Universidade Federal do Paraná (UFPR)
Bacharelado em Informática Biomédica

Boas Maneiras em Aprendizado de Máquinas

David Menotti
www.inf.ufpr.br/menotti/ci171-182

Boas Maneiras

Agenda

- Introdução
- Métricas de Avaliação
- Técnicas de Validação
- Avaliação de Classificadores
- Comparando Classificadores

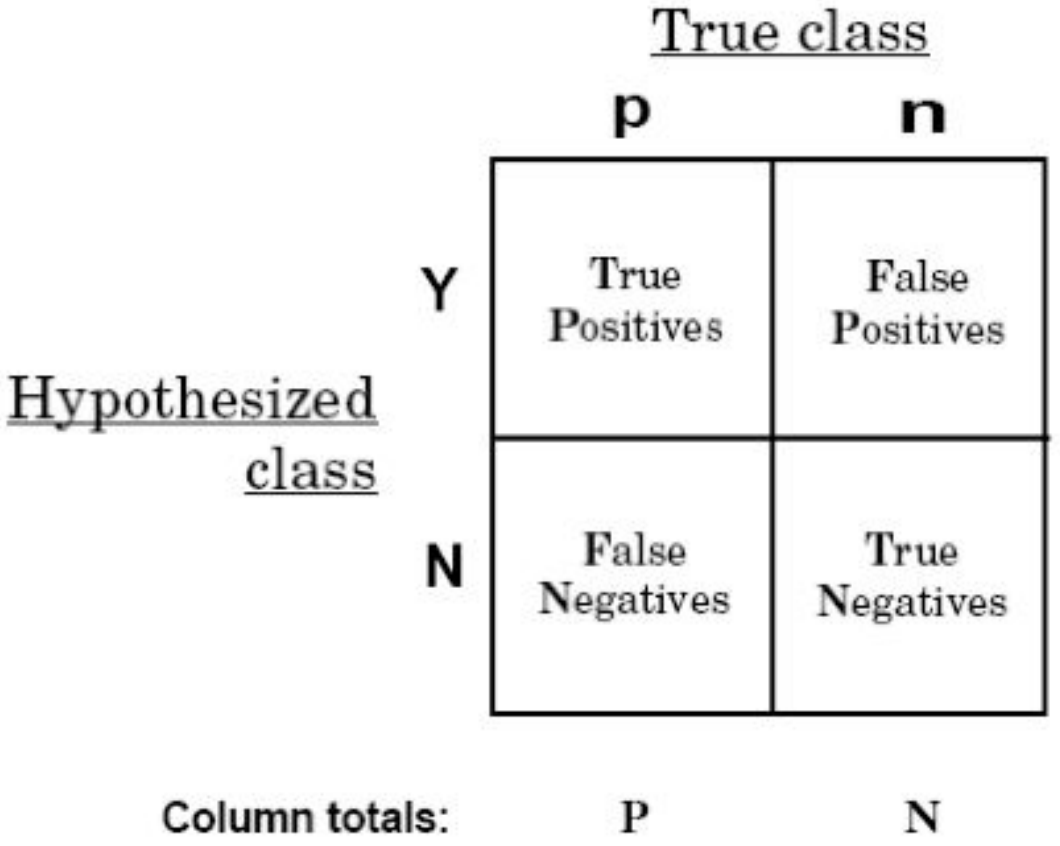
Introdução

- Considere por exemplo, um problema de classificação binário com 99 exemplos na classe positiva e 1 exemplo na classe negativa.
- Considere que o classificador *classificou* corretamente todos os exemplos da classe positiva e errou somente o exemplo da classe negativa.
- Nesse caso temos uma acurácia de 99\%.

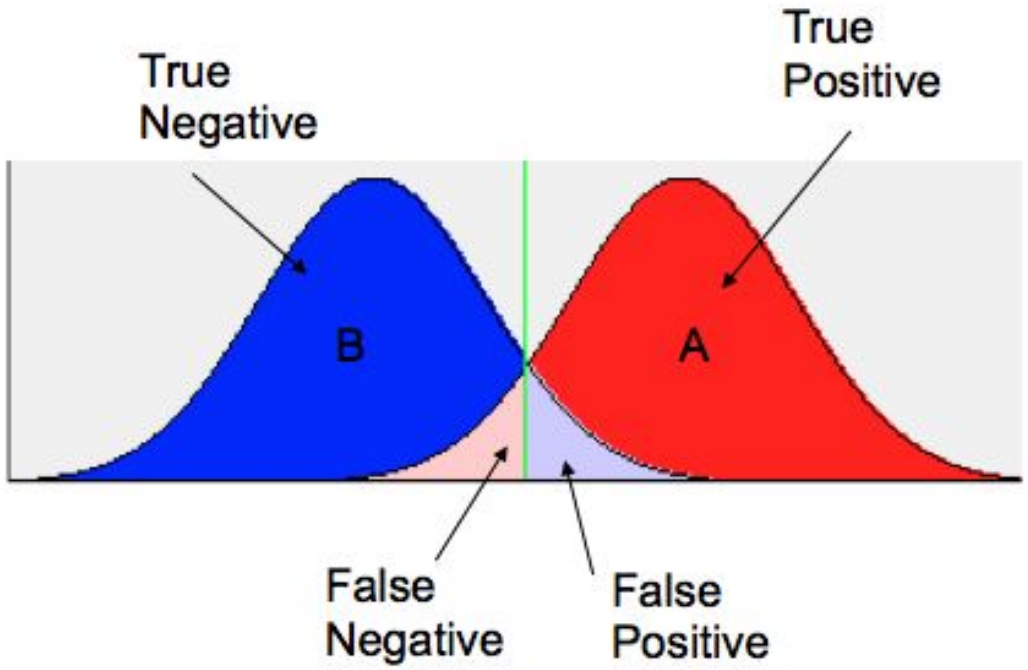
- **Atenção**: Acurácia de um classificador nem sempre é a melhor medida para se avaliar um classificador, principalmente em problemas **desbalanceados**.

