PSI3471 – Fundamentos de Sistemas Eletrônicos Inteligentes

Medidas de desempenho

Magno T. M. Silva e Renato Candido

Escola Politécnica da USP

1 Problemas de classificação

- Vamos voltar ao problema das meias-luas, em que se deseja classificar um determinado ponto como pertencente à Região A ou à Região B
- **Dado um determinado ponto, ele pertence à Região A?**
- Há duas possíveis respostas: Sim ou Não, sendo que quando ocorre a resposta negativa, subentende-se que o ponto pertence à Região B
- Os elementos relevantes são os da Região A.

2 Tipos de erros de classificação

Rótulo verdadeiro	Rótulo Predito	Tipo de erro
Sim	Sim	Verdadeiro Positivo (VP)
Não	Não	Verdadeiro Negativo (VN)
Não	Sim	Falso Positivo (FP)
Sim	Não	Falso Negativo (FN)

3 Tipos de erros de classificação

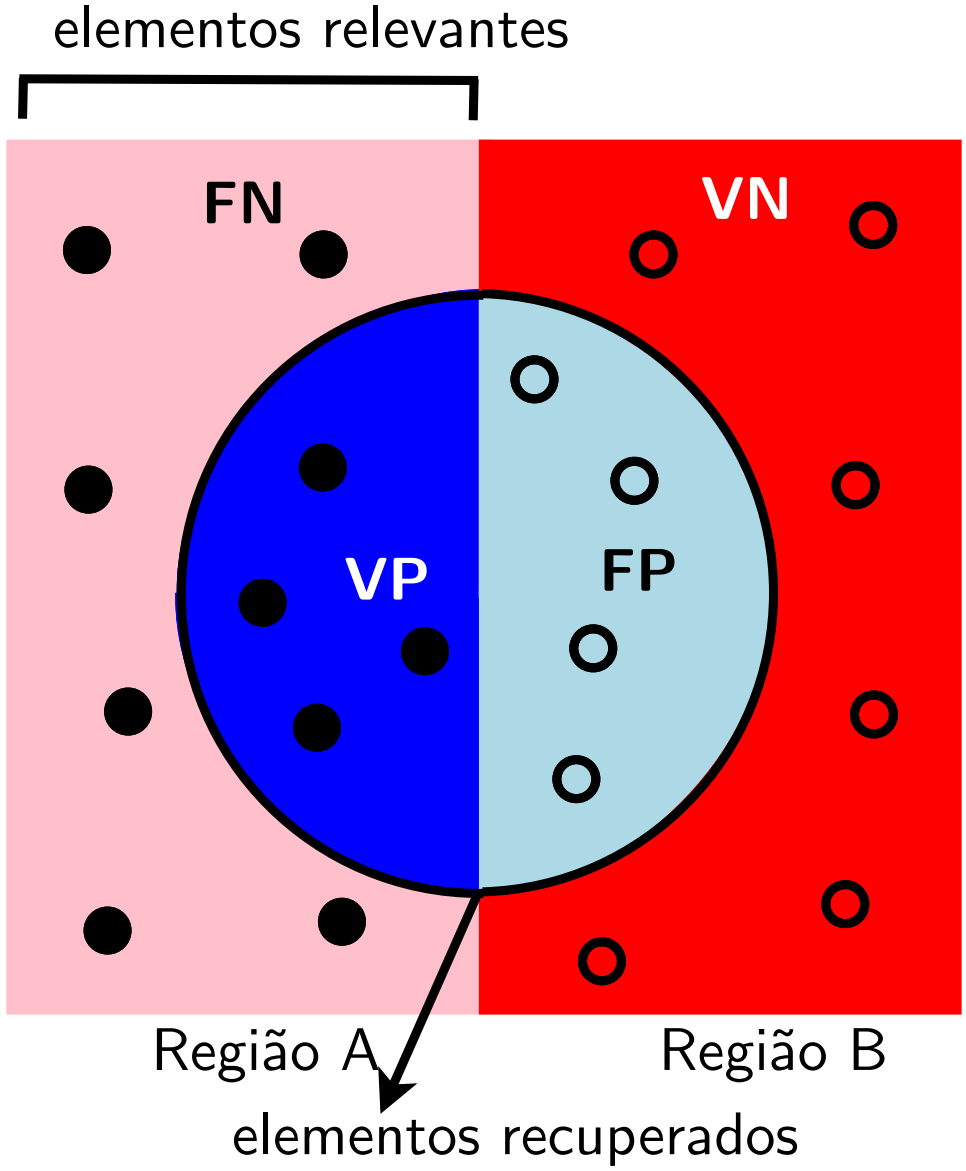
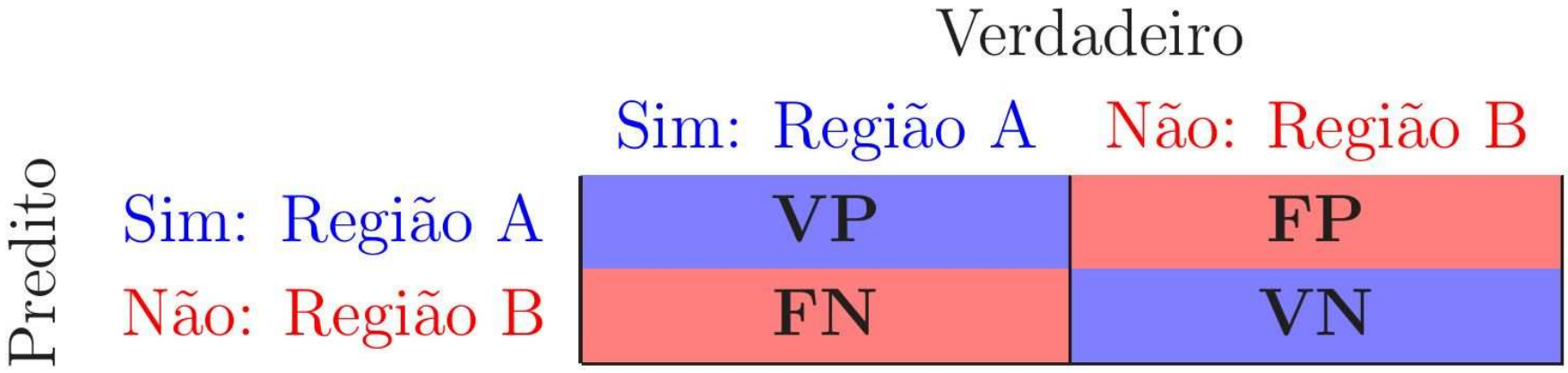


Diagrama dos erros de classificação. Adaptado de
`https://en.wikipedia.org/wiki/Precision_and_recall`

