INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
MATO GROSSO

Engenharia da Computação - Inteligência Artificial

Avaliação de Classificadores

Prof. Dr. Ruy de Oliveira

IFMT

Avaliação de um Classificador

- Qual é a efetividade do modelo criado?
  - *Que medida de desempenho deve ser usada?*

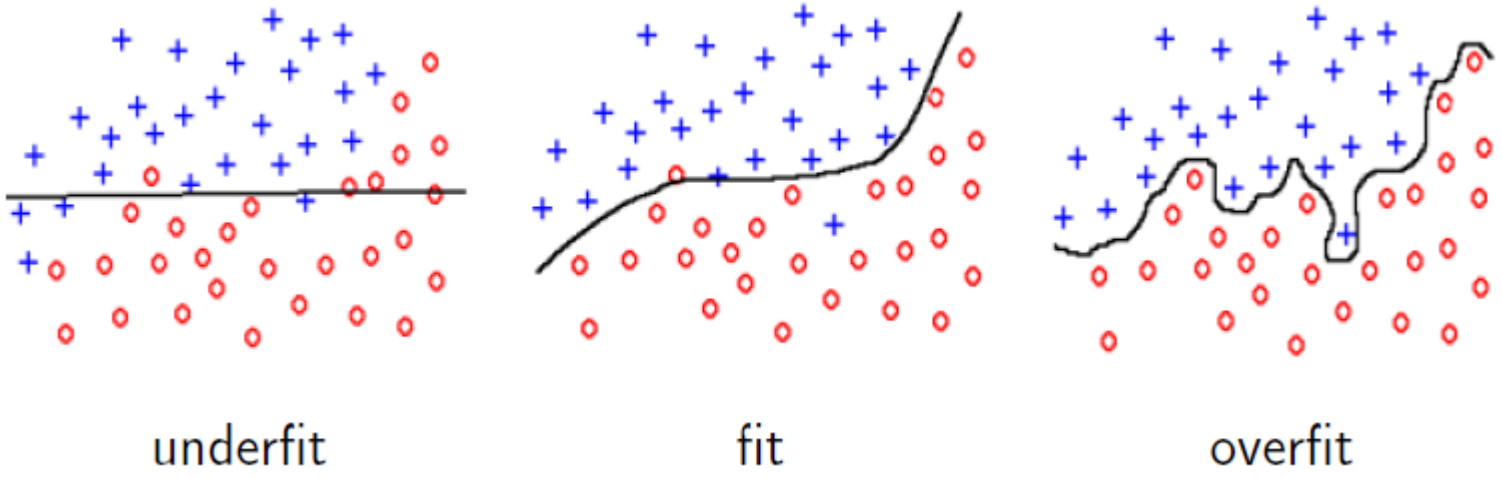
- *Medida de desempenho natural para problemas de classificação: taxa de erro em um conjunto de teste*
  - *Sucesso: exemplo classificado corretamente*
  - *Erro: exemplo classificado erroneamente*
  - *Taxa de erro: proporção de erros cometidos sobre sobre o conjunto de dados por completo*
  - *Exatidão (accuracy): proporção de exemplos classificados corretamente sobre o conjunto de dados por completo*

$$\text{exatidão} = 1 - \text{taxa de erro}$$

Risco da superadaptação

- Memorizar os dados de treinamento muito precisamente usualmente resulta em classificação pobre sobre dados novos

- Classificadores devem ter a habilidade de generalizar!



Dados de treinamento vs. Dados de teste

Problema: apenas uma quantidade finita de dados está disponível e tem de ser usada para as fases de treinamento e teste

- Mais dados de treinamento melhorar a generalização
- Mais dados de teste possibilita melhor estimativa para a probabilidade de erro de classificação
- Nunca se deve avaliar desempenho com os dados de treinamento
  - As conclusões tenderiam a ser muito otimistas