Avaliação

- Dado um classificador binário, as saídas possíveis são as seguintes:



Avaliação

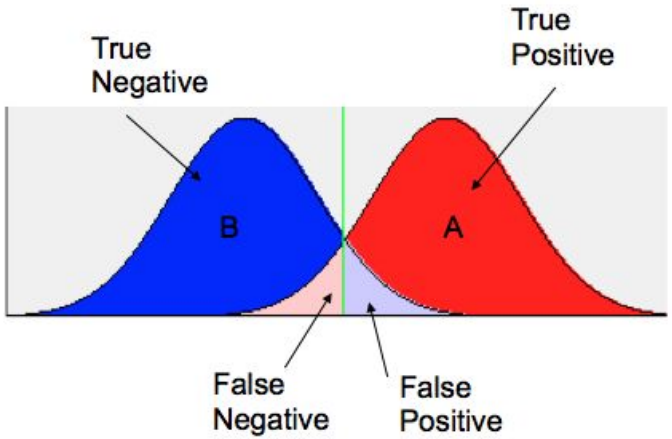


TPF	FPF
0.920	0.088
FNF	TNF
0.079	0.911

TP – Classe é A e classificamos como A
TN – Classe é B e classificamos como B
FP – Classe é B e classificamos como A
FN – Classe é A e classificamos como B

Tipos de Erro

- **Erro Tipo I** (α-erro, falso positivo):
  - Aceita-se como genuíno algo que é falso - **FP**
- **Erro Tipo II** (β-erro, falso negativo):
  - Rejeita-se algo que deveria ter sido aceito - **FN**



TPF	FPF
0.920	0.088
FNF	TNF
0.079	0.911

TP – Classe é A e classificamos como A
TN – Classe é B e classificamos como B
FP – Classe é B e classificamos como A
FN – Classe é A e classificamos como B

Métricas

- FP Rate
  - FP / N
- TP Rate ou hit rate
  - TP / P
- Accuracy
  - (TP + TN) / (P + N)
- P = TP + FN
- N = TN + FP

- *Precision* (Pr)
  - TP / (TP + FP)

Tende a 1 se FP tende a 0

- *Recall* (R) (revogação)
  - TPR = TP / (TP + FN)

Tende a 1 se FN tende a 0

- F-Measure
  - 2 / (1 / Pr + 1 / R)
  - 2 Pr. R / (Pr + R)

Média harmônica de Pr e R, tendo em vista que são grandezas inversamente proporcionais

Métricas

No scikit-learn, essas métricas estão implementadas no método `classification_report` (classe `metrics`)

```
>>> from sklearn.metrics import classification_report
>>> y_true = [0, 1, 2, 2, 0]
>>> y_pred = [0, 0, 2, 2, 0]
>>> target_names = ['class 0', 'class 1', 'class 2']
>>> print(classification_report(y_true, y_pred, target_names=target_names))
             precision    recall  f1-score   support

    class 0       0.67      1.00      0.80         2
    class 1       0.00      0.00      0.00         1
    class 2       1.00      1.00      1.00         2

avg / total       0.67      0.80      0.72         5
```

