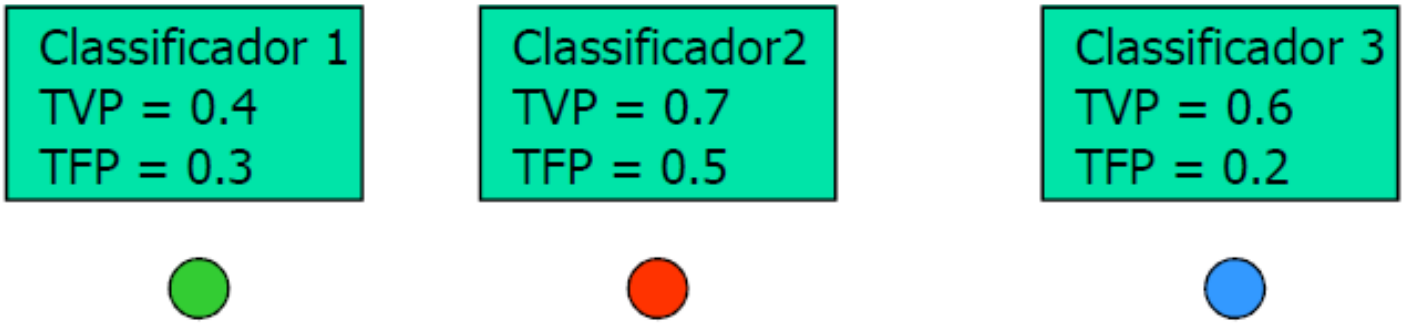

Precisão
Revocação
5
4
3
2
1
5
4
3
2
1
Classificador 1
Classificador 2

Gráficos ROC

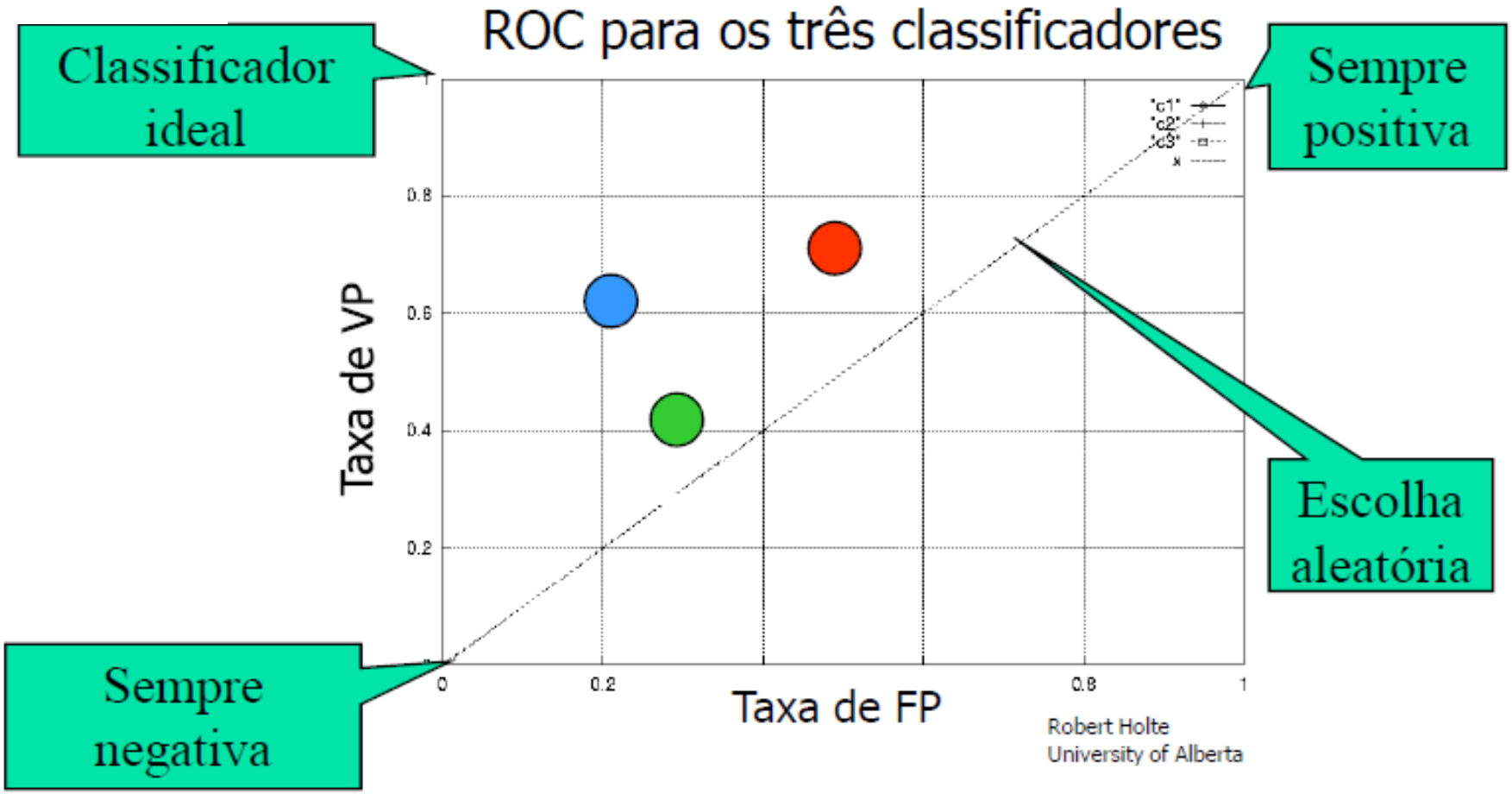
- Do inglês, Receiver Operating Characteristics
- Medida de desempenho originária da área de processamento de sinais
  - Muito utilizada na área médica
  - Mostra relação entre custo (taxa de FP) e benefício (taxa de VP)
    - Taxa de FP = Erro do Tipo I (alarmes falsos)
    - Taxa de VP (Recall, Sensibilidade)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

Exemplo

- Plotar no gráfico ROC os 3 classificadores do exemplo anterior



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)