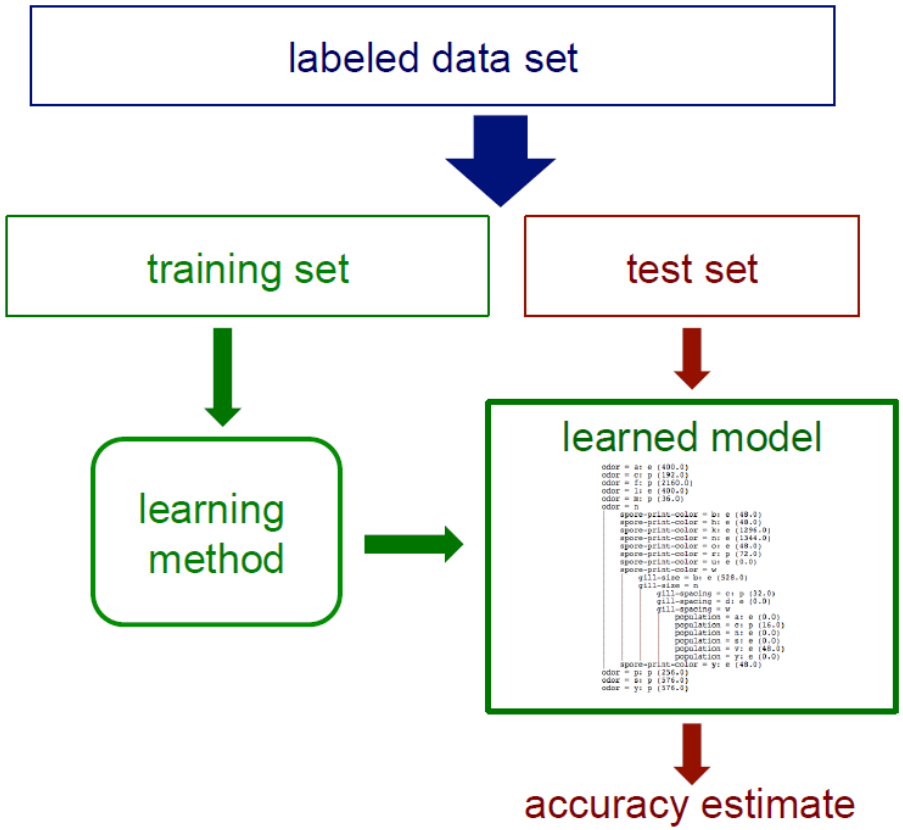
- O problema ocorre quando a base de dados não é grande o suficiente para ser dividida em duas!

Métodos de avaliação de classificador

Todos os métodos utilizam dados separados para as fases de treinamento e teste

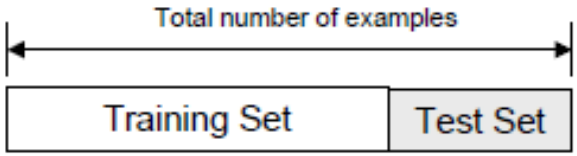
- Base de dados de treinamento
- Base de dados de teste (avaliação)

- Métodos mais populares
  - Hold out
  - Validação cruzada (Cross validation)
  - Bootstrap



O método Hold out

- Divide a base de dados em dois grupos
  - Dados de treinamento: para treinar o classificador
  - Dados de teste: para estimar a taxa de erro do classificador treinado



- Requer base de dados relativamente grande!
- Há o risco de a divisão escolhida ser inadequada (desbalanceada) → taxa de erro enganosa

O método Hold out

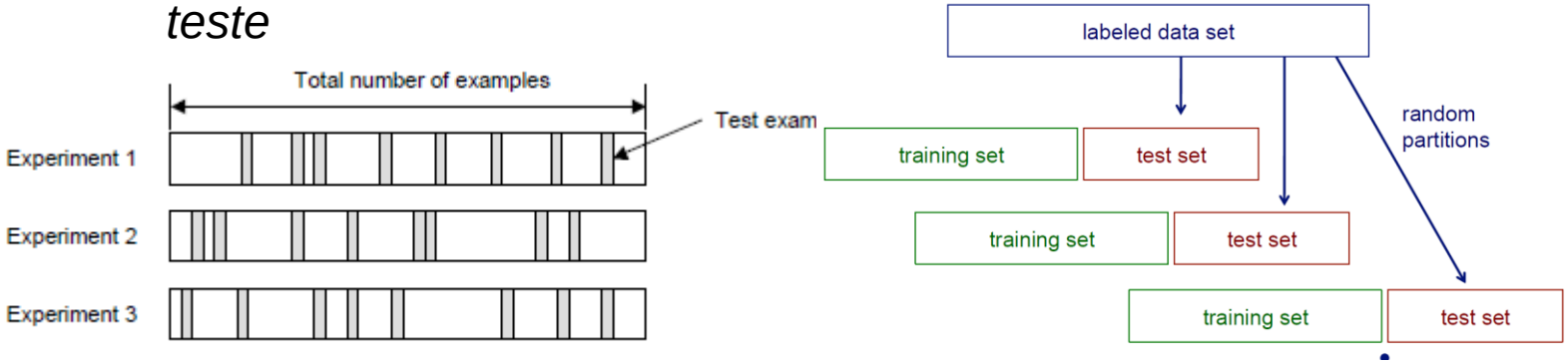
As limitações do método Hold out podem ser superadas com uma família de métodos de reamostragem (*resampling*)