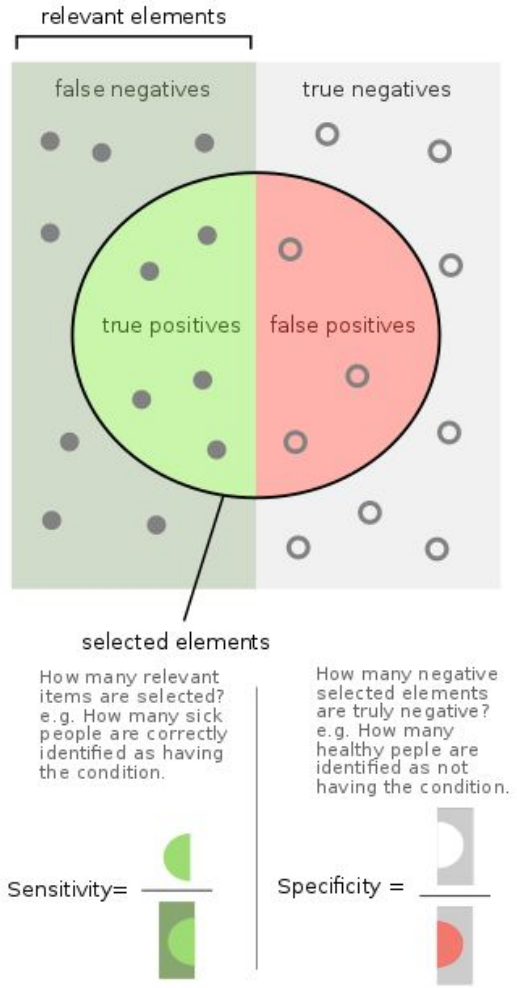
Métricas

- Especificidade / *Specificity*
  - TN / N

- Sensibilidade / *Sensibility*
  - TP / P

    * *Recall*

- P = TP + FN
- N = TN + FP



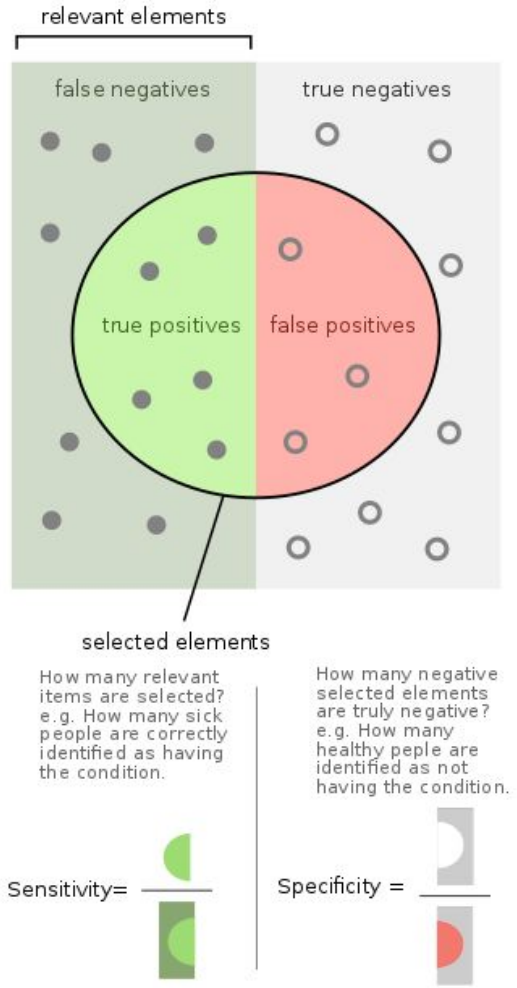
Métricas

- Especificidade / *Specificity*
  - TN / N

- Sensibilidade / *Sensibility*
  - TP / P

    * *Recall*

- P = TP + FN
- N = TN + FP



Técnicas de Validação

- Resubstitution
- Hold-out
- K-fold cross-validation
- LOOCV
- Random subsampling
- Bootstrapping

Resubstitution

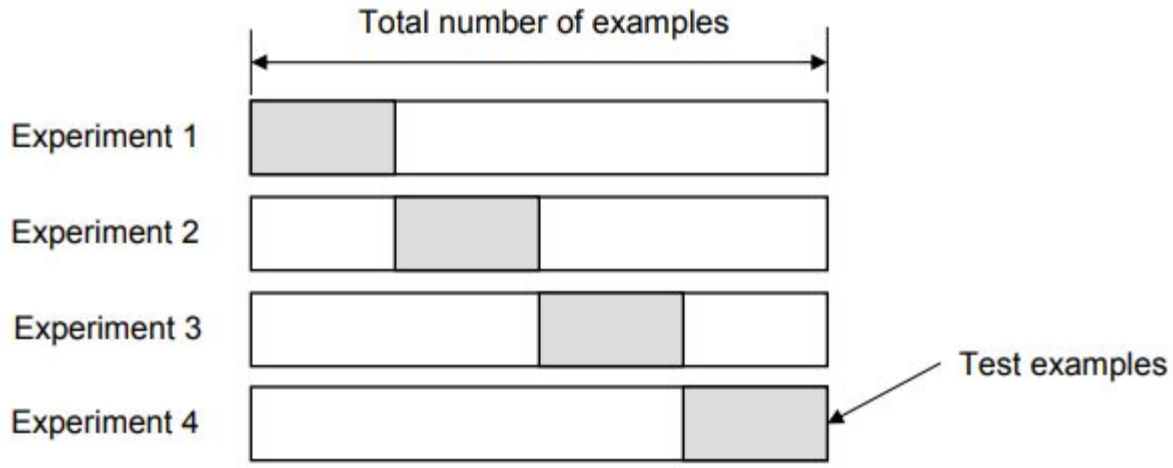
- Se todos os dados forem usados para treinar o modelo e a taxa de erro for avaliada com base no resultado versus valor real do mesmo conjunto de dados de treinamento, esse erro será chamado de **erro de resubstituição** (resubstitution error validation).
- Por exemplo:
  - Análise de *Malware*
  - Aplicações onde necessitam de:
    - *Black list*
    - *White list*

Hold-out

- Para evitar o erro de resubstituição, os dados são divididos em dois conjuntos de dados diferentes rotulados como um conjunto de treinamento e de teste.
- A divisão pode ser: 60% / 40% ou 70% / 30% ou 80% / 20%, etc.
  - Avaliar classificador em diferentes cenários:
    - 5% de train? ou 90% de train?
- Muito provavelmente haverá distribuição desigual das classes (alvos/metas) nos conjunto de dados de treinamento e teste.
  - os conjuntos de dados de treinamento e teste são criados com distribuição igual de diferentes classes de dados.
    - Esse processo é chamado de **estratificação**.