4 Matriz de confusão

- A **matriz de confusão** é uma tabela que contém o resumo dos resultados de um problema de classificação
- Os números de predições corretas e incorretas são resumidos com valores de contagem e divididos por cada classe
- Ela mostra as maneiras pelas quais seu **modelo de classificação fica confuso** quando faz predições, fornecendo **informações sobre tipos de erros** cometidos pelo classificador

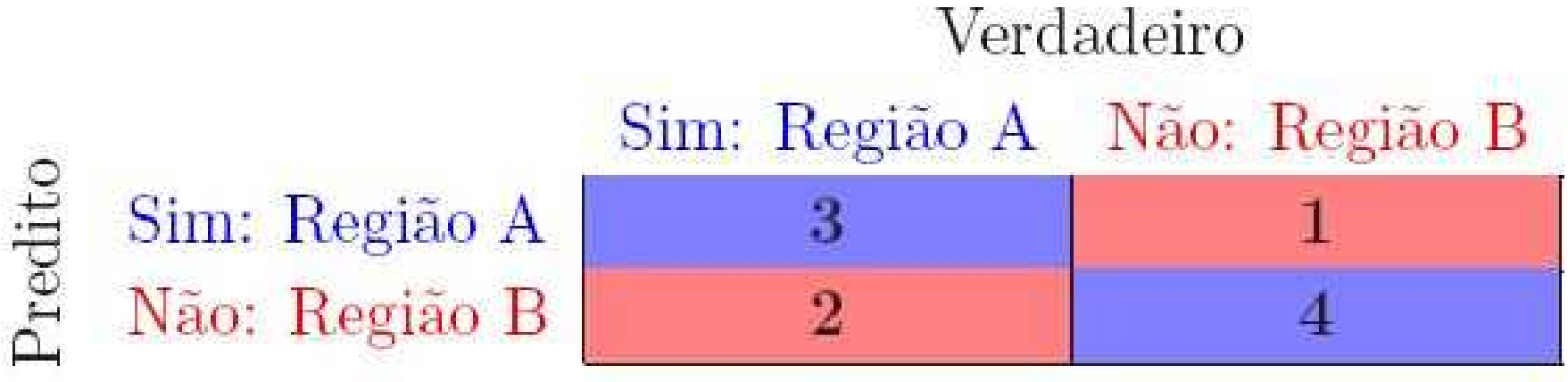


5 Exemplo de um problema de classificação

Considere o exemplo de classificação da tabela abaixo, em que há 10 predições, sendo VP $= 3$, VN $= 4$, FP $= 1$ e FN $= 2$

Rótulo Correto	S	N	N	S	S	N	S	S	N	N
Rótulo Predito	N	N	N	S	S	N	N	S	N	S
Tipo de erro	FN	VN	VN	VP	VP	VN	FN	VP	VN	FP

A matriz de confusão desse exemplo é dada por



6 Matriz de confusão

- Para um bom classificador, a matriz de confusão deve ter **valores na diagonal principal muito maiores que os valores da antidiagonal**, que idealmente deveriam ser iguais a zero
- Apesar de ser a mais usual, **a definição da matriz de confusão pode variar**. O *scikit-learn*, por exemplo, define a matriz de confusão como a transposta da matriz aqui apresentada

7 Acurácia e taxa de erro

$$\boxed{\text{Acurácia} = \frac{\#\text{predições corretas}}{\#\text{predições totais}} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}}$$

$$\boxed{\text{Taxa de erro} = \frac{\#\text{predições incorretas}}{\#\text{predições totais}} = \frac{\text{FP} + \text{FN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}}$$

$$\boxed{\text{Taxa de erro} = 1 - \text{Acurácia}}$$

8 Acurácia e taxa de erro – Exemplo

- Considerando o exemplo de classificação da tabela, temos

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} = \frac{3+4}{10} = 0{,}7 = 70\%$$

$$\text{Taxa de erro} = 1 - \text{Acurácia} = 1 - 0{,}7 = 0{,}3 = 30\%$$

- A **acurácia não é uma métrica adequada** para analisar a classificação quando as classes estão **desbalanceadas**.
- Se 97% dos exemplos de treinamento forem da Região B e apenas 3% da Região A, se o modelo predizer apenas os pontos da Região B, a acurácia será de 97% e nenhum ponto da Região A será detectado
- Portanto, o modelo parece ter um ótimo desempenho com base na acurácia, mas falha ao detectar pontos da Região A

9 Precisão, Sensibilidade e F_1-*score*

- A **precisão** mede a fração dos elementos relevantes entre os elementos recuperados, ou seja,

$$\boxed{\text{Precisão} = \frac{\#\text{amostras positivas preditas corretamente}}{\#\text{todas as amostras preditas como positivas}}}$$

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

- A precisão é uma medida que nos diz qual é a proporção dos pontos que foram classificados como pertencentes à Região A que de fato são da Região A

10 Precisão, Sensibilidade e F_1-*score*

- A **sensibilidade**, também conhecida como **taxa de verdadeiros positivos** (*True Positive Rate* – TPR) ou **revocação** (*recall*), calcula a fração dos elementos relevantes que foram recuperados, ou seja,

$$\text{Sensibilidade} = \frac{\#\text{amostras positivas preditas corretamente}}{\#\text{todas as amostras com rótulos positivos}}$$

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

- A sensibilidade é uma medida que nos diz qual é a proporção dos pontos que de fato são da Região A que foram classificados como pertencentes a essa região.

11 Precisão, Sensibilidade e F_1-*score*

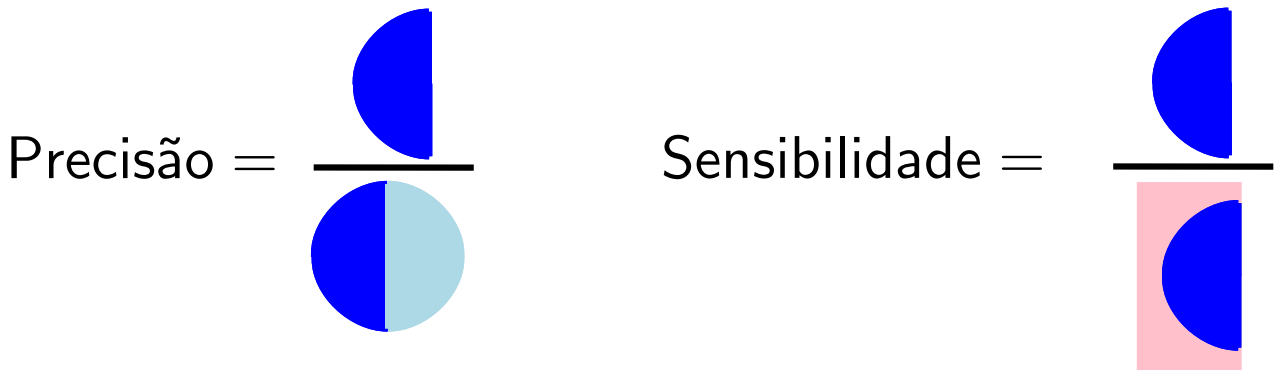


Ilustração do cálculo da precisão e sensibilidade seguindo o esquema
Fonte: Adaptado de
`https://en.wikipedia.org/wiki/Precision_and_recall`