- Validação cruzada
  - Subamostragem aleatória (random subsampling)
  - *K-fold*
  - *Leave-one-out*
- *Bootstrap*

Subamostragem aleatória

- Também conhecido como *Monte Carlo crossvalidation*

- Divide a base de dados em k partes
  - *Cada parte é composta de um número fixo de exemplos, sem reposição*
  - *Para cada parte o classificador é treinado novamente com os exemplos de treinamento e o erro E_i é estimado com exemplos de teste*

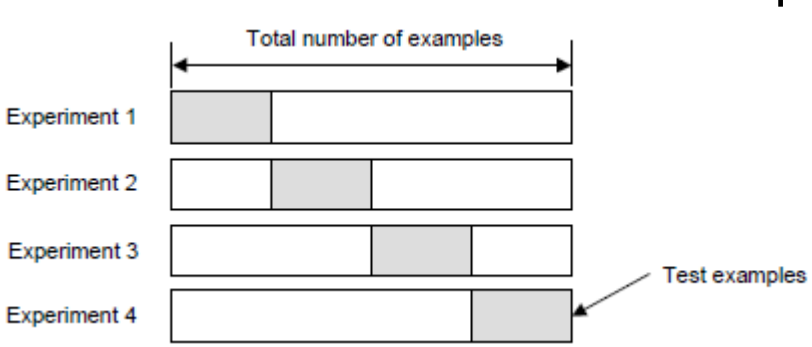


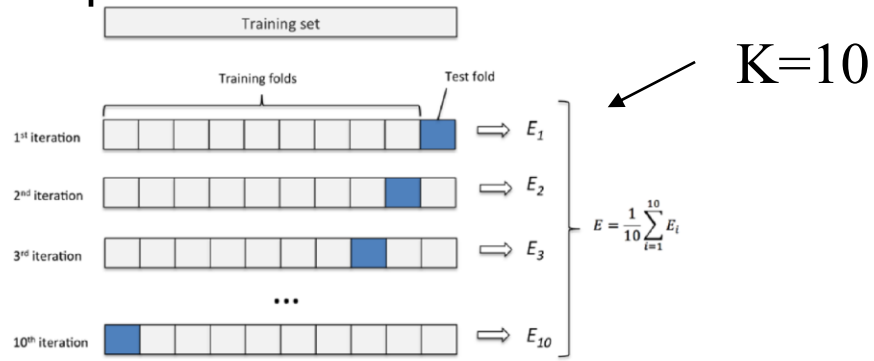
- *A estimativa de erro final é obtida da média dos E_i da partes*
  - *Provê resultado superior ao médodo hold out*

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Método *k-fold*

- Cria k partições da base de dados
  - Para cada um dos k experimentos, usa k-1 partições para o treinamento e as demais partições para o teste





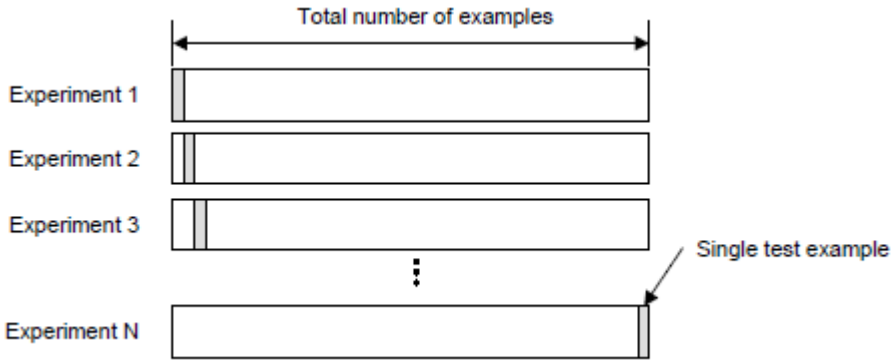
- *K-fold é similar ao método subamostragem aleatória*
  - *A vantagem da validação cruzada k-fold é que no final todos os exemplos da base de dados são usados para ambos o treinamento e o teste*

