

11/08/2021

## Segundo Trabalho de Inteligência Artificial e Sistemas Inteligentes

Prof. Flávio Miguel Varejão

### 1. Descrição

Este trabalho consiste em realizar uma comparação experimental entre um conjunto pré-definido de técnicas de aprendizado para classificação automática, baseadas na ideia de do uso de metaheurísticas para construção de combinados de classificadores, aplicadas a alguns problemas de classificação. As metaheurísticas escolhidas são: Hill Climbing, Simulated Annealing e Genetic Algorithm. As bases de dados a serem utilizadas são digits, wine e breast cancer, todas disponíveis em [https://scikit-learn.org/stable/datasets/toy\\_dataset.html](https://scikit-learn.org/stable/datasets/toy_dataset.html).

Para cada base, o procedimento experimental de treinamento, validação e teste será realizado através de 3 rodadas de ciclos aninhados de validação e teste, com o ciclo interno de validação contendo 4 folds e o externo de teste com 10 folds. A busca em grade (grid search) do ciclo interno deve considerar os valores 3, 5 e 7 do hiperparâmetro n\_samples para cada técnica de aprendizado.

Todos métodos devem ser implementados como variações do método Heterogeneous Pooling desenvolvido no primeiro trabalho. **Duas diferenças deverão ser realizadas. A primeira diferença é que não mais se treinarão classificadores com a base de treino original. Desde a primeira iteração deve ser feito o resampling da base de treino original para gerar os classificadores. A segunda diferença entre esses novos classificadores e o do primeiro trabalho será a seleção de um subconjunto dos classificadores gerados usando as metaheurísticas para formarem o combinado. A ideia é escolher a combinação mais promissora dentre os classificadores gerados.**

Para escolher o melhor subconjunto de classificadores, inicialmente se deve gerar o pool de classificadores da mesma forma que foi feito no classificador Heterogeneous Pooling. Depois deve-se proceder a seleção com a aplicação da metaheurística. Para isso, o conjunto de treinamento original terá que ser usado para estimar a acurácia do subconjunto de classificadores do combinado. O conjunto de classificadores selecionado será o que obtiver a melhor acurácia ao longo do processo de busca.

Os resultados de cada classificador devem ser apresentados numa tabela contendo a média das acurácias obtidas em cada fold do ciclo externo, o desvio padrão e o intervalo de confiança a 95% de significância dos resultados, e também através do boxplot dos resultados de cada classificador em cada fold.

Um exemplo de uma tabela para uma base de dados hipotética é mostrado a seguir.

Método	Média	Desvio Padrão	Limite Inferior	Limite Superior
Hill Climbing	0.95	0.04	0.94	0.96
Simulated Annealing	0.92	0.01	0.91	0.93

Genetic	0.96	0.08	0.90	0.99
---------	------	------	------	------

Os dados utilizados no conjunto de treino em cada rodada de teste, tal como no primeiro trabalho, devem ser padronizados (normalização com z-score). Os valores de padronização obtidos nos dados de treino devem ser utilizados para padronizar os dados do respectivo conjunto de teste.

Além das tabelas e dos gráficos bloxplot, será necessário apresentar também a tabela pareada dos resultados (p-values) dos testes de hipótese entre os pares de métodos. Na matriz triangular superior devem ser apresentados os resultados do teste t pareado (amostras dependentes) e na matriz triangular inferior devem ser apresentados os resultado do teste não paramétrico de wilcoxon. Os valores da célula da tabela que rejeitarem a hipótese nula para um nível de significância de 95% devem ser escritos em negrito.

Um exemplo de uma tabela pareada para uma base de dados hipotética é mostrado a seguir.