K-fold cross-validation

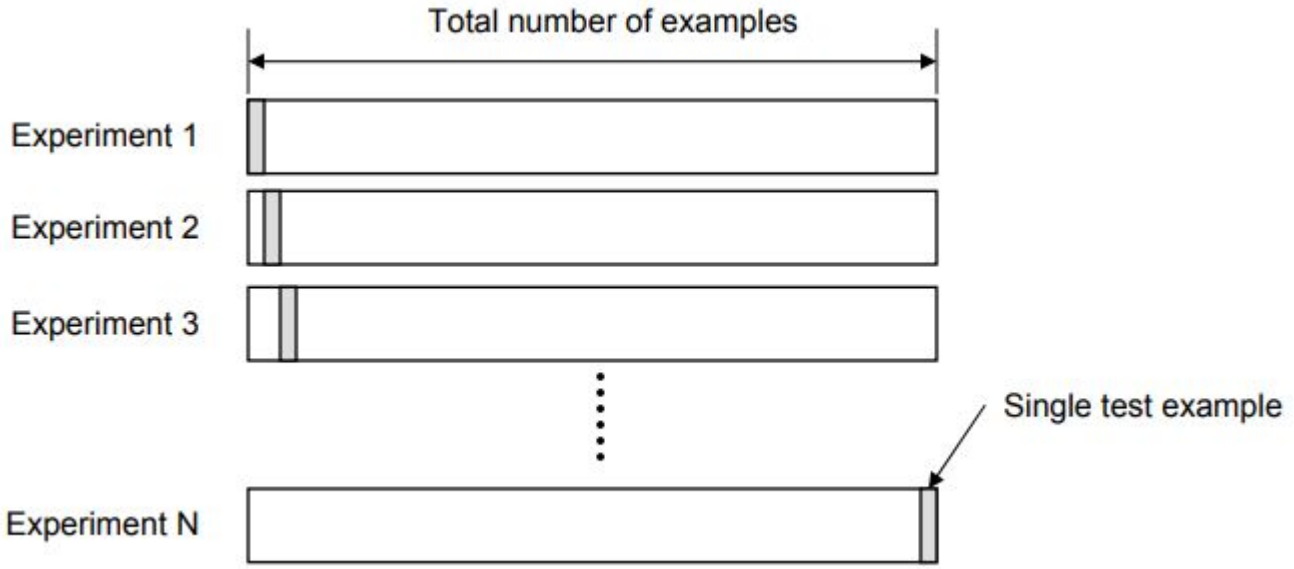
- Nesta técnica, k-1 *folds* são usadas para **treinamento** e o restante (1 *fold*) é usado para **teste**.



- A vantagem é que todos os dados são usados para treinamento.
  - Cada amostra é usada apenas uma vez para **teste**.
- A taxa de erro do modelo é a média
- Pode-se ver esta técnica como um *hold-out* repetido

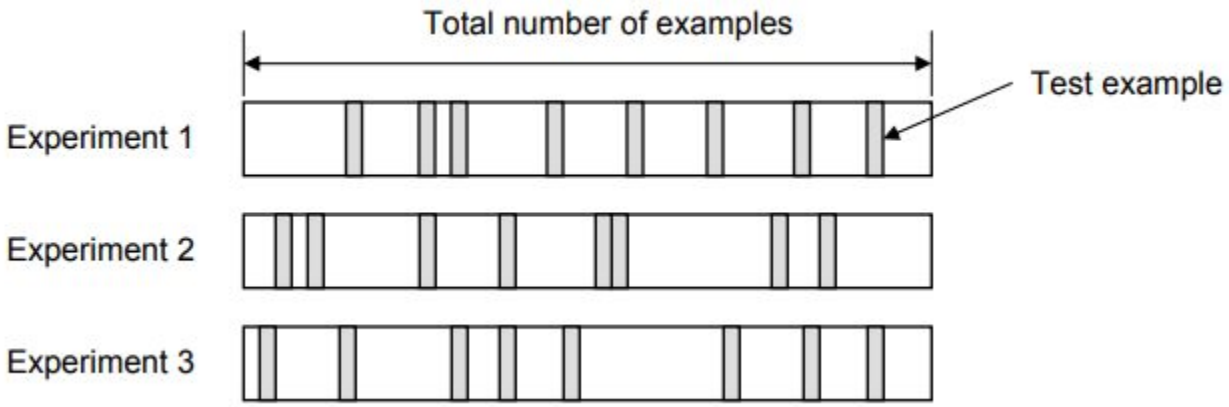
Leave-One-Out Cross-Validation (LOOCV)

- Todos os dados, exceto um registro, são usados para treinamento
  - Um registro é usado para teste.
- Processo é repetido por N vezes se houver registros N.
  - A vantagem é que (N-1)/ são usados para treinamento
- Desvantagem: custo computacional: N **rodadas**



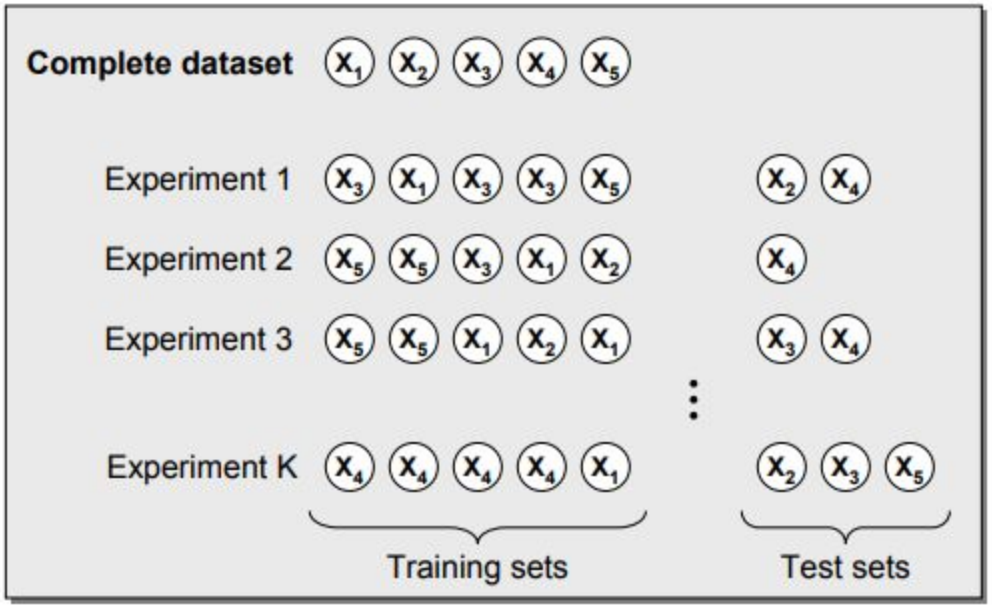
Random subsampling

- Amostras são escolhidos aleatoriamente para formar o ***test set***.
- Os dados restantes formam o ***train set***
- Bem utilizada na prática: pelo menos 30 execuções/rodadas



Bootstrapping

- Train set é selecionado aleatoriamente com **substituição / repetição**.
- Os exemplos restantes que não foram selecionados para treinamento são usados para teste.
- Diferente da validação cruzada de K-fold,
  - é provável que o valor mude de fold para fold.