Gráficos ROC

- Alguns classificadores produzem saídas discretas
  - Por exemplo, ADs e SVMs
    - Atribuem cada exemplo a uma das classes
  - Produzem um **ponto simples no gráfico ROC**

- Outros classificadores produzem como saída um escore (e.g. uma probabilidade) associado a cada classe
  - Por exemplo, Naive Bayes e RNAs
  - Permitem gerar uma **curva no gráfico ROC**

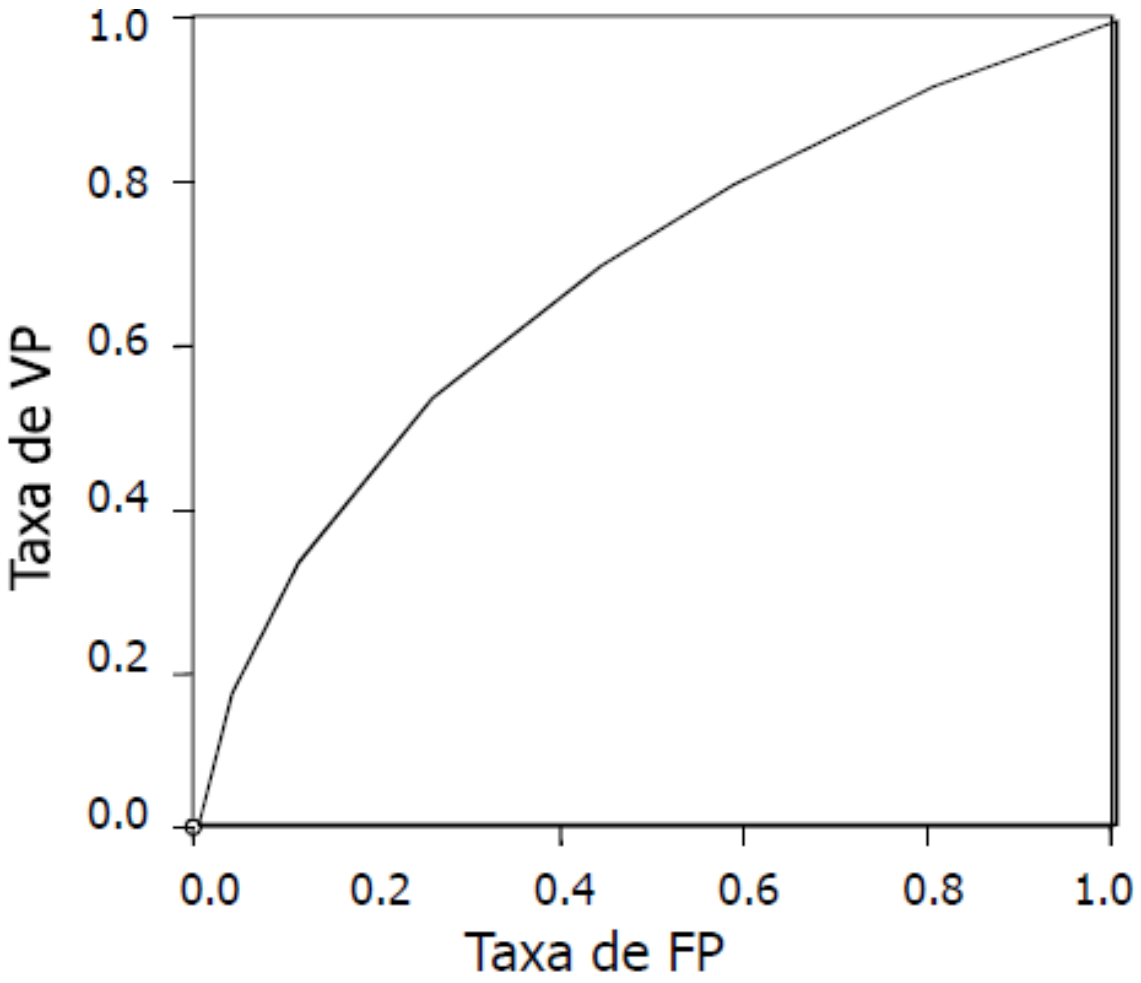
- Curvas ROC permitem uma melhor comparação de classificadores
  - São insensíveis a mudanças na distribuição das classes

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

Classificador
(Escore)



Curvas ROC

Instância	Classe V.	Score P
6	P	0.9
3	P	0.8
2	N	0.7
9	P	0.6
5	P	0.6
1	N	0.5
7	N	0.3
8	N	0.2
4	N	0.2
10	N	0.1

- Ordenar exemplos em ordem decrescente por valor de predição (score) para a classe Positiva (P)
- Para cada limiar de decisão dado por cada valor de score:
  - Classificar todos os padrões
  - Calcular VP, VN, FP, FN
  - Calcular TVP e TFP

$$Classe = \begin{cases} predição \geq \theta, \mathrm{P} \\ predição < \theta, \mathrm{N} \end{cases}$$

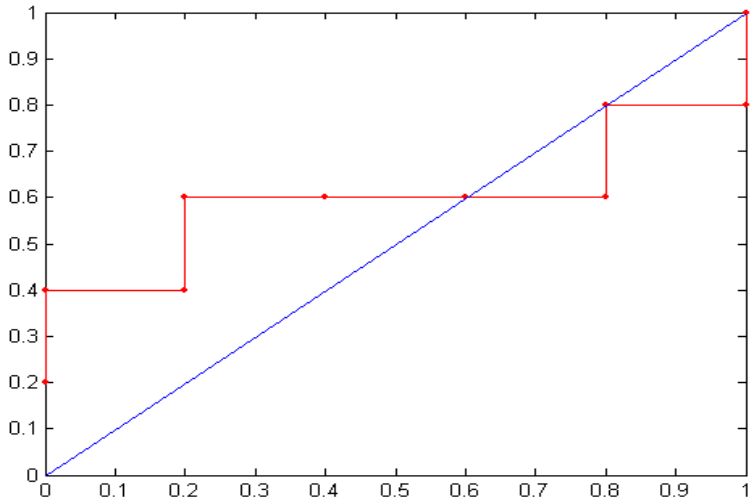
How to Construct an ROC curve

Instance	Pr(+|x)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

- Use classifier that produces posterior probability Pr(+|**x**) for each test instance **x**
- Sort the instances according to Pr(+|**x**) in decreasing order
- Apply threshold at each unique value of Pr(+|**x**)
- Count the number of TP, FP, TN, FN at each threshold
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

