Universidade Federal de Minas Gerais
Disciplina de Computação Natural
Profa. Gisele L. Pappa

Trabalho Prático 1
Programação Genética

Data de Entrega: 30 de setembro de 2018

---

1 Introdução

O principal objetivo deste trabalho é desenvolver conceitos chave para a construção de soluções para problemas usando Programação Genética (GP), envolvendo o entendimento e a implementação dos componentes básicos de um arcabouço de GP, bem como a análise de sensibilidade dos seus parâmetros (como eles afetam o resultado final, a natureza da convergência, etc) e procedimentos para avaliação das soluções alcançadas.

Um dos problemas mais populares que podem ser resolvidos com técnicas de programação genética é a regressão simbólica. Conforme visto em sala de aula, dado um conjunto de m amostras provenientes de uma função *desconhecida* $f : \mathbb{R}^n \mapsto \mathbb{R}$, representadas por uma dupla $\langle X, Y \rangle$ onde $X \in \mathbb{R}^{m \times n}$ e $Y \in \mathbb{R}^m$, o objetivo é encontrar a expressão simbólica de f que melhor se ajusta às amostras fornecidas.

No arcabouço de programação genética a ser desenvolvido, os indivíduos deverão ser representados por árvores, compostas por nós terminais e operadores. Será de sua responsabilidade determinar ambos os conjuntos para solucionar o problema de regressão simbólica fornecido. Lembre-se que é importante considerar a presença de constantes (para a representação de coeficientes), bem como das variáveis do problema.

Um critério de avaliação possível para medir a qualidade de um indivíduo é a raiz quadrada do erro quadrático médio normalizada (NRMSE), que pode ser calculada por:

$$fitness(Ind) = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \text{EVAL}(Ind, x_i))^2}{\sum_{i=1}^{N} (y_i - \bar{Y})^2}}$$

onde N é o número de instâncias fornecidas, *Ind* é o indivíduo sendo avaliado, EVAL(*Ind*, x_i) avalia o indivíduo *Ind* na *i-ésima* instância de X, y_i é a saída esperada para a entrada x_i e $\bar{Y}$ é a média do vetor de saídas Y.

O uso de uma métrica normalizada para avaliação de desempenho das soluções, apesar de esta não ser a ideal[1], facilita a comparação de resultados gerados pelo modelo em bases de dados diferentes.

[1]Apenas como curiosidade, o NRMSE pode não ser muito confiável nos casos de bases de dados muito pequenas. Para diminuir o efeito de tais problemas, uma divisão da raiz quadrada do erro quadrático médio (RMSE) pela amplitude interquartil seria mais confiável, porém mais complexa de se fazer.

Decisões de Implementação:

1. Como representar um indivíduo (genótipo);
2. Como gerar a população inicial;
3. Quais operadores genéticos serão utilizados;
4. Facilidades para variação de parâmetros—parâmetros *hardcoded* no arcabouço certamente dificultarão a avaliação dos parâmetros;

2 Operadores Elitistas

Independentemente dos operadores genéticos escolhidos por você, também será necessária a implementação de versões elitistas dos mesmos. **ATENÇÃO**: Não confunda esses operadores com o uso de elitismo ao longo das gerações, são dois conceitos diferentes!

Operadores elitistas permitem a adição do indivíduo gerado por eles à população apenas se este indivíduo for melhor, em questão de fitness, do que todos os pais que o geraram, i.e., o único pai para mutação e os dois no caso do cruzamento. Você deverá fazer uma análise dos efeitos dos usos desses operadores na qualidade das soluções, como descrito melhor nas próximas seções.

Funcionamento dos Operadores Elitistas

O fluxo de funcionamento dos operadores elitistas é bem simples:

1. Execute a operação genética normalmente.
2. Se o filho gerado for melhor que ambos os pais, adicione-o à próxima geração. Caso contrário, adicione o melhor pai à próxima geração.