Avaliação de Classificadores

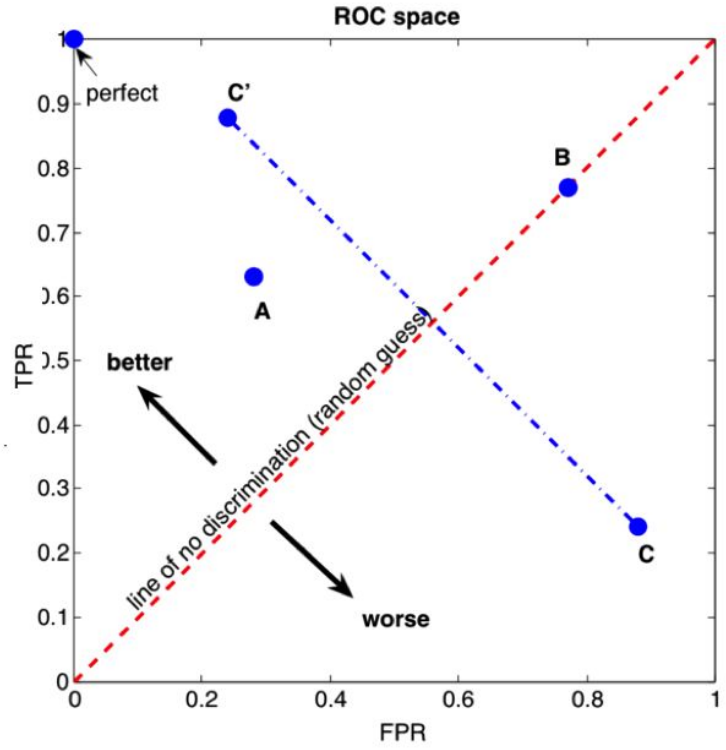
Exemplos de Classificadores

A			B			C			C′		
TP=63	FN=37	100	TP=77	FN=23	100	TP=24	FN=76	100	TP=76	FN=24	100
FP=28	TN=72	100	FP=77	TN=23	100	FP=88	TN=12	100	FP=12	TN=88	100
91	109	200	154	46	200	112	88	200	88	112	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		
F1 = 0.66			F1 = 0.61			F1 = 0.22			F1 = 0.81		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		

Espaço ROC

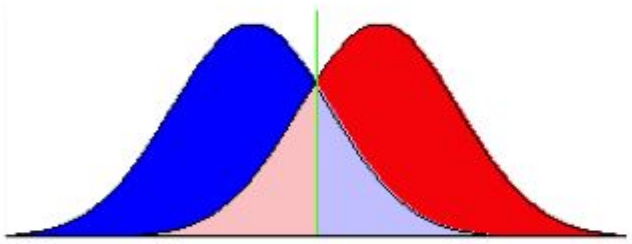
ROC (*Receiver Operating Characteristic*) é uma representação gráfica que ilustra o desempenho de um classificador binário em um determinado ponto de operação.

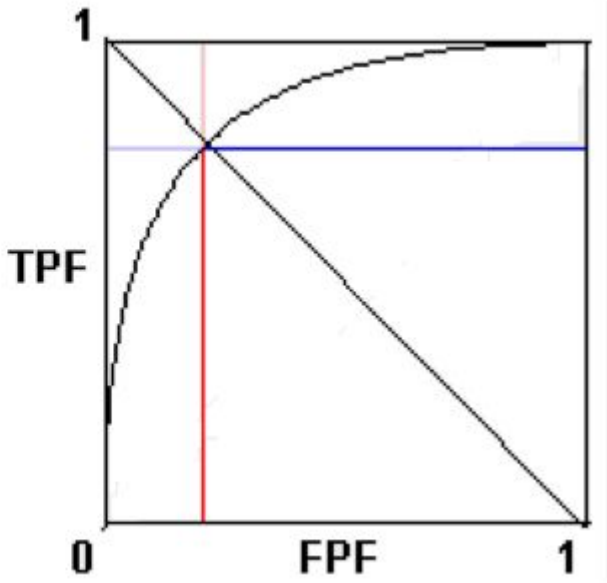
- Classificadores do exemplo anterior no espaço **ROC**.



Curva ROC

- A curva ROC mostra todos os pontos de operação do classificador (relação entre TP e FP)
- Para cada ponto, existe um limiar associado.
- EER (*Equal Error Rate*):
  - Ponto no gráfico no qual FPR é igual a 1-TPR
- Quando não existe um ponto operacional específico, usamos EER.





Curva ROC

- Considere 20 amostras, 10 classificadas corretamente (+) e 10 classificadas incorretamente (-) com suas respectivas probabilidades.