Curvas ROC

Classe Verdadeira	+	-	+	-	-	-	+	-	+	+	
Score +	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
VP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
VN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TVP	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
TFP	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0



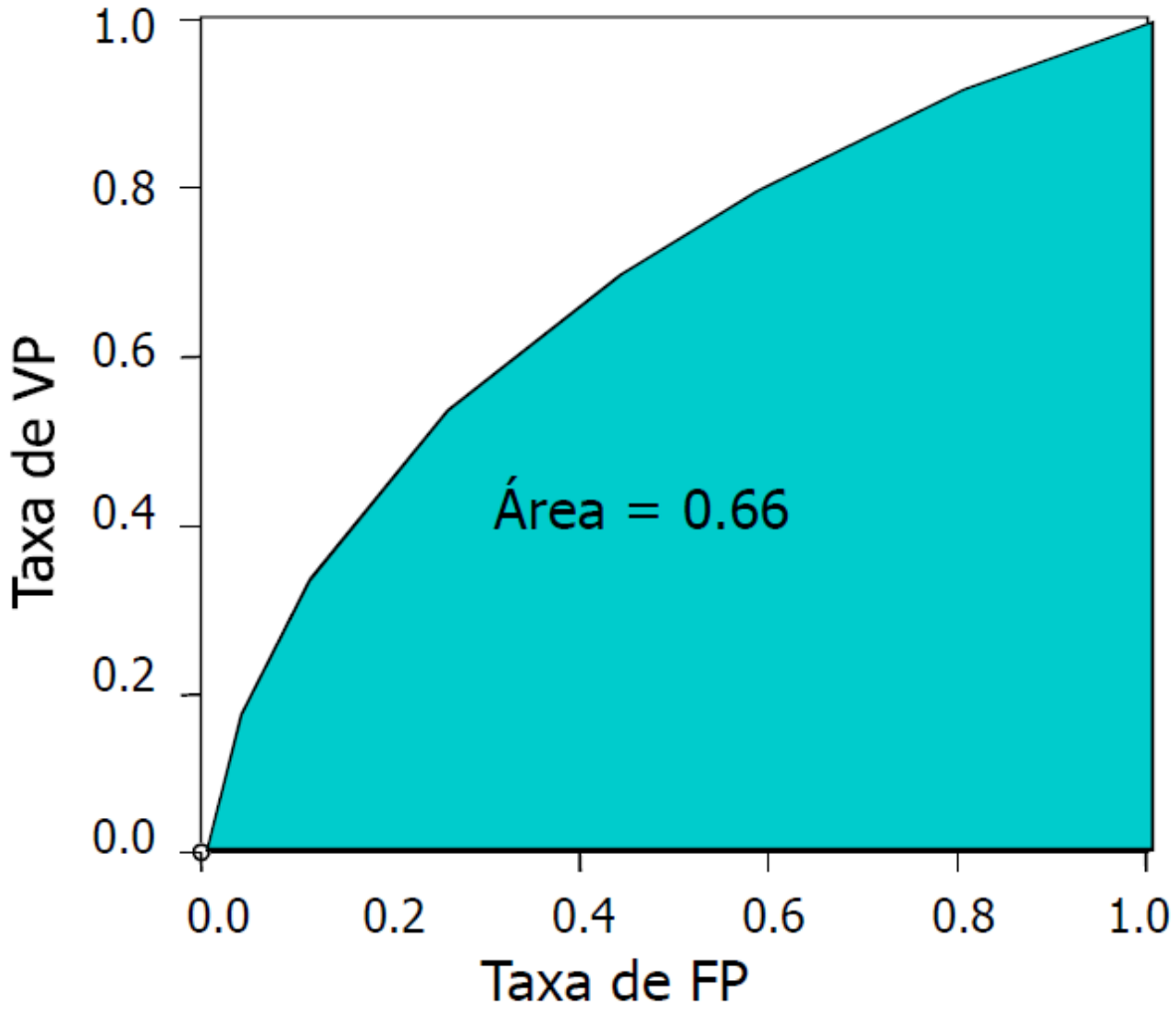
Curvas ROC

- Classificadores que geram valores discretos
  - Podem ser modificados para gerar escores
    - Para ADs, nós folhas podem conter a fração de exemplos de treinamento positivos
    - Para SVMs, saída pode ser distância do exemplo ao limiar de decisão (hiperplano separador)
    - Para K-NN, saída pode ser a fração dos k-vizinhos mais próximos que pertencem à classe Positiva
    - ...

Área Sob a Curva ROC (AUC)