Hill Climbing	0.0941	<b>0.045</b>
<b>0.032</b>	Simulated Annealing	0.074
0.084	0.096	Genetic
0.105	0.084	<b>0.049</b>

O classificador Hill Climbing usará um algoritmo determinístico, apresentado na próxima seção. Os classificadores Simulated Annealing e Genetic usarão suas versões canônicas, ambas não determinísticas. Embora os valores dos hiperparâmetros dessas metaheurísticas deveriam ser determinados através de um processo de busca, para simplificação dos experimentos, você poderá definir valores fixos para esses hiperparâmetros. Você também terá liberdade para escolher e definir as funções usadas pelos algoritmos, por exemplo, a função de vizinhança do Simulated Annealing e as funções de recombinação e mutação do algoritmo genético.

Como a execução das metaheurísticas é computacionalmente pesada e, por conseguinte, demorada, neste trabalho cada execução de metaheurística deve ser limitada a, no máximo, 120 segundos.

O procedimento para classificar uma instância é idêntico ao usado pelo classificador Heterogeneous Pooling. O critério de decisão é votação majoritária, ou seja, deve-se escolher a classe mais escolhida dentre os classificadores que compõem o combinado. Em caso de empate, a classe escolhida deve ser a mais frequente na base de dados de treino original dentre as que empataram na votação.

## 2. Hill Climbing

O classificador Hill Climbing usa o algoritmo de seleção sequencial para frente para escolher o subconjunto de classificadores que comporão o combinado. A ideia consiste em

inicialmente avaliar os combinados compostos por um único classificador. Aquele que obtiver o melhor desempenho é escolhido para compor o combinado resultante. Na próxima iteração se avalia os combinados com dois classificadores, sendo um deles o classificador escolhido na primeira iteração. O combinado com um par de classificadores que obteve o melhor desempenho é selecionado. Caso seu desempenho seja superior **ou igual** ao combinado obtido na iteração anterior, esse combinado é selecionado para ser o resultante e o processo continua de forma análoga. Caso seu desempenho seja inferior ao combinado escolhido na iteração anterior, a busca é interrompida e o combinado escolhido na iteração anterior é retornado como resultado.

### **3. Informações Complementares**

a. Use o valor 36851234 para o parâmetro `random_state` (`random_state=36851234`) nas chamadas a `RepeatedStratifiedKFold`.

b. Use o valor 0 para o parâmetro `random_state` (`random_state=0`) na primeira chamada a `resample`. A partir daí, use o valor corrente incrementado de 1 nas chamadas sucessivas de `resample` (`random_state = 1`, `random_state = 2`, ...), tal como foi feito no primeiro trabalho.

### **4. Artigo**

Após a realização dos experimentos, um artigo descrevendo todo o processo experimental realizado deverá ser escrito em latex usando o software overleaf. O artigo deve ser estruturado contendo os seguintes componentes:

1. Título
2. Resumo
3. Seção 1. Introdução
4. Seção 2. Descrição dos Métodos Implementados (Hill Climbing, Simulated Annealing, Genetic)
5. Seção 3. Descrição dos Experimentos Realizados
  - a. Seção 3.1 Digits
  - b. Seção 3.2 Wine
  - c. Seção 3.3 Breast Cancer
6. Seção 4. Conclusões
  - a. Análise geral dos resultados
  - b. Contribuições do Trabalho
  - c. Melhorias e trabalhos futuros
7. Referências Bibliográficas

A título de comparação dos resultados, devem ser incluídas nas tabelas de média e desvio padrão, os resultados obtidos alcançados pelo método `HeterogeneousPooling` e o que

obteve o melhor resultado em cada base no primeiro trabalho. Isso não necessita ser feito para a tabela com os testes estatísticos.

Na subseção de análise geral dos resultados é importante discutir, dentre outras coisas, se houve diferença estatística significativa entre quais métodos em quais bases e responder se teve um método que foi superior em mais de uma base. Também deve ser feita a análise da comparação com o método Heterogeneous Pooling e o que obteve o melhor resultado em cada base no primeiro trabalho.

## **5. Condições de Entrega**

O trabalho deve ser feito individualmente e submetido pelo sistema da sala virtual até a data limite (08 de setembro de 2021).