#	Classe	Score	#	Classe	Score
1	+	0.90	11	-	0.70
2	+	0.80	12	-	0.53
3	+	0.60	13	-	0.52
4	+	0.55	14	-	0.505
5	+	0.54	15	-	0.39
6	+	0.51	16	-	0.37
7	+	0.40	17	-	0.36
8	+	0.38	18	-	0.35
9	+	0.34	19	-	0.33
10	+	0.30	20	-	0.10

Curva ROC

Exemplo

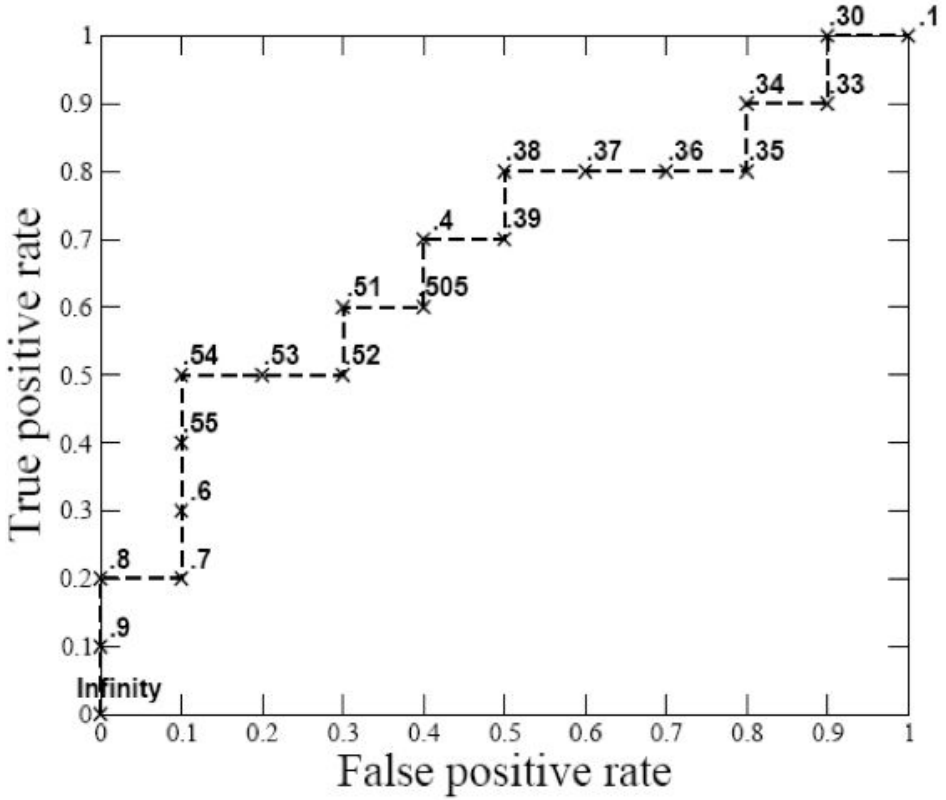
- Considere 20 amostras, 10 classificadas corretamente (+) e 10 classificadas incorretamente (-) com suas respectivas probabilidades.

#	Classe	Score	#	Classe	Score
1	+	0.90	11	-	0.70
2	+	0.80	12	-	0.53
3	+	0.60	13	-	0.52
4	+	0.55	14	-	0.505
5	+	0.54	15	-	0.39
6	+	0.51	16	-	0.37
7	+	0.40	17	-	0.36
8	+	0.38	18	-	0.35
9	+	0.34	19	-	0.33
10	+	0.30	20	-	0.10

Curva ROC

Exemplo

- Após ordenar os dados usando as probabilidades, temos a seguinte curva ROC



Curva ROC

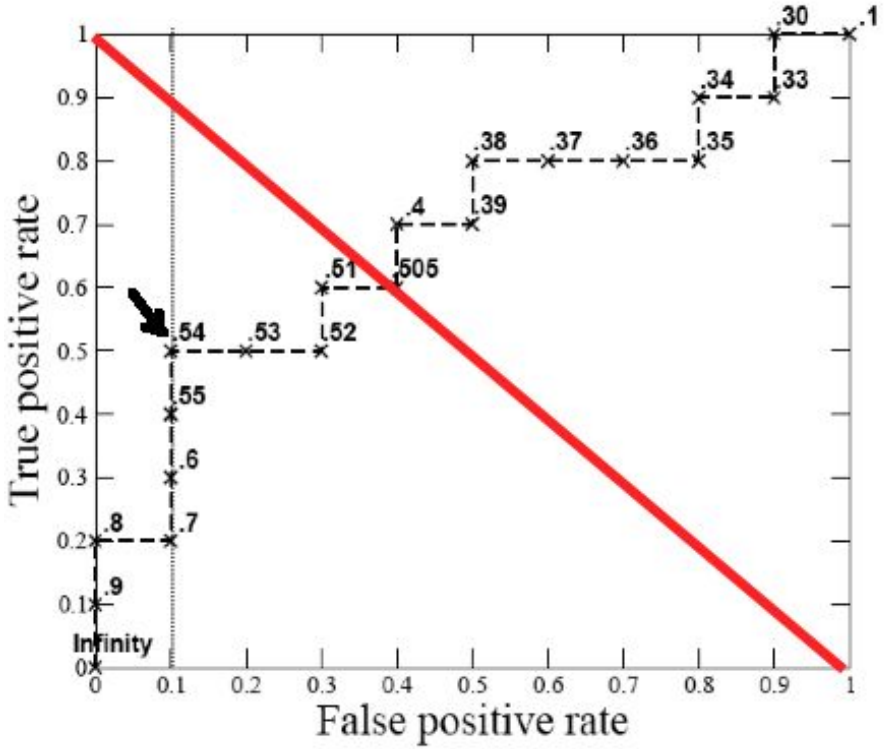
Exemplo

- Suponha que a especificação do seu sistema diga que o máximo **FPR** permitido é de 10%
- Qual seria o ponto de operação do sistema (limiar) ? **[0,7 ; 0,54]**

- Para esse limiar, qual seria a taxa de acerto do sistema?