12 Precisão, Sensibilidade e F_1-*score*

- A *sensibilidade* fornece informação sobre o desempenho do classificador com relação aos **falsos negativos**, enquanto a **precisão** em relação aos **falsos positivos**.
- Existe um **compromisso** entre o desempenho em termos dessas medidas:
  - Quando se busca **aumentar a precisão**, diminuindo o número de falsos positivos, **diminui-se sensibilidade**, pois o número de falsos negativos aumenta e vice-versa.
  - Dependendo do problema, pode ser mais interessante **priorizar uma ou outra métrica**, alterando-se o limiar utilizado para classificar um exemplo positivo
  - **Um classificador com alta precisão tende a deixar alguns exemplos de fora** (falsos negativos) **mas aqueles classificados como positivo tem uma alta qualidade** (poucos falsos positivos)
  - **Um classificador com uma sensibilidade alta, deixa poucos exemplos de fora** (falsos negativos) **mas aqueles classificados como positivo não tem tanta qualidade** (mais falsos positivos)

13 Precisão, Sensibilidade e F_1-*score*

- Uma métrica que **combina a precisão e a sensibilidade** é a F_1-*score*:

$$F_1\text{-}score = 2\frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}$$

- A **média harmônica** é uma espécie de média quando essas medidas são iguais. Mas quando elas são diferentes, essa métrica **fica mais próxima do menor valor** em comparação com o maior
- Se a precisão ou sensibilidade for **muito pequena**, o F_1-*score* **levanta uma bandeira e fica mais próximo do menor valor**, dando ao modelo uma pontuação apropriada em vez de apenas uma simples média aritmética

14 Precisão, Sensibilidade e F_1-*score*

No exemplo, temos

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{3}{3+1} = 0{,}75 = 75\%,$$

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{3}{3+2} = 0{,}6 = 60\%,$$

$$F_1\text{-}score = 2\frac{0{,}75 \times 0{,}60}{0{,}75 + 0{,}60} = 0{,}6667 = 66{,}67\%.$$

Na literatura, existem outras medidas do tipo F que envolvem a precisão e sensibilidade, mas a F_1-*score* é a mais utilizada.

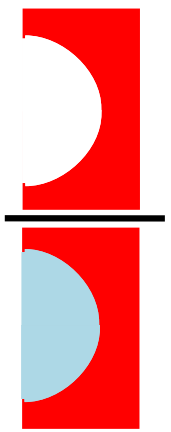
15 Especificidade e taxa de falsos positivos

- A **especificidade**, em contraste à sensibilidade, é uma medida que fornece a proporção de pontos que não pertencem à Região A e que foram previstos pelo modelo como não pertencentes à essa região:

$$\text{Especificidade} = \frac{\#\text{amostras negativas preditas corretamente}}{\#\text{todas as amostras com rótulos negativos}}$$

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

$$\text{Especificidade} =$$



16 Especificidade e taxa de falsos positivos

- Também é comum definir a taxa de falsos positivos (*False Positive Rate* – FPR):

$$\boxed{\text{FPR} = \frac{\text{FP}}{\text{VN} + \text{FP}} = 1 - \text{Especificidade}}$$

- Essa taxa fornece a proporção da classe negativa (pontos da Região B), que foi classificada incorretamente
- Uma FPR baixa é desejável uma vez que se deseja classificar corretamente os elementos da classe negativa
- No exemplo, temos

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{4}{4+1} = 0{,}8 = 80\%$$

17 Coeficiente de correlação de Matthew

- O **coeficiente de correlação de Matthew** (*Matthews Correlation Coefficient* – MCC) trata as classes verdadeira e prevista como duas variáveis aleatórias binárias e calcula seu coeficiente de correlação:

$$\mathrm{MCC} = \frac{\mathrm{VP} \times \mathrm{VN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{VP} + \mathrm{FP})(\mathrm{VP} + \mathrm{FN})(\mathrm{VN} + \mathrm{FP})(\mathrm{VN} + \mathrm{FN})}}.$$

- O **MCC fica sempre entre** -1 **e** 1, com 0 significando que o classificador não é melhor uma moeda honesta
- **Classificador perfeito** $(\mathrm{FP} = \mathrm{FN} = 0)$, $\mathrm{MCC} = 1$
- **Classificador sempre classifica mal** $(\mathrm{VP} = \mathrm{VN} = 0)$, $\mathrm{MCC} = -1$ (neste caso, você pode simplesmente inverter o resultado)
- O MCC é perfeitamente simétrico: nenhuma classe é priorizada
- Um valor alto do MCC significa que ambas as classes são bem preditas

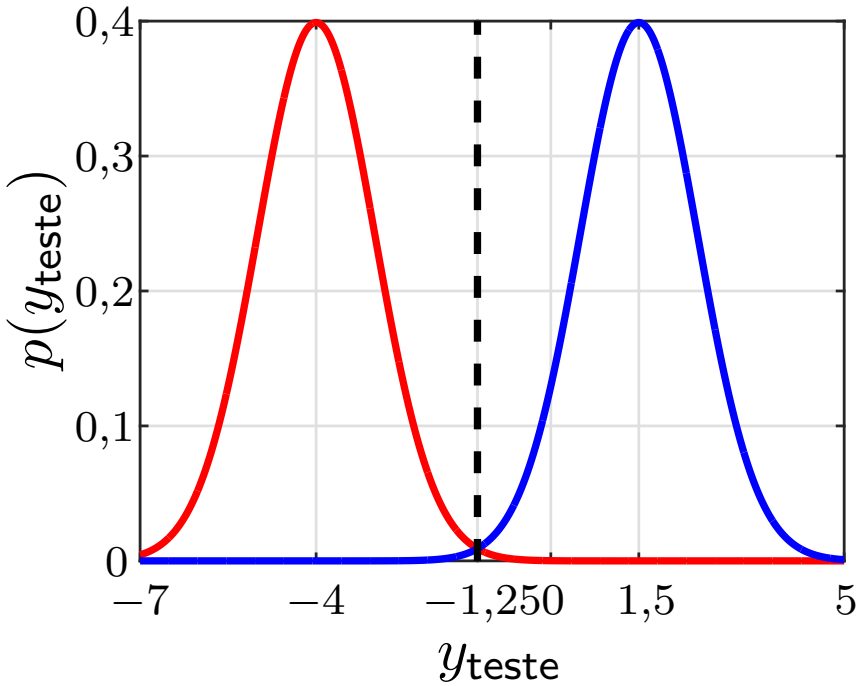
18 Coeficiente de correlação de Matthew

No exemplo, temos

$$\text{MCC} = \frac{3 \times 4 - 1 \times 2}{\sqrt{(3+1)(3+2)(4+1)(4+2)}} = \frac{10}{\sqrt{600}} = 0{,}408$$