3 Bases de Dados

Três conjuntos de dados serão utilizados neste trabalho prático, sendo dois problemas sintéticos e um real. Estão disponíveis no Moodle dois arquivos para cada base de dados:

1. <nome-da-base>-train.csv
2. <nome-da-base>-test.csv

O primeiro deverá ser usado para evoluir às soluções até o número máximo de gerações ser alcançado. Finalizada esta etapa, as melhores soluções encontradas deverão ser avaliadas utilizando a base de teste. Neste momento é importante comparar os indicadores de teste com os indicadores de treino a fim de detectar algum tipo de anomalia (overfitting, por exemplo).

Todos os arquivos estão no formato CSV, e a última coluna contém a saída desejada. Esta saída deverá ser comparada com a saída estimada.

4 Metodologia Experimental

O GP deverá ser executado nas 3 bases de dados fornecidas. Uma descrição das características de cada uma pode ser encontrada na Tabela 1 e visualizações das bases sintéticas em 3 dimensões podem ser vistas nas Figuras 1 e 2. Olhe para os gráficos e tente supor quais as operações matemáticas necessárias para gerar tais funções. Baseado nisso, escolha **uma** das bases sintéticas e realize um estudo dos parâmetros como descrito abaixo:

- Definir o tamanho máximo do indivíduo como 7. Esse parâmetro não precisa ser obrigatoriamente variado.
- Escolher o tamanho da população e o número de gerações apropriados. O tamanho da população pode ser testado, por exemplo, utilizando 50, 100, 500 indivíduos. O número de gerações pode também ser escolhido usando esses mesmos números. Mas como saber se o escolhido é o mais apropriado? Você pode avaliar como o aumento no número da população ou de gerações melhora a solução encontrada (em termos do erro gerado), se a população converge, etc.
- Testar duas configurações de parâmetros para cruzamento e mutação. Na primeira, a probabilidade de cruzamento (p_c) deve ser alta (por exemplo, 0.9), e a probabilidade de mutação (p_m) deve ser baixa (por exemplo, 0.05). Na segunda, p_c deve ser mais baixa (por exemplo, 0.6) e p_m mais alta (por exemplo, 0.3). Para ambas as configurações, deve-se avaliar o efeito do cruzamento e da mutação na evolução, isto é, em quantos casos esses operadores contribuem positivamente (os filhos gerados são melhores que os pais) ou negativamente para a evolução? A partir desse estudo inicial, que valores finais você proporia?
- Analisar as mudanças ocorridas quando o tamanho do torneio aumenta de 2 para 5 ou 3 para 7, dependendo do tamanho inicial da população.
- Analisar o impacto de usar somente operadores elitistas.
- Existe uma forma simples de medir bloating no seu algoritmo?

Lembrem-se que ao mexer em um dos parâmetros, todos os outros devem ser mantidos constantes, e que a análise dos parâmetros é de certa forma interativa. A configuração de parâmetros raramente vai ser ótima, mas pequenos testes podem melhorar a qualidade das soluções encontradas.

Utilizando os parâmetros considerados mais apropriados para a base escolhida, execute o GP para a base sintética restante e avalie a performance do algoritmo.

Agora que você já sabe quais parâmetros tiveram maior impacto nos resultados das bases sintéticas, escolha dois deles e faça um estudo mais detalhado para a base de dados real. Os parâmetros que levaram à melhor performance neste caso se parecem com os encontrados anteriormente? Tente encontrar justificativas para sua resposta.

Guia para execução dos experimentos

1. Escolha do tamanho da população e número de gerações (utilizar tamanho máximo do indivíduo como 7, elitismo, torneio de tamanho 2 e $p_c = 0.9$ e $p_m = 0.05$).

2. Após alguns testes, defina o tamanho da população e o número de gerações e varie p_c e depois p_m. Os parâmetros escolhidos no passo 1 ainda são apropriados?