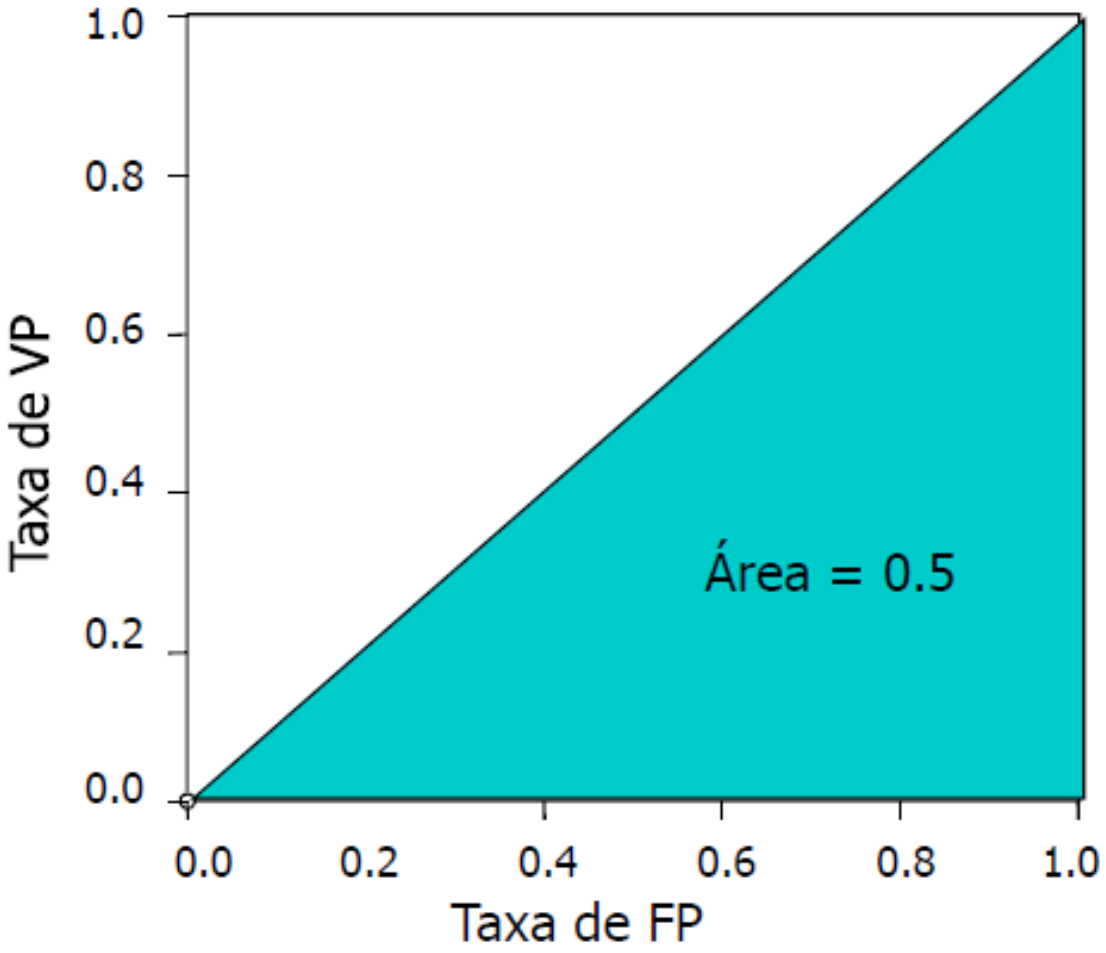
- Estimativa do desempenho de classificadores
- Gera um valor contínuo no intervalo [0,1]
  - Quanto maior melhor
  - Adição de áreas de sucessivos trapezóides
    - Possível provar que equivale à probabilidade do classificador atribuir um score Pr(+|x) maior a um exemplo x positivo (classe +) escolhido aleatoriamente que a um exemplo x negativo escolhido aleatoriamente