O trabalho deve ser submetido em três arquivos: um arquivo pdf com o artigo produzido no trabalho, um arquivo ipynb com o notebook jupyter para ser carregado e executado no jupyter e todos os arquivos com código fonte em python utilizados em um arquivo zip. Tanto o arquivo pdf quanto os arquivos ipynb e zip devem possuir o mesmo nome Trab1\_Nome\_Sobrenome.

Note que a data limite já leva em conta um dia adicional de tolerância para o caso de problemas de submissão via rede. Isso significa que o aluno deve submeter seu trabalho até no máximo um dia antes da data limite. Se o aluno resolver submeter o trabalho na data limite, estará fazendo isso assumindo o risco do trabalho ser cadastrado no sistema após o prazo. Em caso de recebimento do trabalho após a data limite, o trabalho não será avaliado e a nota será ZERO. Plágio ou cópia de trabalhos serão verificadas automaticamente por sistemas como o moss. Trabalhos em que se configure cópia receberão nota zero independente de quem fez ou quem copiou.

## **6. Requisitos da implementação**

- a. Modularize seu código adequadamente.
- b. Crie códigos claros e organizados. Utilize um estilo de programação consistente, Comente seu código.
- c. Os arquivos do programa devem ser lidos e gerados na mesma pasta onde se encontram os arquivos fonte do seu programa.