- *Analogamente ao caso anterior, o erro total é estimado como a média do erro de cada experimento*

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Método *leave-one-out*

- Cria k partições da base de dados, mas k é escolhido como o número total de exemplos
  - *Para uma base de dados com N exemplos → N experimentos*
  - *Para cada experimento, usa N-1 exemplos para treinamento e o exemplo restante para teste*



- *Como usual, o erro final é computado como a média do erro dos experimentos*

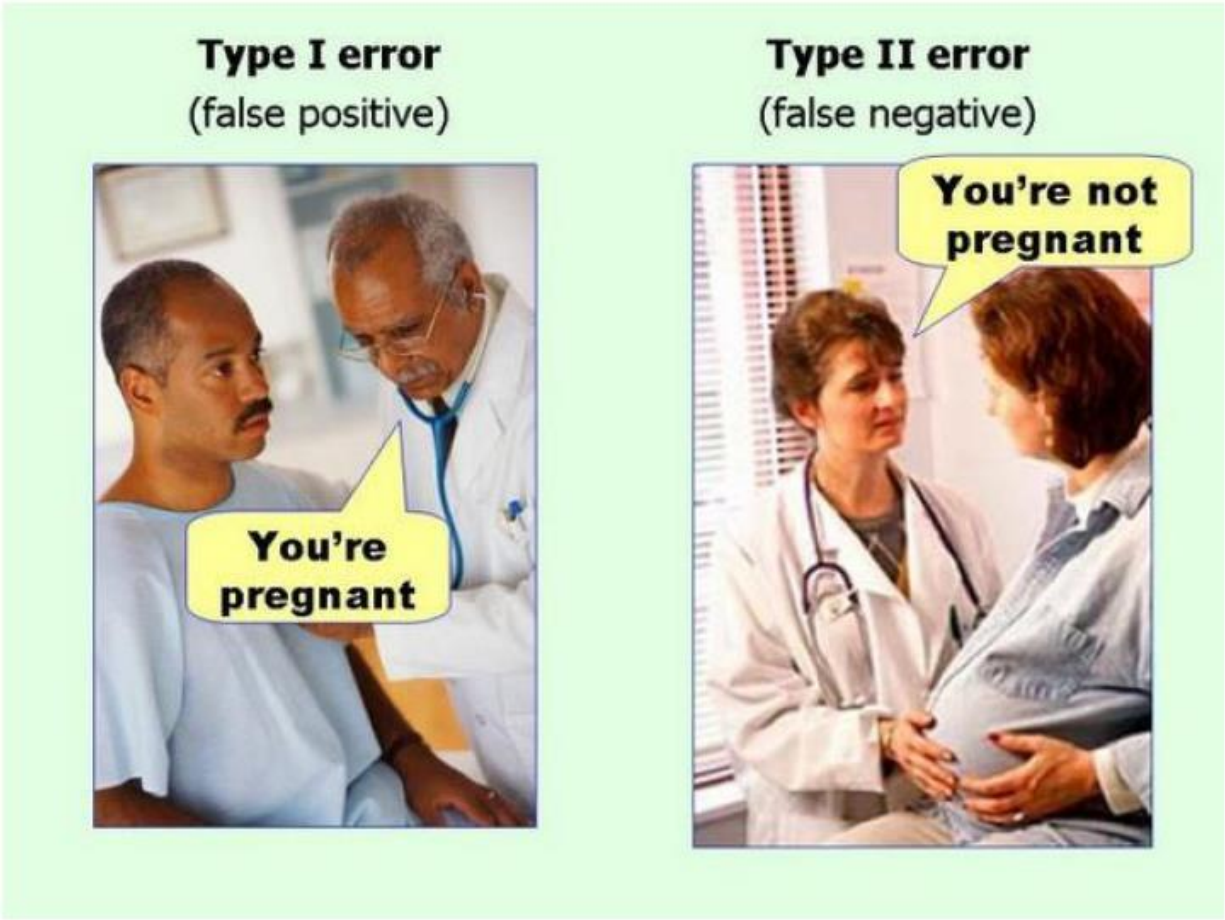
$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Qual é o número ideal de partições?

