

Experimentação, Parametrização e Análise de Algoritmos Bio-inspirados Parte 2/2

Gisele L. Pappa

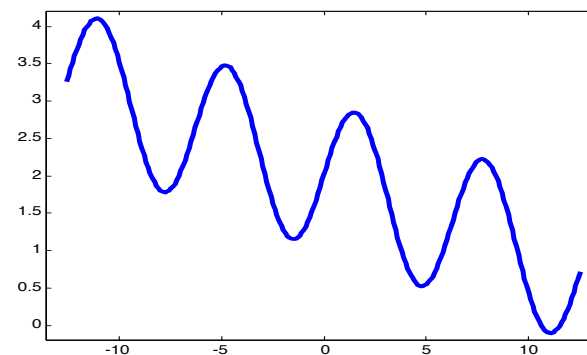
Otimização versus Aprendizado

Trabalhando com dados

- A maioria dos métodos de computação natural trabalha com dados
- A metodologia de experimentação para otimização é muito diferente da de aprendizado
- Em otimização o objetivo é, através de um conjunto de dados, encontrar a solução ótima para o problema
 - A fitness é calculada utilizando a base de dados completa

Exemplo

x	saída
-10	3.6
-8	2.1
-6	2.7
-4	2.6
.	.
.	.



Trabalhando com Dados

- Aprendizado de máquina
 - O objetivo é criar um modelo que depois vai ser utilizado em um novo conjunto de dados
- O tipo do modelo criado depende do problema sendo resolvido, incluindo:
 - Classificação (Predição)
 - Regressão
 - Agrupamento

Classificação

Conjunto de
Treinamento

A_1, A_2, A_3, C

0, 0, 1, 1

1, 1, 0, 1

....

0, 1, 0, 0



Algoritmo de
Aprendizado



Modelo criado
a partir dos
dados

Classificação

Conjunto
de teste

A_1, A_2, A_3, C
1, 0, 0, ?
0, 1, 0, ?
....
1, 1, 1, ?



Modelo criado
a partir dos
dados



A_1, A_2, A_3, C
1, 0, 0, 1
0, 1, 0, 0
....
1, 1, 1, 1

Agrupamento

Conjunto de
Treinamento

A_1, A_2, A_3

0, 0, 1

1, 1, 0

....

0, 1, 0



Algoritmo de
Agrupamento



Modelo criado
a partir dos
dados

- Nesse caso, a quantidade de classes é desconhecida. Cabe ao algoritmo determinar e agrupar os dados de acordo com suas características

Trabalhando com Dados em Computação Natural

- Dados são divididos em 2 ou 3 partes (dependendo do problema)
 - Treinamento – utilizados para criar o modelo, normalmente antes do/durante o cálculo da fitness
 - Validação – utilizado para avaliar o modelo criado, e determinar a medida de fitness
 - Teste – utilizado para testar o melhor indivíduo retornado

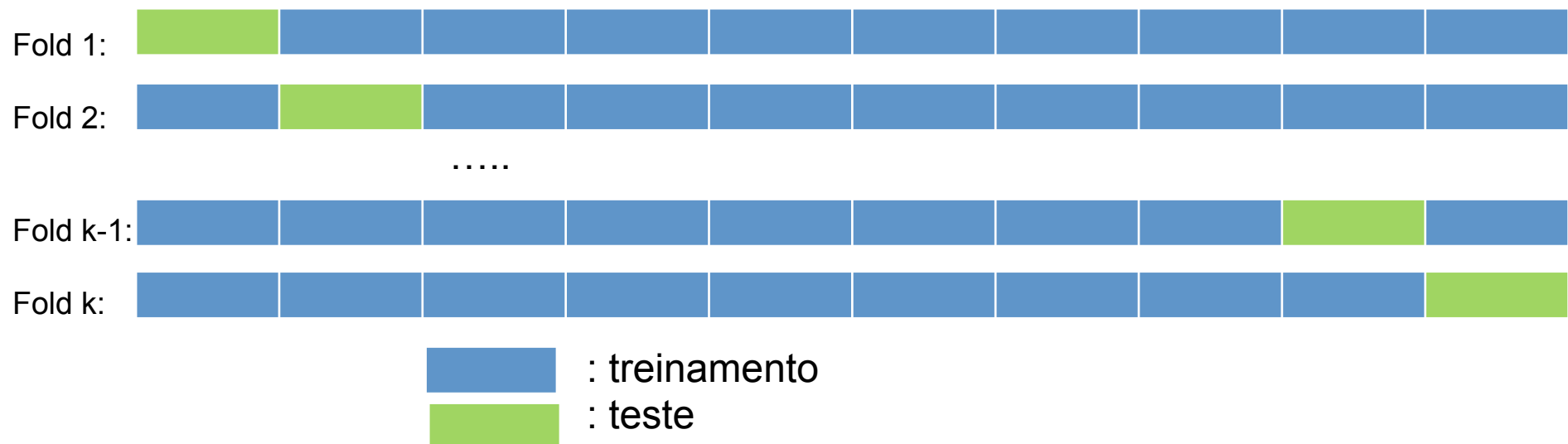
Aprendizado

- Em tarefas de aprendizado, os dados devem ser divididos em pelo menos dois conjuntos:
 - Conjunto de treinamento, onde a fitness é calculada
 - Conjunto de teste, onde o melhor indivíduo retornado é avaliado

Validação Cruzada

- Garante validação estatística dos resultados
- Como funciona?
 - Passo 1: os dados são divididos em k subconjuntos de mesmo tamanho
 - Passo 2: em cada instante um subconjunto é usado para teste e os demais para treinamento
- Chamada de *validação cruzada de fator k*

Validação Cruzada de K-partições



Validação Cruzada de K-partições

- Normalmente os subconjuntos são estratificados (divididos de forma que a distribuição das classes sejam mantidas) antes de realizar a validação cruzada
- Faz-se a média das estimativas de erro para obter o erro estimado geral

Validação Cruzada

- Método padrão para avaliação: validação cruzada estratificada com fator 5/10
- A realização de vários experimentos tem demonstrado que 10 é a melhor escolha
- A estratificação reduz a variância da estimativa
- Opção ótima: validação cruzada estratificada com fator 10 repetida (10 x)
 - Nem sempre viável devido ao custo computacional

Tuning automático de parâmetros

- Existem diversas ferramentas que testam diferentes combinações de parâmetros automaticamente
- Ferramenta muito utilizada atualmente: iRace
 - <http://iridia.ulb.ac.be/irace/>
- Baseado no algoritmo Iterated F-race (I/F-Race)
- Funciona em 2 fases:
 - Tuning
 - Teste

iRace (pacote R)

- Usa o teste estatístico não-paramétrico de Friedman para analisar a variância de várias configurações de parâmetros através de um ranking
 - Cada parâmetro a ser otimizado segue uma distribuição normal (atributos numéricos) ou discreta (atributos categóricos)
- Ao longo das iterações, as distribuições são atualizadas de forma a aumentar a probabilidade de amostrar parâmetros próximos aos melhores encontrados até o momento

Leitura Recomendada

- **A study of cross-validation and bootstrap for accuracy estimation and model selection**, R. Kohavi, Proc. of the 14th International Joint Conference on Artificial Intelligence 2 (12): 1137–1143, 1995
- The irace Package: Iterated Racing for Automatic Algorithm Configuration, Relatório técnico
 - <http://iridia.ulb.ac.be/IridiaTrSeries/link/IridiaTr2011-004.pdf>

Experimentação, Parametrização e Análise de Algoritmos Bio-inspirados Parte 2/2

Gisele L. Pappa