3. Definidos o tamanho da população, número de gerações, p_c e p_m, aumente o tamanho do torneio.

4. Usando os valores ótimos encontrados para os parâmetros, compare os resultados encontrados quando utilizamos os operadores elitistas com e sem elitismo na população. As diferenças são significativas entre os dois cenários?

5. Se desejar, teste outras características, como métodos para garantir a diversidade da população.

Por ser um método estocástico, a avaliação experimental do algoritmo baseado em GP deve ser realizada com *repetições*, de forma que os resultados possam ser reportados segundo o valor médio obtido e o respectivo desvio-padrão. A realização de 30 repetições pode ser um bom ponto de partida (lembrando que desvio-padrão alto sugere um maior número de repetições, e um desvio muito baixo possibilita diminuir esse número).

Estatísticas importantes

Estas estatísticas devem ser coletadas para todas as gerações.

1. Fitness do melhor e pior indivíduos

2. Fitness média da população

3. Número de indivíduos repetidos na população

4. Número de indivíduos gerados por cruzamento melhores e piores que a fitness média dos pais

O que deve ser entregue...

- Código fonte do programa
- Documentação do trabalho:
  - Introdução
  - Implementação: descrição sobre a implementação do programa, incluindo detalhes da representação, fitness e operadores utilizados
  - Experimentos: Análise do impacto dos parâmetros no resultado obtido pelo AE.
  - Conclusões
  - Bibliografia

A entrega DEVE ser feita pelo Moodle na forma de um único arquivo zipado, contendo o código e a documentação do trabalho.

Considerações Finais

- Os parâmetros listados para execução dos experimentos são sugestões iniciais, e podem ser modificados a sua conveniência.
- Depois da entrega do trabalho, faremos uma competição em sala de aula para avaliar as diversas decisões de implementação do algoritmo e como a otimização dos parâmetros podem levar ao sucesso ou fracasso do algoritmo.

Tabela 1: Bases de dados disponibilizadas.

Base	**# atributos**	**# instâncias**		**Tipo**
		Treino	**Teste**	
synth1	3	60	600	Sintética
synth2	3	300	1000	Sintética
concrete[1]	9	824	206	Real

[1] Encontrada em `http://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength`.

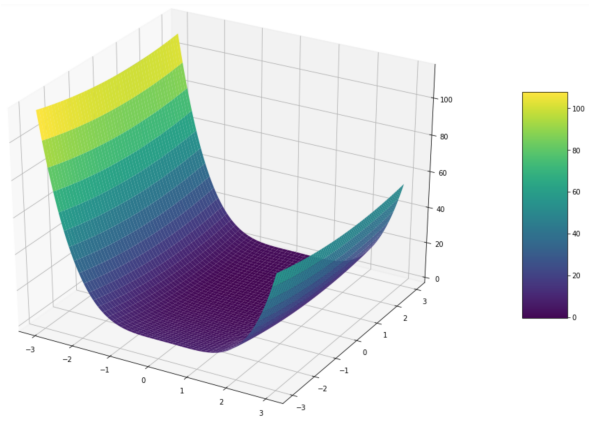

Figura 1: Base de dados synth1.

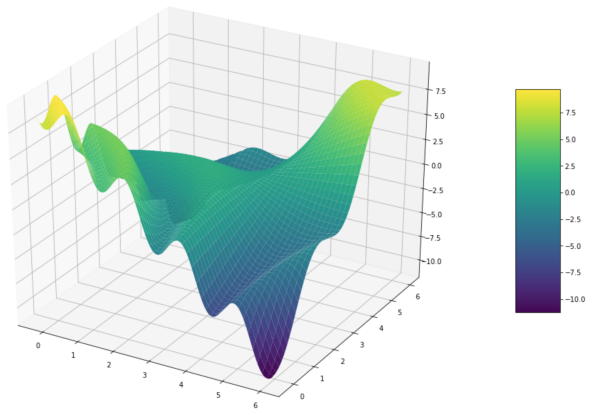

Figura 2: Base de dados synth2.