19 Área sob a curva ROC

- No problema das meias-luas, adotamos o limiar igual a 0: quando a saída do classificador é ≥ 0, o ponto é classificado como da Região A e caso contrário como da Região B
- Esse limiar é razoável neste caso, uma vez que adotamos $d \in \{-1, +1\}$ como sinal desejado
- Suponha que a saída de um determinado classificador siga duas distribuições gaussianas com $\sigma^2 = 1$ e $\mu_1 = -4$ e $\mu_2 = 1{,}5$



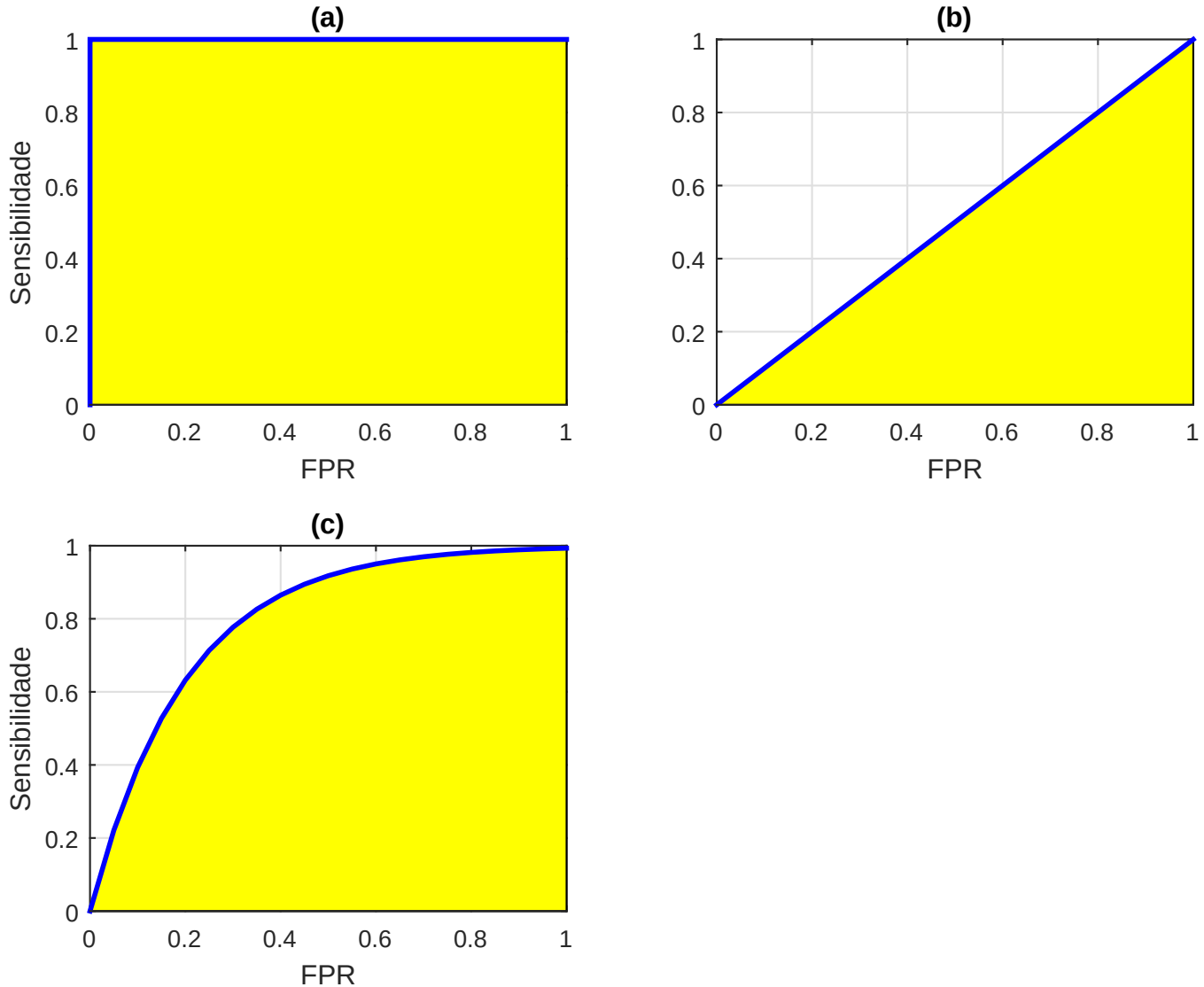
20 Área sob a curva ROC

- A gaussiana com média negativa corresponde aos pontos que devem ser classificados como pertencentes à Região B
- A gaussiana com média positiva corresponde aos pontos que devem ser classificados como pertencentes à Região A
- Para minimizar o erro de classificação, o limiar deve ser igual a $-1{,}25$ e não 0.
- **O limiar influencia nas medidas de desempenho** do classificador.
- Neste exemplo, qualquer outro limiar diferente de $-1{,}25$ levará a um maior erro de classificação.
- Na escolha do limiar, devem ser feitos testes para cada valor, ou seja, pode-se gerar a matriz de confusão e comparar as métricas discutidas até agora.
- O melhor a se fazer é considerar a **área sob a curva ROC**

21 Área sob a curva ROC

- Dado um limiar, **a curva ROC (Receiver Operator Characteristic)** é um gráfico da taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (FPR $= 1 -$ Especificidade).
- A área sob essa curva **(area under the curve – AUC)** é uma medida da habilidade do classificador de distinguir entre as classes e é usada frequentemente como métrica.
- Quanto maior a AUC, melhor o desempenho do classificador.

22 Área sob a curva ROC



Exemplos de curvas ROC com três valores distintos de AUC: (a) AUC $= 1$, (b) AUC $= 0{,}5$ e (c) AUC $= 0{,}8$.