$$\frac{(Pos \times TPR) + (Neg - (FPR \times Neg)}{N}$$

- 70% para um limiar de 0,54



Curva ROC

Classes desbalanceadas

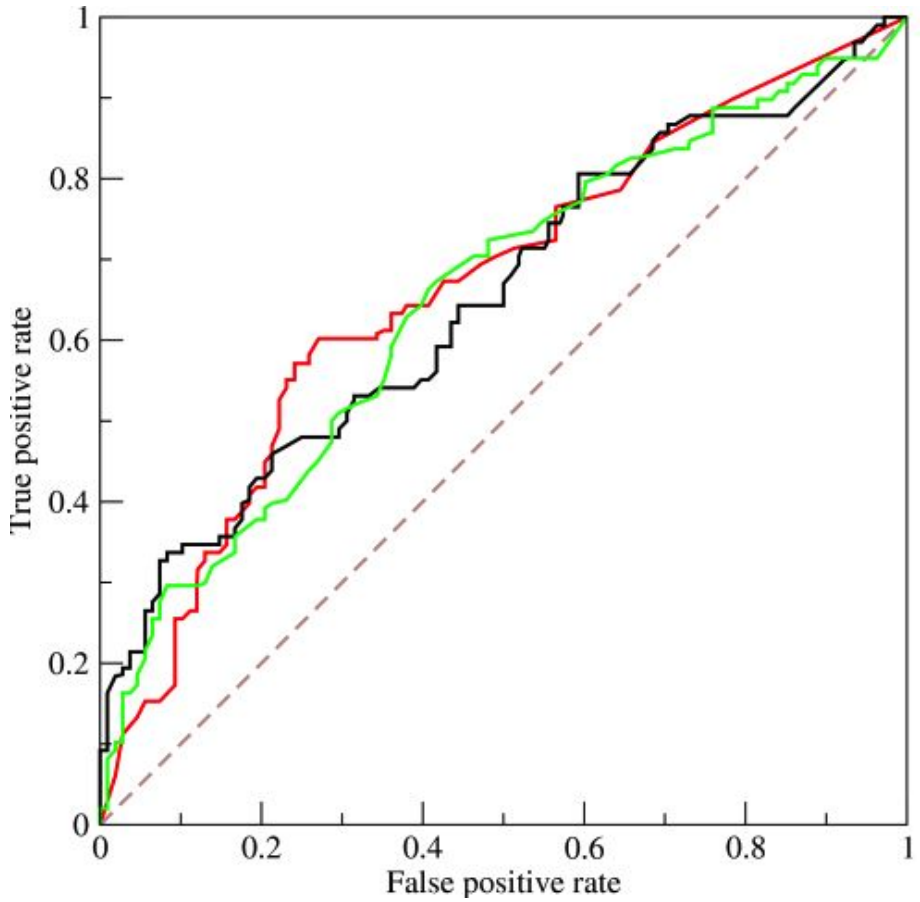
- Curvas ROC têm uma propriedade interessante:
  - não são sensíveis a mudanças de distribuição das classes

- Se a proporção de instâncias negativas e positivas na base de teste muda, a curva continua a mesma.
  - A curva é baseada em TP e FP.

- Isso permite uma fácil visualização do desempenho dos classificadores independentemente da distribuição das classes.

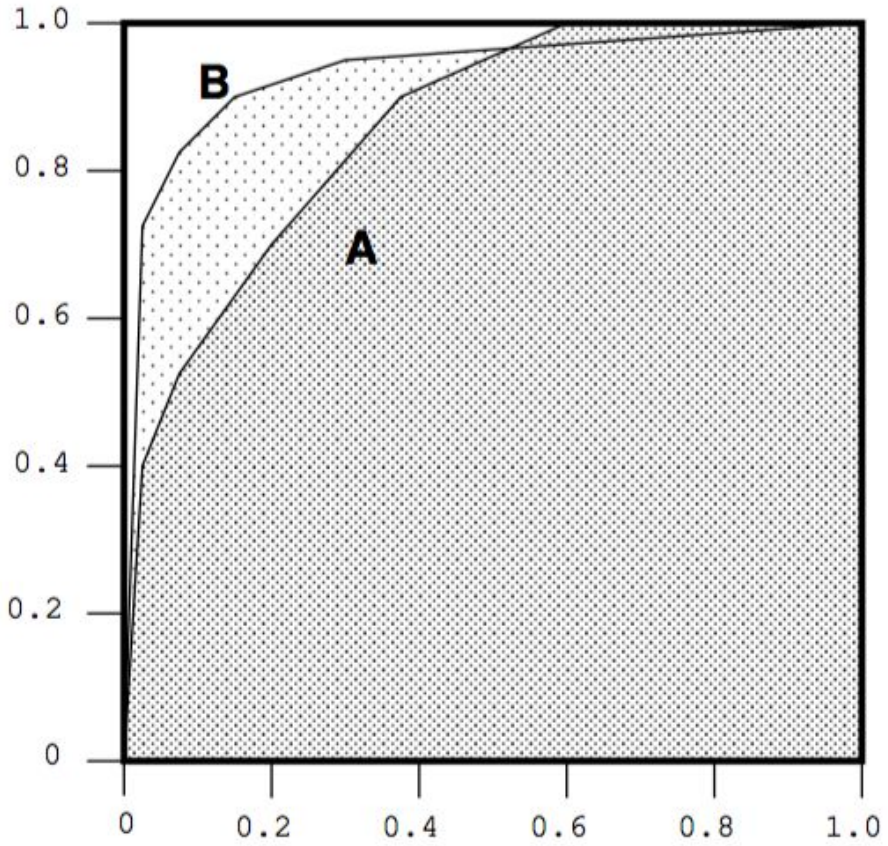
Comparando Classificadores

Qual seria o melhor classificador?