- Número grande de partições
  - O vício (bias) da taxa de erro real será pequeno
  - A variância do erro será grande
  - O tempo de processamento será muito grande
- Número pequeno de partições
  - O número de experimentos (e tempo de processamento) serão reduzidos
  - A variância do erro será pequena
  - O vício (bias) da taxa de erro real será grande
- Na prática, a escolha do número de partições depende do tamanho da base de dados
  - Para base de dados grande, até 3-fold proverá resultado preciso
  - Para base de dados esparsas, deve-se usar o método leave-one-out, de modo a treinar o máximo possível de exemplos
- Uma escolha comum para validação cruzada k-fold é k=10

Falso Positivo e Falso Negativo (tipo 1 e 2)

- O classificações erradas pode ser muito alto!



Matriz Confusão

- A matriz confusão (MF) resume os resultados da estimativa de um classificador
- A MF mostra o quão “confuso” o classificador está quando fazendo as estimativas
- A MF provê informações não apenas acerca dos erros, mas também sobre os **tipos de erros que o classificador está cometendo**, o que é ainda mais importante

		Classe Estimada	
		Pos	Neg
Classe Real	Pos	*TP*	*FN*
	Neg	*FP*	*TN*

$$\text{exatidão (accuracy)} = \frac{TP + TN}{TP + FP + FN + TN}$$

Exemplos de Matriz Confusão

- A matriz confusão (MF) de duas e três classes

predicted → / real ↓	*Class_pos*	*Class_neg*
Class_pos	114	86
Class_neg	7	93

predicted → / real ↓	*Class_1*	*Class_2*	*Class_3*
Class_1	94	16	10
Class_2	21	113	16
Class_3	4	4	92

Predicted

Actual	Iris-setosa	Iris-versicolor	Iris-virginica	Σ
Iris-setosa	100.0 %	0.0 %	0.0 %	50
Iris-versicolor	0.0 %	88.7 %	6.4 %	50
Iris-virginica	0.0 %	11.3 %	93.6 %	50
Σ	50	53	47	150

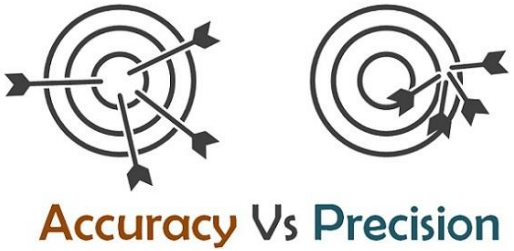
A medida exatidão (*accuracy*)

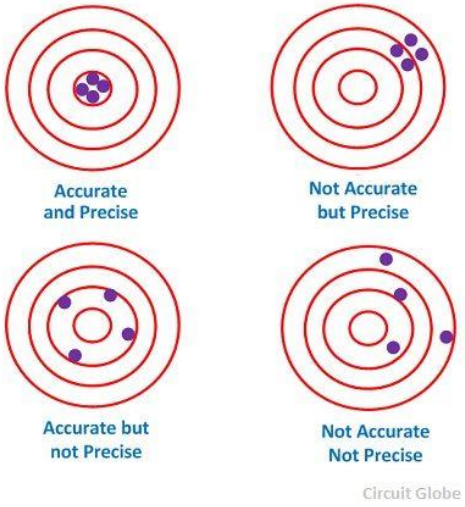
- A medida exatidão pode ser inútil em casos onde:
  - Há uma variação significativa na quantidade das classes
    - 98% de exatidão é bom, se 97% dos exemplos (instâncias) são negativos?
  - Há diferentes custos para a classificação errada (um positivo errado custa mais do que um negativo errado)
    - Na área médica, um falso positivo resulta num exame extra, mas um falso negativo resulta numa falha referente ao tratamento da doença