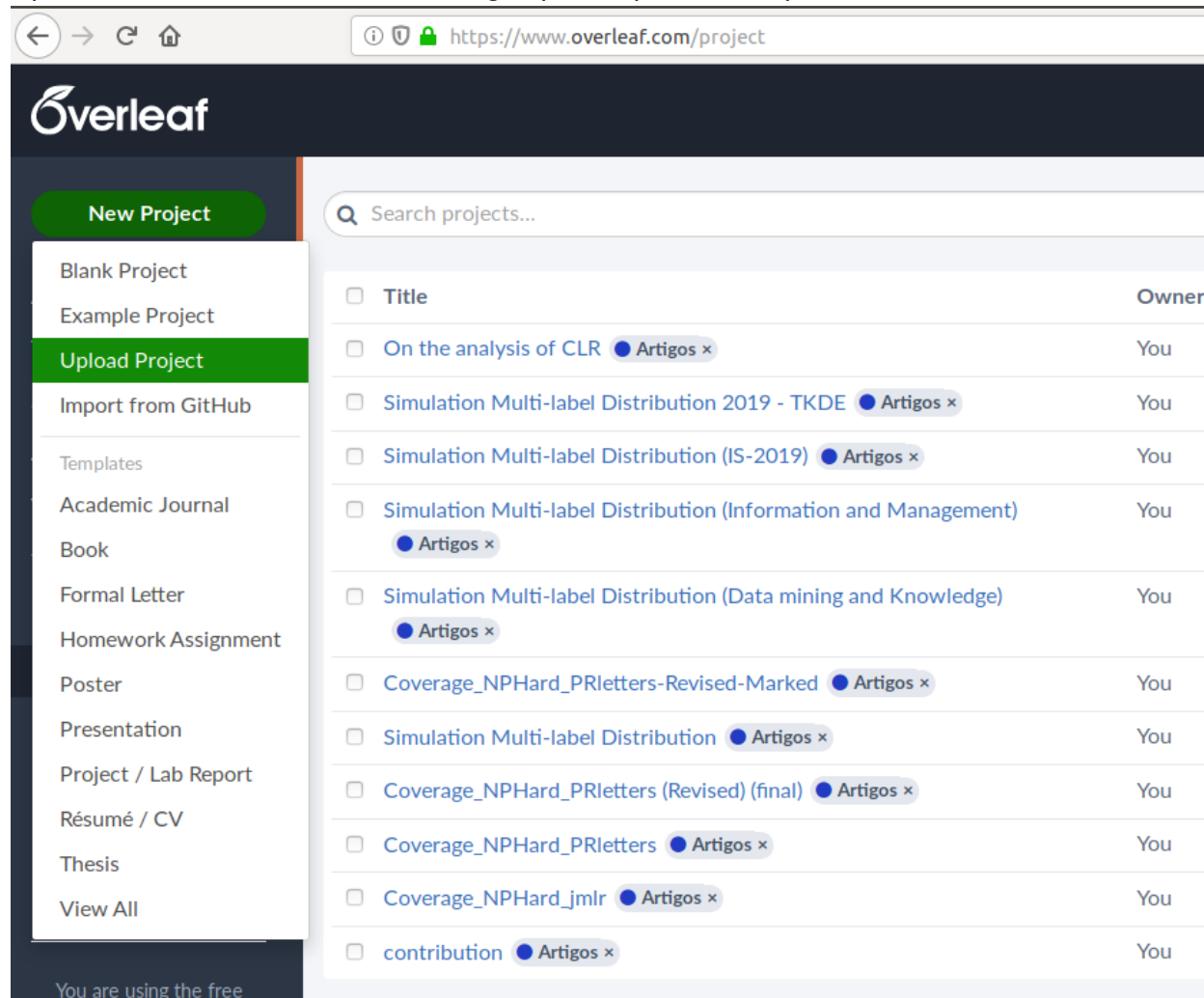
### **Observação importante**

**Caso haja algum erro neste documento, serão publicadas novas versões e divulgadas erratas em sala de aula. É responsabilidade do aluno manter-se informado, freqüentando as aulas ou acompanhando as novidades na página da disciplina na Internet.**

### **Apêndice. Artigo em Latex usando Overleaf**

Juntamente com este enunciado foi disponibilizado um arquivo zip com o template de latex

para confecção do artigo. O primeiro passo a ser feito é criar uma conta pessoal no Overleaf (<https://www.overleaf.com/register>). Uma vez criada sua conta, deve-se entrar nela. Para incluir o template no overleaf, basta apenas selecionar "New Project>Upload Project" e selecionar o arquivo zip, como mostrado na figura abaixo. Não é necessário descompactar, faça o upload do zip direto. Lembrar de renomear o artigo após o upload do arquivo.



The screenshot shows the Overleaf website interface. The browser address bar displays <https://www.overleaf.com/project>. The Overleaf logo is in the top left. A sidebar on the left contains a 'New Project' button, which is open, showing a dropdown menu with options: 'Blank Project', 'Example Project', 'Upload Project' (highlighted in green), 'Import from GitHub', and a 'Templates' section with 'Academic Journal', 'Book', 'Formal Letter', 'Homework Assignment', 'Poster', 'Presentation', 'Project / Lab Report', 'Résumé / CV', 'Thesis', and 'View All'. The main content area has a search bar 'Search projects...' and a table of projects.

<input type="checkbox"/>	Title	Owner
<input type="checkbox"/>	On the analysis of CLR <span>● Artigos x</span>	You
<input type="checkbox"/>	Simulation Multi-label Distribution 2019 - TKDE <span>● Artigos x</span>	You
<input type="checkbox"/>	Simulation Multi-label Distribution (IS-2019) <span>● Artigos x</span>	You
<input type="checkbox"/>	Simulation Multi-label Distribution (Information and Management) <span>● Artigos x</span>	You
<input type="checkbox"/>	Simulation Multi-label Distribution (Data mining and Knowledge) <span>● Artigos x</span>	You
<input type="checkbox"/>	Coverage_NPHard_PRletters-Revised-Marked <span>● Artigos x</span>	You
<input type="checkbox"/>	Simulation Multi-label Distribution <span>● Artigos x</span>	You
<input type="checkbox"/>	Coverage_NPHard_PRletters (Revised) (final) <span>● Artigos x</span>	You
<input type="checkbox"/>	Coverage_NPHard_PRletters <span>● Artigos x</span>	You
<input type="checkbox"/>	Coverage_NPHard_jmlr <span>● Artigos x</span>	You
<input type="checkbox"/>	contribution <span>● Artigos x</span>	You

You are using the free