AUC

Area Under the Curve



Comparando Classificadores

Comparando Classificadores

- Suponha que você tenha dois classificadores que produziram as seguintes taxas de erro nos 10 folds de validação da sua base de dados.

$$C_1 = [10,15,7,6,13,11,12,9,17,14]$$

e

$$C_2 = [17,14,12,16,23,18,10,8,19,22]$$

- O erro médio de C_1 = 11,4% enquanto o erro médio de C_2 = 15,9%
- Podemos afirmar que C_1 é melhor do que o C_2?
- A média é bastante sensível a outliers.
  - O desempenho muito bom em um fold pode compensar o desempenho ruim em outro.

Comparando Classificadores

- Uma ferramenta bastante útil nesses casos é o teste de hipóteses.
  - t-Teste
    - Requer que as diferenças entre as duas variáveis comparadas sejam distribuídas normalmente.
    - Tamanho das amostras seja grande o suficiente (~30).

- A natureza do nosso problema não atende nenhum dos critério acima.

- Alternativa é o ***Wilcoxon signed-rank test***

Wilcoxon Signed-Rank test

- Teste não paramétrico que ordena as diferenças de desempenho de dois classificadores, ignorando os sinais, e compara a posição das diferenças (positivas e negativa).
- O teste consiste em:
  - Definir as hipóteses
    (H_0 - Não existe diferença e H_1 - Existe diferença)
  - Calcular as estatísticas do teste
  - Definir o nível de significância do teste (α)
  - Rejeitar ou aceitar H_0, comparando o valor da estatística calculada com o valor crítico tabelado.
- **Nível de significância**: Em aprendizagem de máquina, $\alpha = 0,05$ é um valor bastante utilizado. Isso quer dizer que existe 5 chances em 100 da hipótese ser rejeitada quando deveria ter sido aceita, ou seja, uma probabilidade de erro de 0,05 = 5%.

Wilcoxon Signed-Rank test

- Considere o exemplo anterior dos classificadores C_1 e C_2

C_1	C_2	Diff
10	17	7
15	14	1
7	12	5
6	16	10
13	23	10
11	18	7
12	10	2
9	8	1
17	19	2
14	22	8

Differenças	1	1	2	2	5	7	7	8	10	10
	1	2	3	4	5	6	7	8	9	10
Posição	1.5	1.5	3.5	3.5	5	6.5	6.5	8	9.5	9.5
Sinal	+	+	+	-	-	-	-	-	-	-

- Estatísticas: $N = 10$,
- $W^+ = 6{,}5$ (Soma das posições com sinal +)
- $W^+ = 48{,}5$ (Soma das posições com sinal -)
- Comparar os valores das estatísticas calculadas com os valores tabelados.

Wilcoxon Signed-Rank test

Valores críticos bi-caudais para o teste de Wilcoxon

Rejeição de H_0 para $\sum R-$ ou $\sum R+$ fora do intervalo dado pelos valores delimitados entre Inferior e Superior

	α							
	0,1		0,05		0,02		0,01	
número de pares com diferenças não-nulas	Inferior	Superior	Inferior	Superior	Inferior	Superior	Inferior	Superior
1								
2								
3								
4								
5	0	15						
6	2	19	0	21				
7	3	25	2	26	0	28		
8	5	31	3	33	1	35	0	36
9	8	37	5	40	3	42	1	44
10	10	45	8	47	5	50	3	52
11	13	53	10	56	7	59	5	61
12	17	61	13	65	9	69	7	71
13	21	70	17	74	12	79	9	82
14	25	80	21	84	15	90	12	93
15	30	90	25	95	19	101	15	105

- Comparar os valores das estatísticas calculadas ($W^+ = 6{,}5$ e $W^- = 48{,}5$) com os valores tabelados.
- Os valores críticos de W para N=10 e $\alpha = 0{,}05$ são 8 e 47.
- Como os valores calculados estão fora do intervalo crítico, rejeita-se a hipótese que não há diferença (H_0), ou seja, aceita-se H_1 (existe diferença estatística entre os dois classificadores).

Wilcoxon Signed-Rank test

O teste de Wilcoxon pode ser encontrado em diversas ferramentas

- Matlab (`signedrank`)
- Python Scipy (`from scipy import stats`)

```
Python
>>> c1 =[10, 15, 7, 6, 13, 11, 12, 9, 17, 14]
>>> c2 =[17, 14, 12, 16, 23, 18, 10, 8, 19, 22]
>>> from scipy.stats import wilcoxon
>>> wilcoxon(a,b)
(6.5, 0.031865021445197136)

MATLAB
>> [p,h,stats] = signrank(c1,c2)
p = 0.0332
h = 1
stats = signedrank: 6.5000
```

Wilcoxon Signed-Rank test

- Um teste **não paramétrico** utilizado para comparar diversos classificadores é o teste de **Friedman**
  - Semelhante ao ANOVA

Referências

- J. Demsar,
  Statistical Comparisons of Classifiers over Multiple Data Sets,
  J. of M. Learning Research, 7:1-20, 2006.

- S. Salzberg,
  On Comparing Classifiers: Pitfalls to avoid and recommended approach,
  Data Mining and Knowledge Discovery, 1:317-327, 1997.

- Luiz E. S. Oliveira,
  Avaliando Classificadores,
  Notas de Aulas, DInf / UFPR, 2017.

Referências

- Wikipedia, **Sensitivity & Specificity**
  https://en.wikipedia.org/wiki/Sensitivity_and_specificity
- DZone - Ai Zone
  Machine Learning: Validation Techniques
  https://dzone.com/articles/machine-learning-validation-techniques
-