23 Área sob a curva ROC

- Quando AUC $= 1$ [Fig. (a)], o classificador é capaz de distinguir perfeitamente os pontos da Região A dos pontos da Região B
- Se AUC $= 0$, o modelo classifica todos os pontos da Região A como pertencentes a Região B e vice-versa
- Quando AUC $= 0{,}5$ [Fig. (b)], o classificador não é capaz de distinguir entre as classes (moeda)
- Para $0{,}5 <$ AUC < 1 [Fig. (c)], há uma grande chance de que o classificador seja capaz de distinguir as classes. Isso ocorre porque o classificador detecta mais VP e VN do que FN e FP
- Um valor mais alto do eixo x indica um maior número de FP do que de VN
- Um valor mais alto do eixo y indica um maior número de VP do que FN
- A escolha do limiar deve levar em conta um compromisso entre o número de FP e FN

24 Problemas de regressão – Erro absoluto médio

- O **erro absoluto médio** (*mean absolute error* – MAE) fornece uma medida de quão longe as predições estão dos valores verdadeiro, mas não fornece nenhuma ideia da direção do erro, ou seja, se o modelo está subestimando ou superestimando os dados
- O erro absoluto médio é calculado como

$$\boxed{\mathrm{MAE} = \frac{1}{N_{\text{teste}}} \sum_{n=1}^{N_{\text{teste}}} |d_n - y_n|}$$

25 Problemas de regressão – Erro quadrático médio

- O **erro quadrático médio** (*mean square error* – MSE) é semelhante ao erro absoluto médio
- A única diferença é que o MSE toma a média do quadrado da diferença entre os valores desejados e os valores preditos
- Como se calcula o quadrado do erro, o efeito de erros maiores se torna mais pronunciado do que o de erros menores
- O MSE é calculado como

$$\mathrm{MSE} = \frac{1}{N_{\text{teste}}} \sum_{n=1}^{N_{\text{teste}}} (d_n - y_n)^2$$

26 Problemas de regressão – Erro quadrático médio

- É comum usar também a **raiz quadrada do MSE** (*root mean square error* – RMSE) como métrica, ou seja,

$$\boxed{\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N_{\text{teste}}} \sum_{n=1}^{N_{\text{teste}}} (d_n - y_n)^2}.}$$

- A vantagem é que tanto essa métrica como o MAE possuem a mesma unidade da variável predita, o que torna sua interpretação mais simples

27 Problemas de regressão – R^2

- A métrica R^2 é usada para fins explicativos e fornece uma indicação da qualidade ou ajuste de um conjunto de valores de saída previstos aos valores desejados
- Essa métrica é calculada como

$$R^2 = 1 - \frac{\sum_{n=1}^{N_{\text{teste}}} (d_n - y_n)^2}{\sum_{n=1}^{N_{\text{teste}}} (\bar{d} - d_n)^2},$$

em que

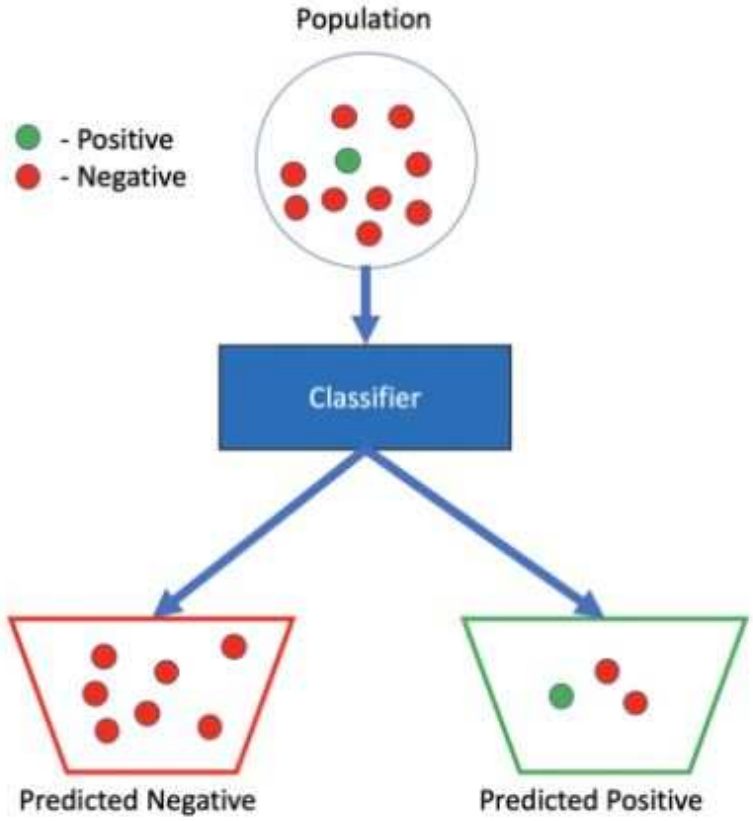
$$\bar{d} = \frac{1}{N_{\text{teste}}} \sum_{n=1}^{N_{\text{teste}}} d_n$$

é a média dos valores desejados do conjunto de teste

28 Problemas de regressão – R^2

- O denominador da fração que aparece na definição de R^2 é proporcional à variância dos dados de teste
- No melhor caso, os valores preditos são exatamente iguais aos valores desejados, o que leva a $R^2 = 1$
- Caso o modelo leve a $y_n = \overline{d}$, $n = 1, 2, \cdots, N_{\text{teste}}$, que é conhecido como modelo base, teremos $R^2 = 0$
- Modelos cujas predições são piores que as do modelo base podem levar a R^2 negativo

Teste 4 – Exemplo 1



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	2	0	7	80	33	100	50	78	22

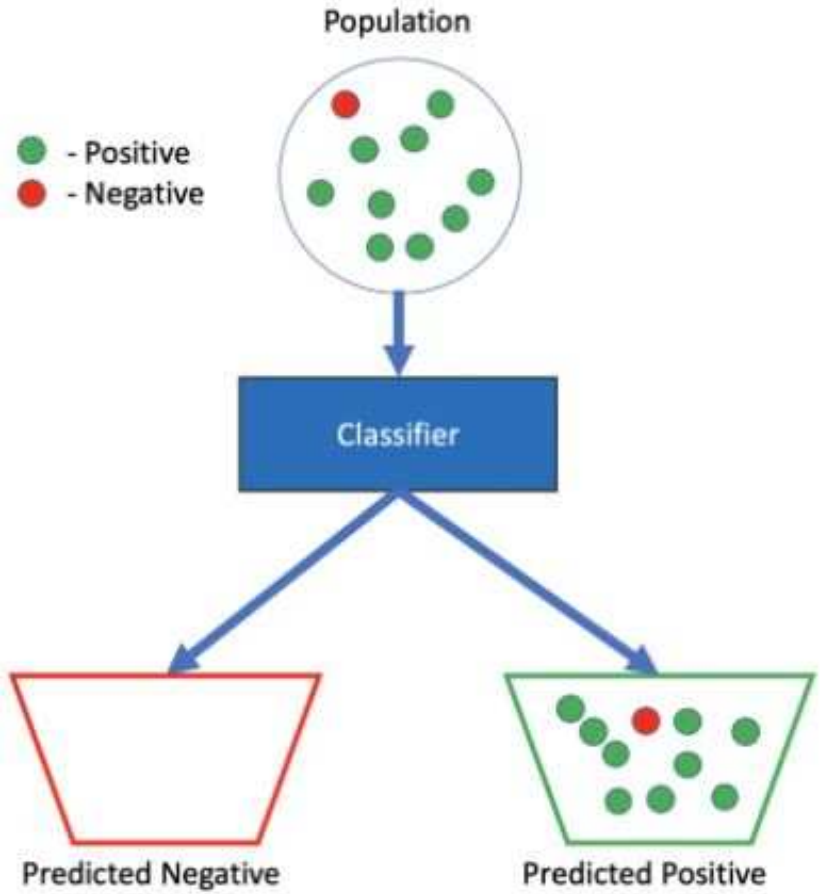
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 1

