

Trabalho Prático 1
Programação Genética

Data de Entrega: 25 de setembro de 2017

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.

Uma 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 (RMSE)

$$f(Ind) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\text{EVAL}(Ind, x) - y)^2}$$

onde Ind é o indivíduo sendo avaliado, $\text{EVAL}(Ind, x)$ avalia o indivíduo Ind no conjunto de entrada fornecido x , y é a saída correta da função para a entrada x , e N é o número de exemplos fornecidos.

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 Bases de Dados

Três conjuntos de dados serão utilizados neste trabalho, e estão todos disponíveis no Moodle. Cada base de dados possui dois arquivos:

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 afim de detectar algum tipo de anomalia (overfitting, por exemplo).

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

3 Metodologia Experimental

O GP deve ser testado nas 3 bases de dados descritas na Tabela 1. A avaliação experimental descrita abaixo deve ser feita para uma das bases representando problemas sintéticos e para o problema real. Os parâmetros considerados mais apropriados para o base sintética escolhida devem ser novamente utilizados para a outra bases sintética (correspondente aos problemas listados na Tabela 1).

A parte de escolha e estudo dos parâmetros deve ser feita da seguinte forma:

- 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? Vocês podem 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 crossover e mutação. Na primeira, a probabilidade de crossover (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 crossover 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.
- Utilizar elitismo.
- 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.

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).

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. Escolha os melhores parâmetros dos anteriores e retire o elitismo. Os resultados obtidos são os mesmos?
5. Se desejar, teste outras características, como métodos para garantir a diversidade da população.

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 crossover 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.

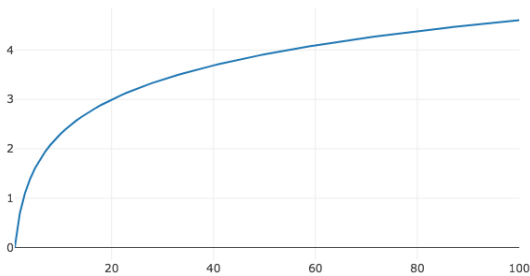
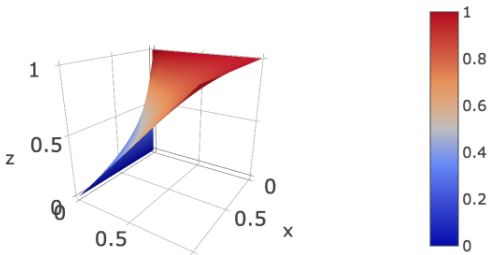
Considerações Finais

- 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.

Referências

- [1] harlfoxem. *House Sales in King County, USA*. <https://www.kaggle.com/harlfoxem/housesalesprediction>. Accessed: 2017-09-01.
- [2] James McDermott et al. “Genetic programming needs better benchmarks”. Em: *Proceedings of the 14th annual conference on Genetic and evolutionary computation*. ACM. 2012, pp. 791–798.

Tabela 1: Bases de Dados

Nome	Tipo	# de Instâncias		Atributos
		Treino	Teste	
Keijzer-7 [2]	Sintética	100	100	x, y
				
Keijzer-10 [2]	Sintética	100	100	x_0, x_1, y
				
House Sales Prediction (Processada) [1]	Real	5000	15000	x_0 : número de banheiros, x_1 : número de quartos, x_2 : área construída, x_3 : área do terreno, x_4 : condição, x_5 : avaliação, x_6 : área acima do solo, x_7 : área abaixo do solo, y : preço
Função Desconhecida				