Área Sob Curvas ROC



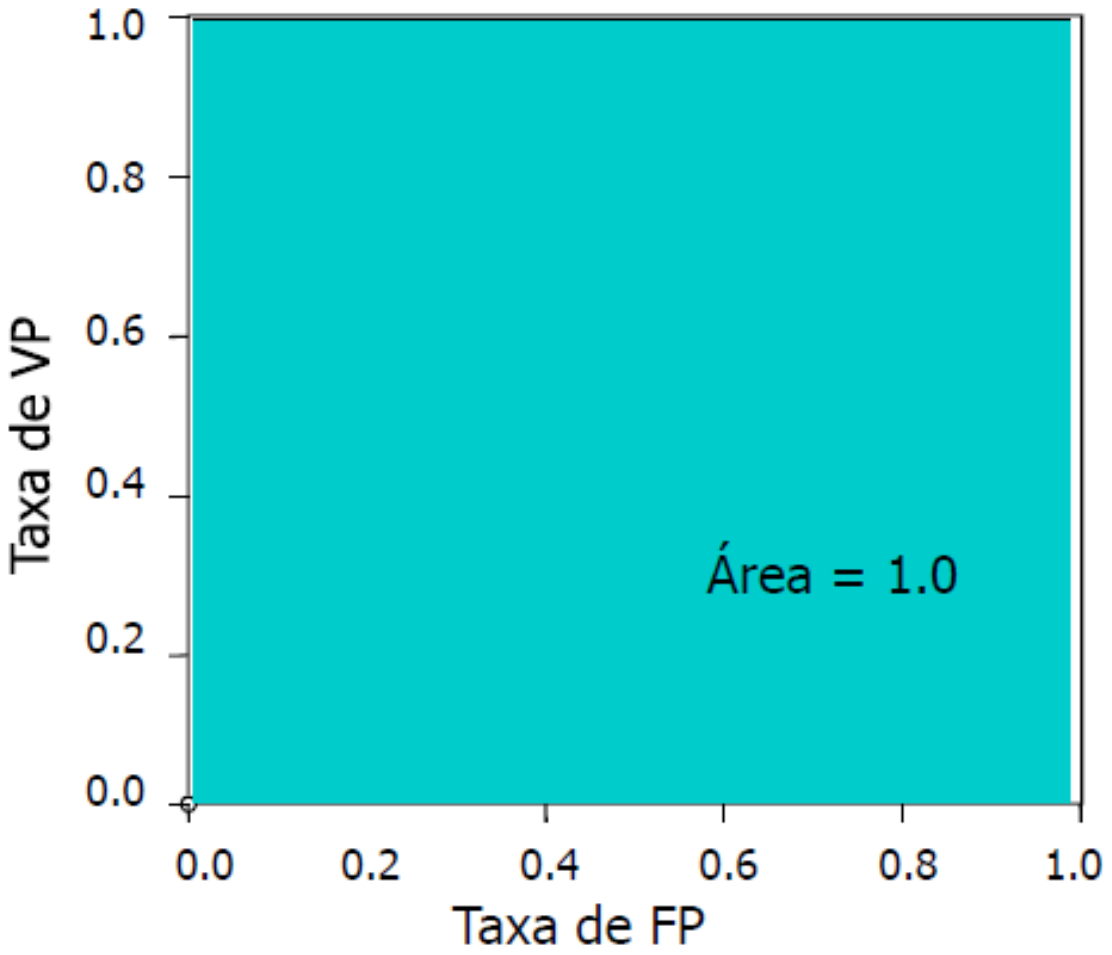
Área Sob Curvas ROC

Nenhuma
Discriminação

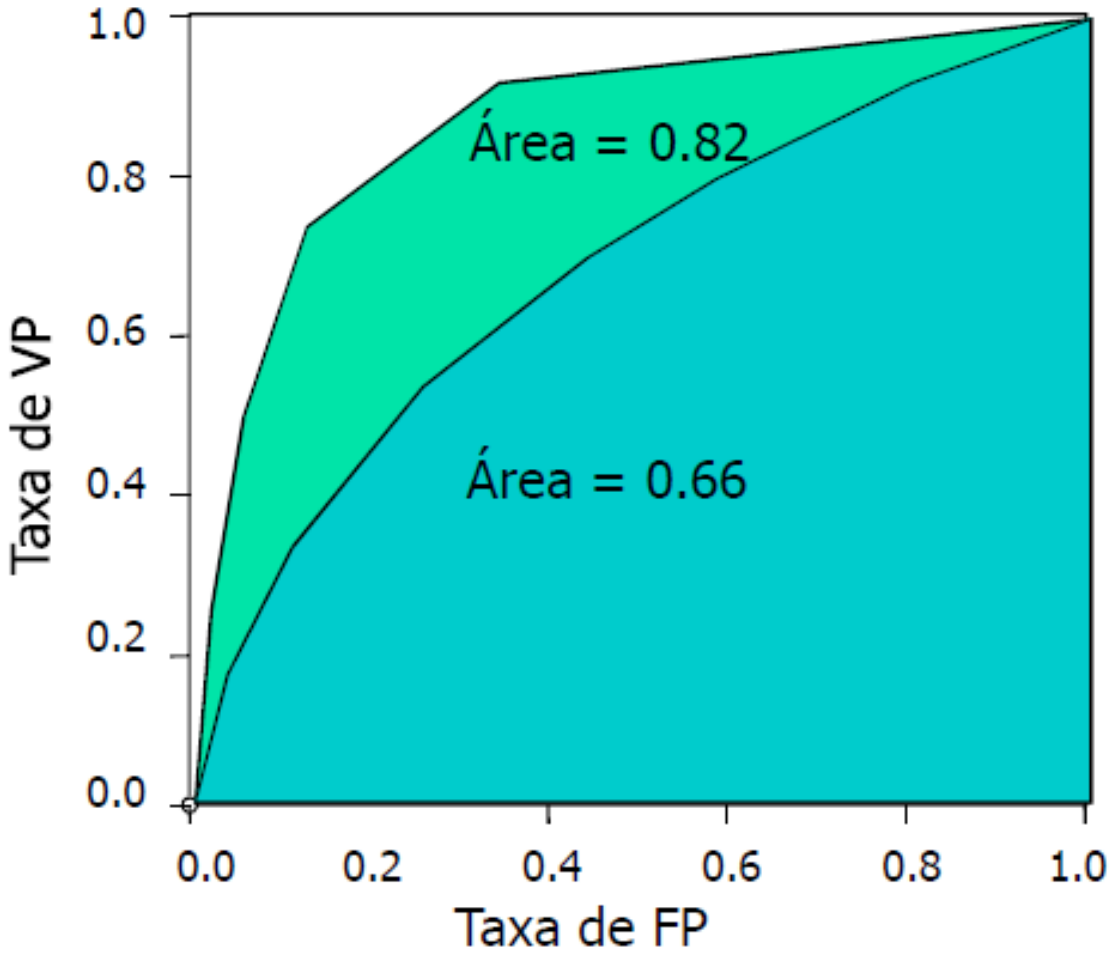


Área Sob Curvas ROC

Discriminação Perfeita



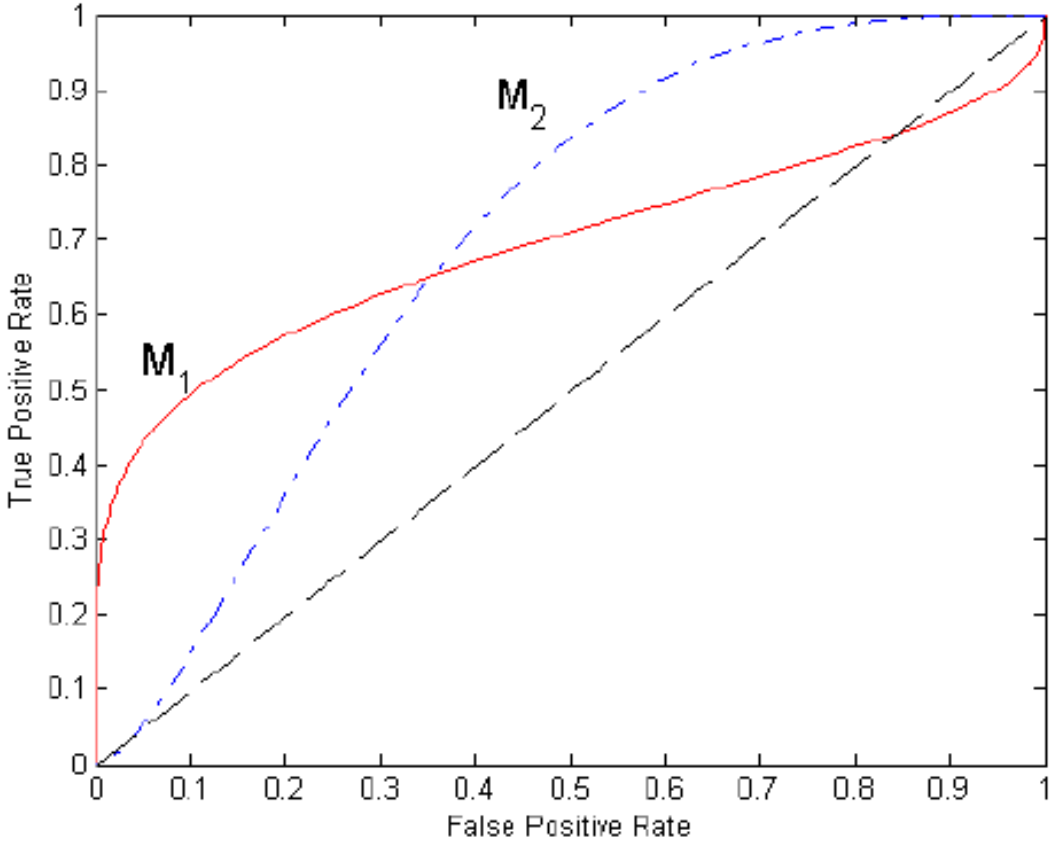
Área Sob Curvas ROC



Área Sob a Curva ROC

- Nota: um classificador com maior AUC pode apresentar AUC pior em trechos da curva...
  - AUC não deve ser vista como um critério absoluto
  - Deve ser vista como uma medida de desempenho auxiliar às demais vistas anteriormente!

Using ROC for Model Comparison



- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

Área Sob a Curva ROC

- Nota:
  - Para maior confiabilidade da análise, usualmente utiliza-se algum dos procedimentos de avaliação de desempenho vistos anteriormente (e.g. validação cruzada) para gerar múltiplas curvas ROC
    - AUC mais confiável é tomada a partir de uma curva gerada a partir de algum tipo **de média das curvas ROC**
    - **Variância** das curvas é um outro fator de avaliação

Análise ROC

- Cuidado !
  - Curva ROC ideal (AUC = 1) não é obtida apenas por classificador que discrimine perfeitamente as classes
  - Qualquer classificador que produza scores maiores para os exemplos positivos que para os exemplos negativos (sem exceção) possui AUC = 1!!!
    - Exercício: Pense e explique o porquê !

Análise ROC

- Nota:
  - Distribuição das classes é dada pela proporção entre os valores da 1ª e 2ª colunas da matriz de confusão
  - Ao contrário de outras medidas, o gráfico ROC não se modifica com alterações nessa proporção
    - Sendo taxas, a Taxa de VP e a Taxa de FP são insensíveis às quantidades de exemplos P e N nas respectivas colunas da matriz
    - Logo, são insensíveis à distribuição das classes
      - (Des)balanceamento não afeta o gráfico ROC!
  - Quais medidas de desempenho vistas possuem essa propriedade ?