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	2	0	7	80	33	100	50	78	22

- Baixa precisão, alta sensibilidade e alta especificidade
- Se o resultado da classificação for negativo, pode-se concluir que o exemplo é negativo
- Se o exemplo for negativo, não há garantia que que ele será predito como negativo (E $= 78\%$)
- Se o resultado da classificação for positivo, não se pode concluir que o exemplo é positivo (P $= 33\%$)
- Se o exemplo for positivo, pode-se confiar no classificador (S $= 100\%$)

Teste 4 – Exemplo 2



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
9	1	0	0	90	90	100	95	0	100

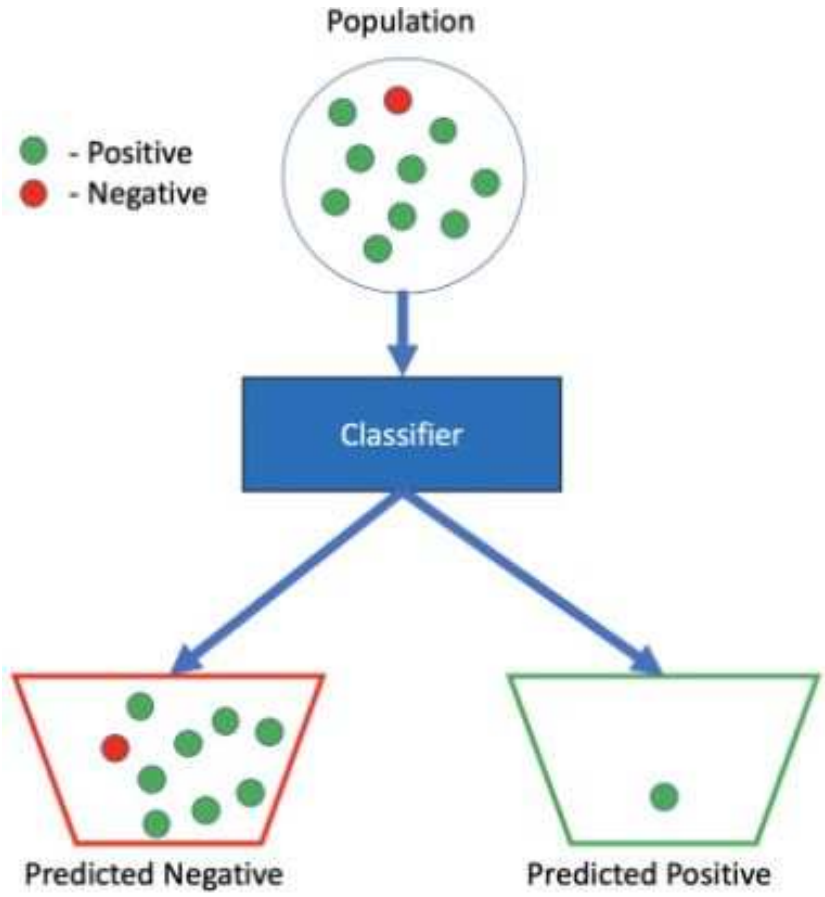
Fonte:https://towardsdatascience.com/

29 Teste 4 – Exemplo 2

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
9	1	0	0	90	90	100	95	0	100

- Alta precisão, alta sensibilidade e baixa especificidade
- Predizer tudo como positivo não é uma boa ideia
- Como a população é desbalanceada, a precisão é relativamente alta
- A sensibilidade é igual a 100% porque os exemplos positivos são preditos como positivos
- A especificidade é 0% porque nenhum exemplo negativo é predito como negativo

Teste 4 – Exemplo 3



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	0	8	1	20	100	11	20	100	0

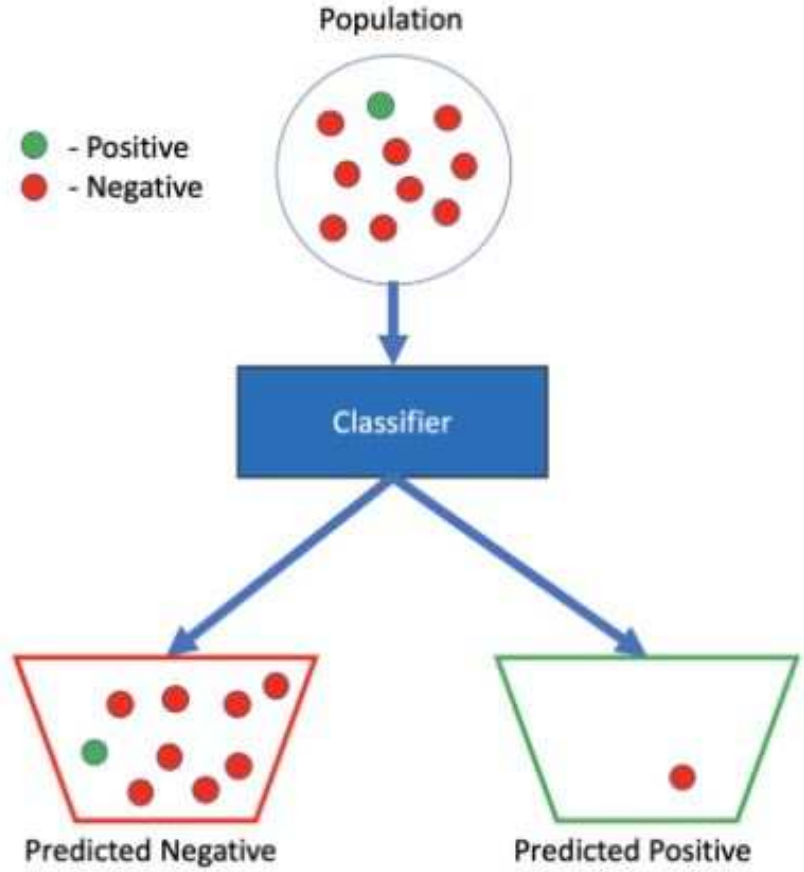
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 3