		Classe Estimada	
		Pos	Neg
Classe Real	Pos	*TP*	*FN*
	Neg	*FP*	*TN*

Principais métricas para avaliação de classificadores

- *Exatidão (Accuracy) =* ***(TP+TN)/(P+N)***
  - *Proximidade da medida com o valor correto*
- *Erro =* ***(FP+FN)/(P+N)***
- *Precisão (Precision) =* ***TP/(TP+FP)***
  - *Proximidade entre as várias medidas*
  - *Quantidade de exemplos que o classificador identificou como positivo e são realmente positivos*





		Classe Estimada	
		Pos (P)	Neg (N)
Classe Real	Pos	*TP*	*FN*
	Neg	*FP*	*TN*

Principais métricas para avaliação de classificadores

- *Revocação (Recall) – Sensibilidade =* ***TP/(FN+TP)***
  - *Fração de instâncias relevantes que são recuperadas*
  - *Que porcentagem de exemplos positivos resultou em positivo*
  - *Ex.: essa métrica pode nos informar a proporção de pacientes que realmente tinha câncer e foi diagnosticado pelo algoritmo como tendo câncer*

Classe Estimada

Classe Real	Pos (P)	Neg (N)
Pos	*TP*	*FN*
Neg	*FP*	*TN*

O custo do falso negative pode ser muito alto!!!

Outra métricas para avaliação de classificadores

- *Especificidade = **TN/TN+FP***
  - *Representa o contrário do Recall*
  - *Ex.: esta métrica pode nos informar a proporção de pacientes que não tiveram câncer, e foram selecionados pelo classificador como não cancerígenos*

		Classe Estimada	
		Pos	Neg
Classe Real	Pos	*TP*	*FN*
	Neg	*FP*	*TN*

Exercício1: matriz confusão

- Calcule para a matriz confusão abaixo:
  - Exatidão
  - Erro
  - Precisão
  - Sensibilidade
  - Especificidade

	Predicted					
Ground Truth		Class1	Class2	Class3	Class4	Class5
	Class1	92	3	2	2	1
	Class2	2	92	2	2	2
	Class3	1	1	92	6	0
	Class4	0	1	1	92	6
	Class5	1	4	2	1	92

Exercício2: avaliação do classificador

- Modelar um classificador Árvore de decisão para a base de dados Íris, disponível no link abaixo
  - (https://archive.ics.uci.edu/ml/datasets/iris)
- O classificador deve utilizar inicialmente a métrica Ganho de informação e depois a métrica índice GINI, como medida de incerteza
- Depois de treinando, o classificador deve ser avaliado pelos métodos: *hold out* e *k-fold (k=5 e 10)*
  - Compare: exatidão, precisão e taxa de erro
- Mostre os resultados, considerando as duas métricas de incerteza acima, e também na matriz confusão
- Compare os resultados da avaliação com os dados de teste e com os dados originais

Base de dados

- Repositório online com várias bases de dados
  - https://archive.ics.uci.edu/ml/datasets.html
- Iris
  - *https://gist.github.com/curran/a08a1080b88344b0c8a7#file-iris-csv*
  - *https://archive.ics.uci.edu/ml/datasets/Iris*
- Breast Cancer Wisconsin (Diagnostic)
  - *https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29*
- Sonar (Mines vs. Rocks)
  - https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Sonar,+Mines+vs.+Rocks)
- Titanic
  - https://www.kaggle.com/c/titanic/data

Links interessantes

- K-fold
  - https://www.talend.com/blog/2017/05/15/machine-learning-algorithms-with-k-fold-cross-validation/

Bibliografia

- RUSSEL, S. Inteligência Artificial. 3ª ed. Campus, 2013
- BRAGA, A. de P. *Redes Neurais Artificiais*. 2ª ed. Rio de Janeiro: LTC, 2007
- HAYKIN, S. S. Redes Neurais. 2a ed. Porto Alegre: Bookman, 2000
- COPPIN, B. Inteligência Artificial. 1a ed. LTC, 2010
- LUGER, G. F. Inteligência Artificial. 6a ed. Pearson, 2013

Bibliografia

- NORVIG, P. Inteligência Artificial. 2a ed. Campus, 2004
- ROSA, J. L. G. Fundamentos da Inteligência Artificial. 1a ed. LTC, 2011
- SILVA, I. N. da Redes Neurais Artificiais para Engenharia e Ciências Aplicadas. 1a ed. São Paulo: Artliber, 2010