Análise ROC

- Nota:
  - Existem análises ROC para problemas multi-classes, porém são muito mais complexas que para problemas binários
  - Por exemplo, pode-se considerar as relações ROC existentes entre cada par de classes...
  - Ou considerar as relações ROC existentes entre cada classe e as demais classes
    - Uma classe vista como positiva e as demais como classe negativa

Comparações de Classificadores

- Métodos vistos até aqui permitem avaliar e portanto comparar o desempenho de dois ou mais classificadores em um mesmo conjunto de teste, de uma mesma base de dados
- Para uma avaliação mais confiável, com rigor estatístico, envolvendo diferentes conjuntos de teste e/ou bases de dados
  - vide próxima aula…

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 AM não conseguem ou têm dificuldade para lidar com esse problema
  - Tendência a classificar na(s) classe(s) majoritária(s)

Classes Difíceis / Desbalanceadas

- Principais Alternativas:
  - **Balanceamento Artificial**
    - sobre-amostragem, sub-amostragem, híbrido
  - **Classificação com Custos Associados**
  - **Classificação com 1 Classe** (1 Class Problem)
  - **Múltiplos Classificadores** (Ensembles)
    - bagging, boosting, random-forests, ...

Sobre-Amostragem

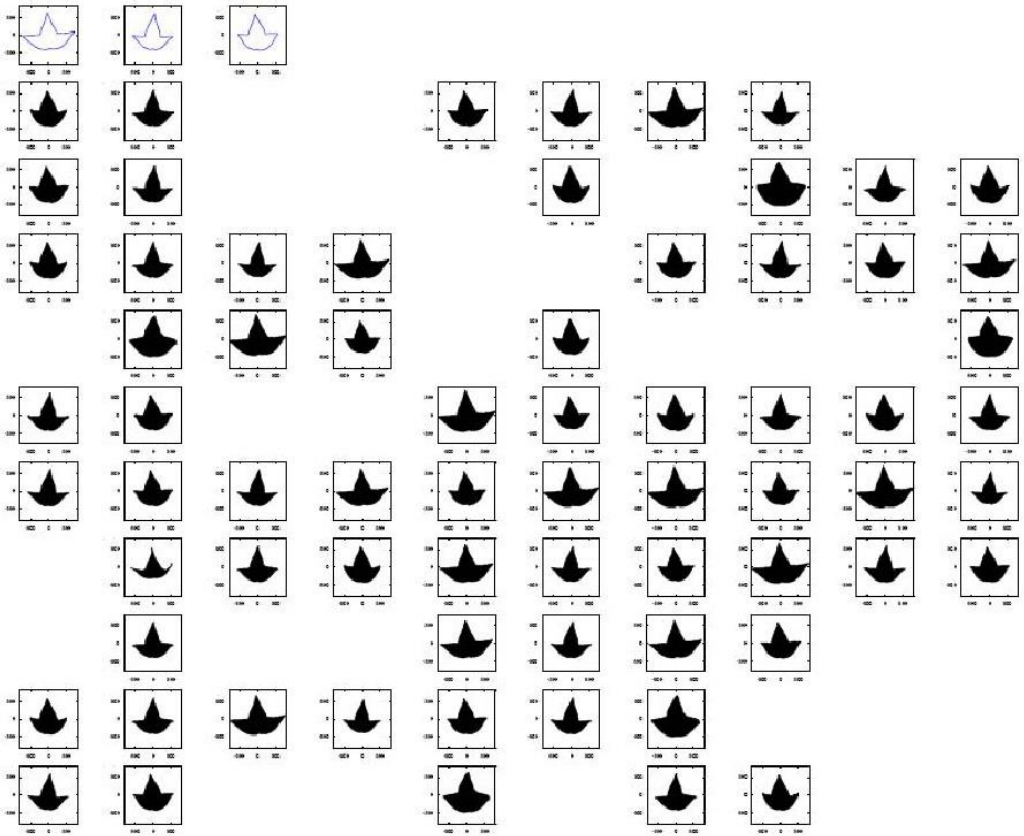
- Sobre-amostragem (**oversampling**) é uma técnica de balanceamento artificial dos dados
  - Consiste em aumentar artificialmente os exemplos da classe minoritária (positiva) até que os dados de treinamento estejam balanceados
  - Duas Abordagens:
    - Replicação
    - Repovoamento
  - Pode potencializar 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 de AM
  - Repovoamento:
    - Cria padrões novos intermediários aos padrões já existentes e seus k vizinhos mais próximos. Logo, insere informação nova, porém artificial ...