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	0	8	1	20	100	11	20	100	0

- Alta precisão, baixa sensibilidade e alta especificidade
- Este classificador pode ser útil
- Se ele prediz que o exemplo é positivo, ele é positivo (pode-se confiar)
- Se ele prediz que ele é negativo, não se pode confiar (há uma grande possibilidade de ele ainda ser positivo)

Teste 4 – Exemplo 4



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
0	1	1	8	80	0	0	NaN	89	11

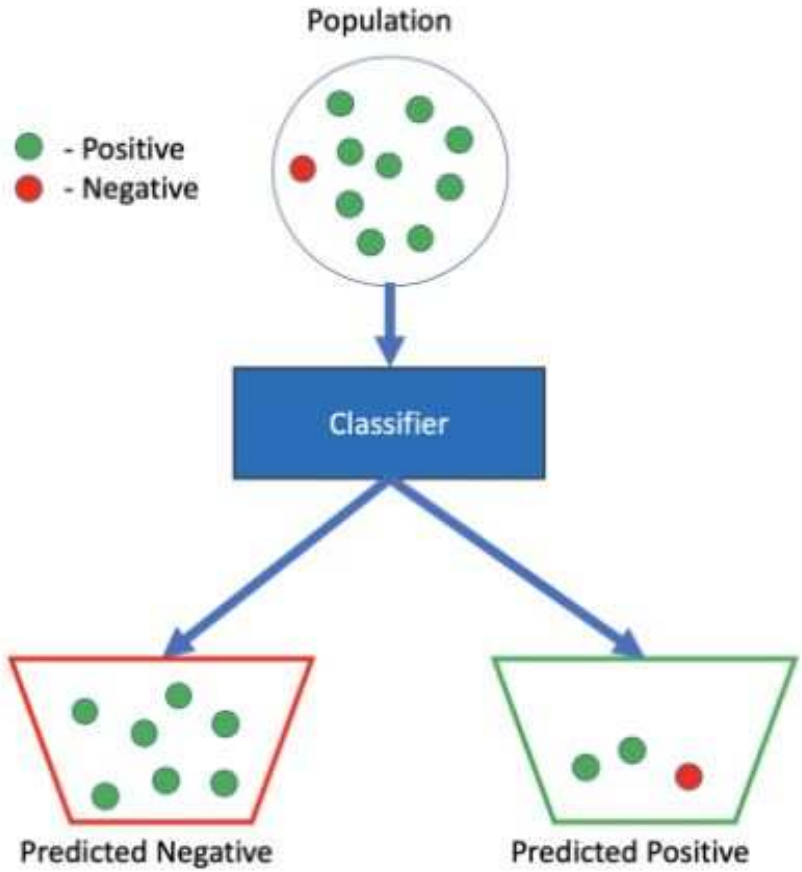
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 4

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
0	1	1	8	80	0	0	NaN	89	11

- Baixa precisão, alta sensibilidade e alta especificidade
- Este classificador é muito ruim
- Ele prediz quase todos como negativo
- Quando a predição é positiva, o classificador erra
- Usar o oposto do classificador é mehor

Teste 4 – Exemplo 5



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
2	1	7	0	20	67	22	33	0	100

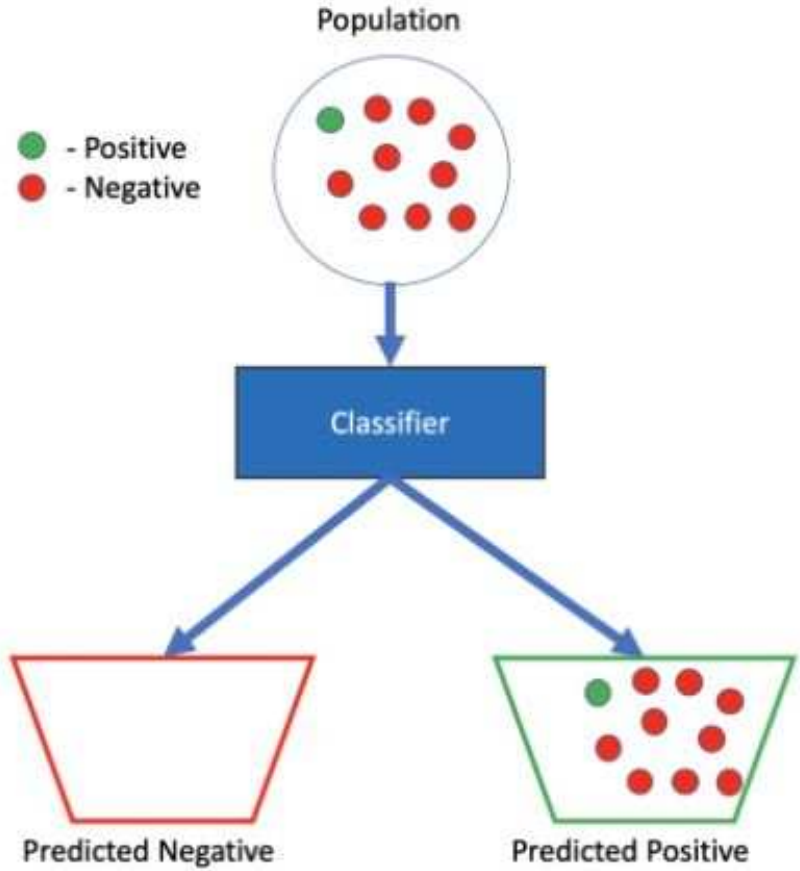
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 5

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
2	1	7	0	20	67	22	33	0	100

- Alta precisão, baixa sensibilidade e baixa especificidade
- Considerar o oposto do que esse classificador prediz seria melhor

Teste 4 – Exemplo 6



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	9	0	0	10	10	100	18	0	100

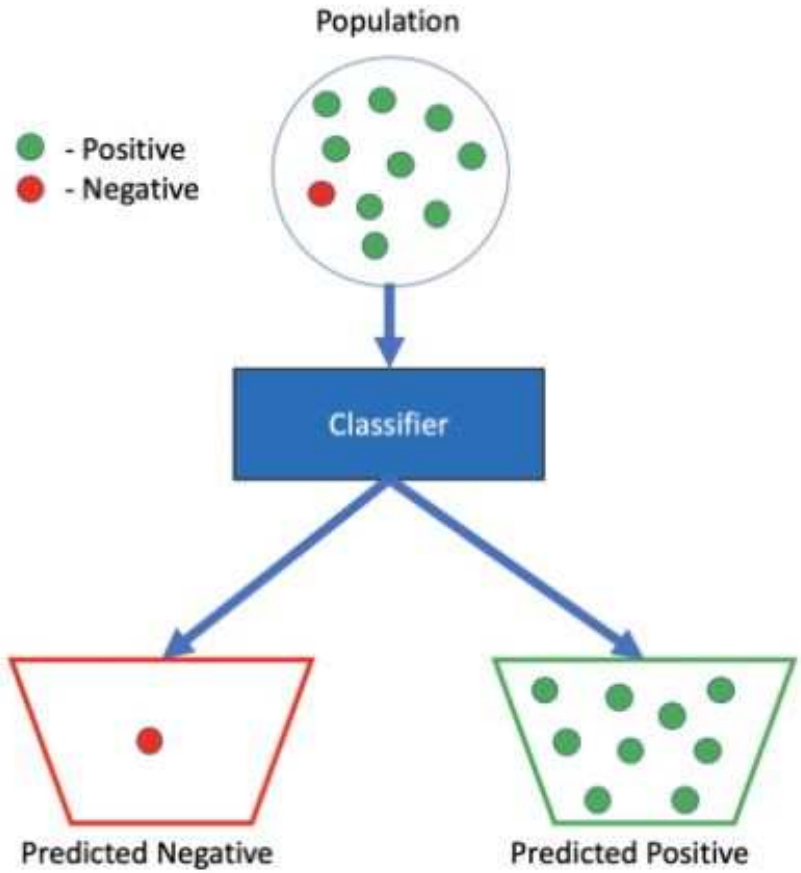
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 6

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
1	9	0	0	10	10	100	18	0	100

- Baixa precisão, alta sensibilidade e baixa especificidade
- Esse classificador não tem muita utilidade
- Ele prediz tudo como positivo
- Ele pode detectar perfeitamente todas as amostras positivas (S $= 100\%$), mas não se obtém nenhuma informação útil ao utilizá-lo

Teste 4 – Exemplo 7



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
9	0	0	1	100	100	100	100	100	0

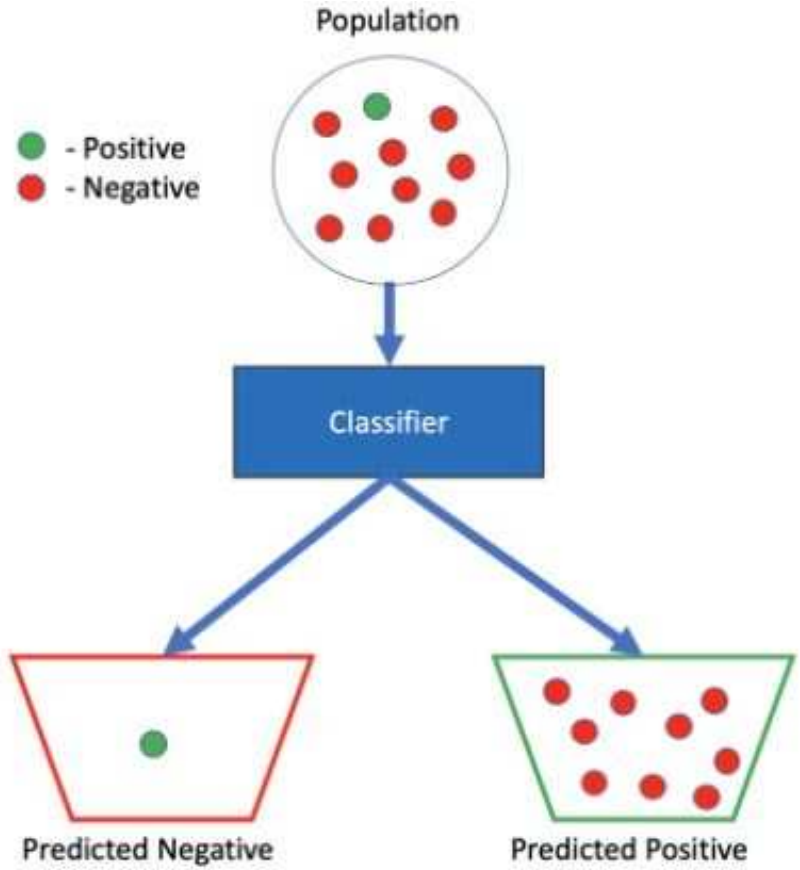
Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 7

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
9	0	0	1	100	100	100	100	100	0

- Alta precisão, alta sensibilidade e alta especificidade
- Esse é o classificador perfeito
- Ele classifica todos os exemplos positivos como positivos e negativos como negativos

Teste 4 – Exemplo 8



VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
0	9	1	0	0	0	0	NaN	0	100

Fonte:https://towardsdatascience.com/

Teste 4 – Exemplo 8

VP	FP	FN	VN	Acc(%)	P(%)	S(%)	F1(%)	E(%)	FPR(%)
0	9	1	0	0	0	0	NaN	0	100

- Baixa precisão, baixa sensibilidade e baixa especificidade
- Esse classificador parece ser muito ruim
- Ele classifica todos os exemplos positivos como negativos e negativos como positivos
- Se considerarmos a classificação oposta ele será perfeito

Exemplos de Aplicação

1. Filtro de spam
   - VP marcado como spam, é spam
   - FP marcado como spam, não é spam
   - VN não marcado como spam, não é spam
   - FN não marcado como spam, é spam

   Resultado de um FP: perder um email genuíno (crítico)

   Resultado de um FN: receber um email que é spam

Exemplos de Aplicação

1. Detecção de fraude
   - VP marcado como fraude, é fraude
   - FP marcado como fraude, não é fraude
   - VN não marcado como fraude, não é fraude
   - FN não marcado como fraude, é fraude

   Resultado de um FP: perder uma negociação

   Resultado de um FN: ter um prejuízo em uma negociação (crítico)

Exemplos de Aplicação

3. Detecção de Covid
   - VP marcado como Covid, é Covid
   - FP marcado como Covid, não é Covid
   - VN não marcado como Covid, não é Covid
   - FN não marcado como Covid, é Covid

   Resultado de um FP: um paciente em isolamento, sem necessidade

   Resultado de um FN: um paciente transmitindo a doença (crítico)

Exemplos de Aplicação

Quais classificadores dos exemplos você usaria em cada aplicação?

- a ideia é considerar a otimização para minimização de FPs ou FNs, usando as medidas:
  - Precisão: importante para os casos em que é importante diminuir os FPs
  - Sensibilidade: importante para os casos em que é importante diminuir os FNs