Sobre-Amostragem

Mauricio Falvo, Joao Batista Florindo, and Odemir Martinez Bruno. **A Method to Generate Artificial 2D Shape Contour Based in Fourier Transform and Genetic Algorithms**. Advanced Concepts for Intelligent Vision Systems,207-215,2011



Sub-Amostragem

- Sub-amostragem (**undersampling**) é uma técnica de balanceamento artificial dos dados
  - Consiste em diminuir artificialmente os exemplos da classe majoritária (negativa) até que os dados de treinamento estejam balanceados
  - Pode descartar informação útil sobre a classe majoritária, especialmente se houver apenas um número 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

Classificação com Custos Associados

Cost Matrix

		PREDICTED CLASS	
	C(i|j)	**Class=Yes**	**Class=No**
ACTUAL CLASS	**Class=Yes**	C(Yes|Yes)	C(No|Yes)
	Class=No	C(Yes|No)	C(No|No)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Confusion Matrix

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Accuracy = 80%

Cost = 3910

Confusion Matrix

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy = 90%

Cost = 4255

Cost vs Accuracy

Count	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

Cost	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	p	q
	Class=No	q	p

Cost is a linear function of Accuracy if

1. C(Yes|No)=C(No|Yes) = q
2. C(Yes|Yes)=C(No|No) = p

$N = a + b + c + d$

$Accuracy = (a + d)/N$

$$\begin{aligned} Cost &= p\,(a + d) + q\,(b + c) \\ &= p\,(a + d) + q\,(N - a - d) \\ &= q\,N - (q - p)(a + d) \\ &= N\,[q - (q - p) \times Accuracy] \end{aligned}$$

Aprendizado Sensível a Custo

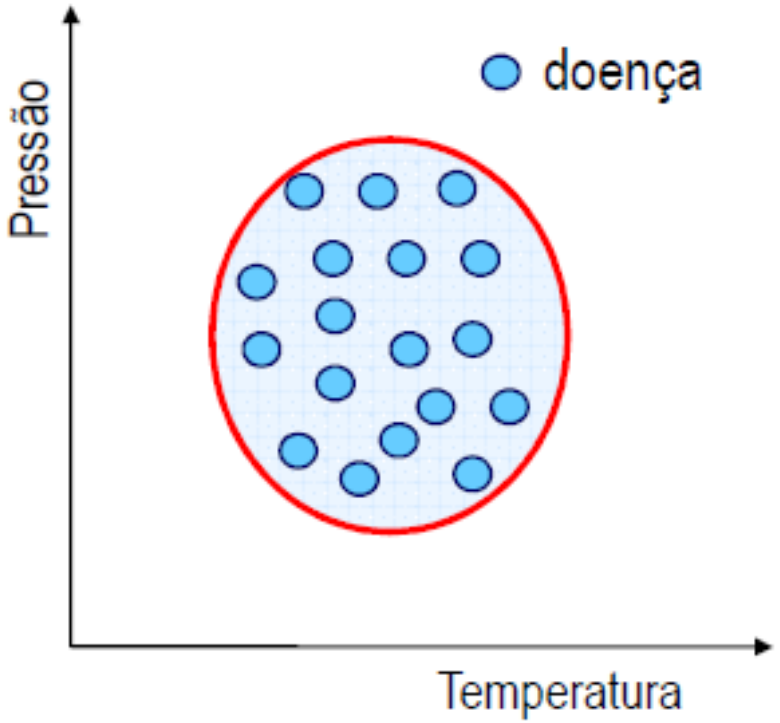
- Existem diferentes maneiras de incorporar custos em um sistema classificador
- Em árvores de decisão, por exemplo, custos podem ser incorporados de diferentes maneiras:
  - Nas medidas de escolha dos atributos
    - minimizar custos, não somente a impureza
  - Nos critérios para poda da árvore ou regras
  - Na determinação do limiar de decisão em cada folha

ADs com Custos Associados

- ADs com custos associados às classes via escolha do limiar de decisão nas folhas:
  - Uma das formas mais simples e intuitivas
  - Consiste em atribuir a cada nó folha o rótulo da classe com custo total mínimo, ao invés do rótulo da classe da maioria. Por exemplo:
    - Nó com 3 exemplos positivos com custo de classificação incorreta = 10 e 15 exemplos negativos com custo = 1
      - Rótulo da Maioria (negativo)=> Custo = 30
      - Rótulo de Custo Mínimo (positivo) => Custo = 15

Classificação com 1 Classe

- Classes raras
  - Obter dados positivos:
    - Difícil; e/ou
    - Custoso
  - Por exemplo:
    - Mulheres grávidas de gêmeos



Classificação com 1 Classe

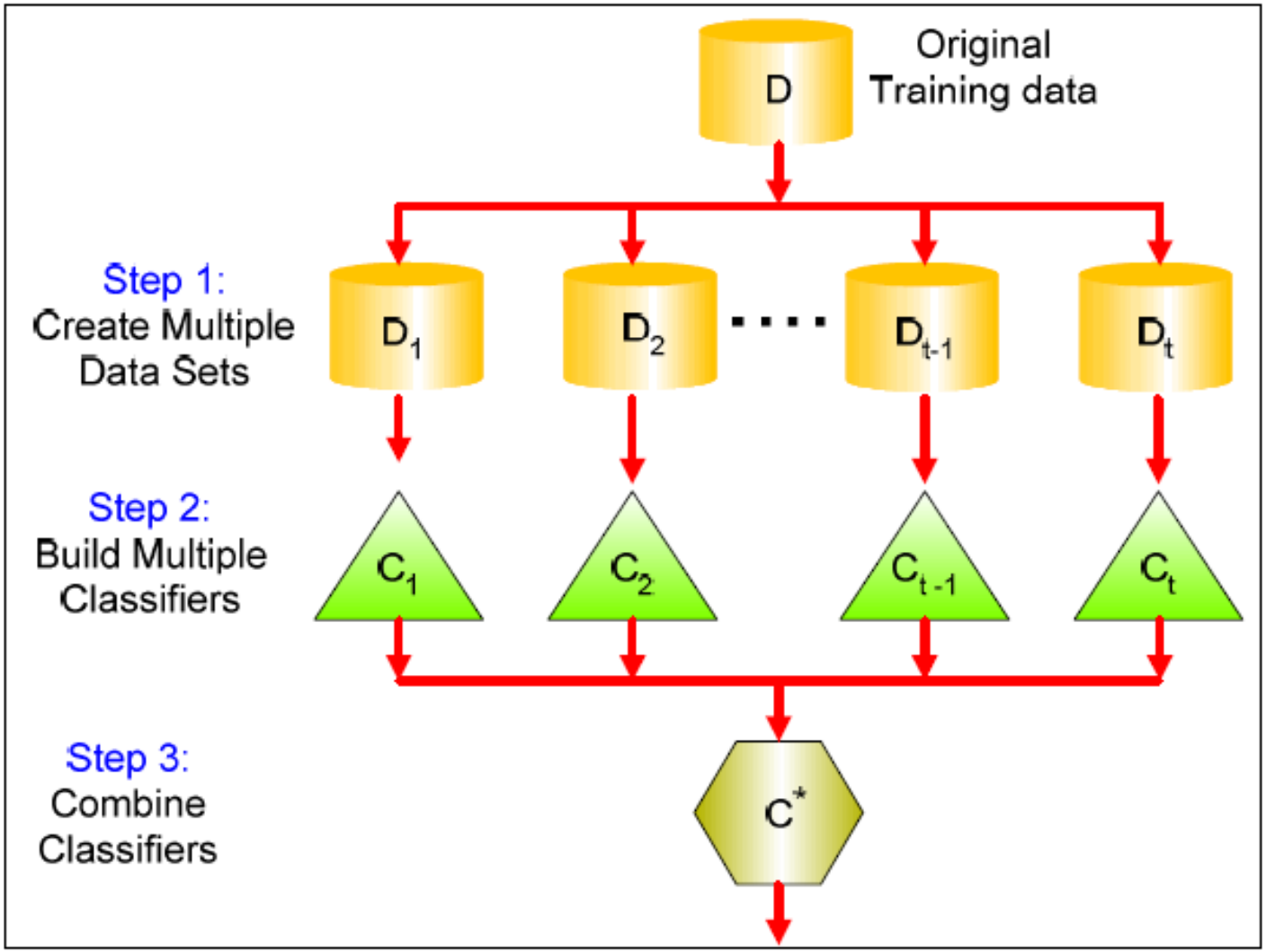
- Aprendizado apenas de uma das classes
  - ou de ambas as classes separadamente
- Normalmente interesse maior é pelo aprendizado da classe rara (positiva)
  - Por exemplo, indução de regras que descrevam somente os exemplos positivos
    - Qualquer exemplo que não satisfaça as premissas das regras induzidas é classificado como negativo por default

Ensemble Methods

- Construct a set of base classifiers from the training data

- Predict class label of previously unseen records by aggregating predictions made by multiple classifiers
  - Diversity is required

General Idea



Why does it work?

- Suppose there are 25 base classifiers
  - Each classifier has error rate, ε = 0.35
  - Assume classifiers are independent
  - From a Bernoulli Trial perspective, the probability that the ensemble classifier makes a wrong prediction is:

$$\sum_{i=13}^{25} \binom{25}{i} \varepsilon^i (1-\varepsilon)^{25-i} = 0.06$$

Ensembles

- Existem diferentes maneiras de introduzir a diversidade necessária aos classificadores base
  - Depende do(s) tipo(s) de classificador(es) usados
  - Algumas das abordagens principais são:
    - Bagging
    - Boosting
    - Random Forests

Bagging

- Realiza bootstrap para gerar uma coleção de bases de dados com o mesmo tamanho da base original de treinamento (~37% réplicas)
- Utiliza classificadores base **instáveis**
  - Sensíveis a perturbações pequenas nos dados
    - portanto às diferentes amostras de dados bootstrap
    - Exemplos: K-NN (K pequeno), ADs sem poda, RNAs,...
  - Caso contrário, ensemble pode desempenhar pior que classificador único treinado na base original!

Boosting

- Procedimento iterativo voltado para melhorar o desempenho de classificação através do enfoque progressivo em classes mais difíceis
  - A cada rodada um classificador é treinado
  - Exemplos classificados incorretamente recebem um aumento nos seus pesos (inicialmente todos iguais)
  - Pesos podem ser usados na rodada subsequente:
    - Por classificador com custos distintos associados aos exemplos
    - Como distribuição de probabilidade para amostragem (bootstrap)
- Ao contrário de bagging, são muito susceptíveis a overfitting por enfocar exemplos particulares, especialmente em dados com ruído
- Exemplo: AdaBoost

Random Forests

- São ensembles constituídos de ADs
- ADs são algoritmos determinísticos, porém aleatoriedade pode ser inserida. Por exemplo:
  - Bagging com ADs
  - Forest-RI: restrição dos atributos teste elegíveis em cada nó a um sub-conjunto dos atributos candidatos originais (sub-conjunto selecionado aleatoriamente)
  - Seleção aleatória do atributo teste em cada nó dentre os F melhores atributos, ao invés do melhor