IMD1101 - Aprendizado de Máquina 2020.2

Check-point Final

Alunas: Ana Caroline da Silva Dantas
Luana Brenda Pontes Ferreira
Tayane da Costa Varela

1. Objetivos

O principal objetivo desta parte do trabalho prático é analisar como os métodos e técnicas supervisionados vistos em sala de aula se comportam quando combinados em comitês de classificadores, em uma aplicação prática. Para tal, o aluno terá que responder às seguintes perguntas de pesquisa:

1. Qual o impacto do número de classificadores base em comitês homogêneos?
2. Qual o impacto do número de classificadores base em comitês heterogêneos?
3. Qual o impacto do tipo de classificadores base em comitês homogêneos?
4. Qual o impacto do tipo de classificadores base em comitês heterogêneos?
5. Qual a melhor escolha de tipo de comitê, homogêneo ou heterogêneo?

2. Metodologia

Inicialmente, utilize a base de dados que você vem utilizando no decorrer da disciplina, contendo o atributo classe. Aplique as estratégias de aprendizado Boosting e Bagging, vistos em sala de aula, para esta base de dados, contendo 10, 15 e 20 classificadores base. Para tal, utilize todos os 4 algoritmos vistos em sala de aula: AD, NB (RF para quem trabalhou com regressão), k-NN e MLP. Então monte duas tabelas, uma para Bagging e uma para Boosting, no seguinte formato:

Tabela 1 - Estratégia de Aprendizado: Boosting.

Estratégia	10	15	20	Média (Class)
AD	0.933333 ± 0.010	0.93483 ± 0.007	0.937211 ± 0.009	0.935125 ± 0.009
k-NN	0.899409 ± 0.007	0.899409 ± 0.007	0.899409 ± 0,007	0.899409 ± 0.007
NB	0.520884 ± 0.108	0.535442 ± 0.062	0.565102 ± 0.059	0.540476 ± 0.076
MLP	0.937007 ± 0.005	0.940816 ± 0.009	0.938095 ± 0.006	0.938639 ± 0.007
RF	Não se aplica	Não se aplica	Não se aplica	Não se aplica
Média (TAM)	0.822658 ± 0.033	0.827624 ± 0,021	0.834954 ± 0.020	

Prec = Precisão e DP = Desvio Padrão

Tabela 2 - Estratégia de Aprendizado: Bagging.

Estratégia	10	15	20	Média (Class)
AD	0.900680 ± 0.011	0.902245 ± 0.006	0.899524 ± 0.007	0.900816 ± 0.008
k-NN	0.909320 ± 0.004	0.9136735 ± 0.006	0.914355 ± 0.007	0.912449 ± 0.006
NB	0.820000 ± 0.006	0.821020 ± 0.013	0.825306 ± 0.011	0.822109 ± 0.010
MLP	0.929728 ± 0.006	0.931020 ± 0.004	0.933401 ± 0.006	0.931383 ± 0.005
RF	Não se aplica	Não se aplica	Não se aplica	Não se aplica
Média (TAM)	0.889932 ± 0.007	0.89199 ± 0.007	0.893146 ± 0.008	

Prec = Precisão e DP = Desvio Padrão

Nesta tabela, a última linha representa a média de precisão e desvio padrão para um determinado tamanho, assim como a última coluna representa a média de precisão e desvio padrão para um algoritmo de classificação. Ao analisar as duas tabelas, responda às seguintes perguntas:

Tabela 3 - Valores individuais para os algoritmos.

	AD	k-NN	NB	NN	RF
Base Reduzida 1	0.897959	0.91449	0.836735	0.928571	Não se aplica

Tabela 4 - Média (Class) para os algoritmos pelo método Boosting.

	AD	k-NN	NB	MLP	RF
Média (Class)	0.935125 ± 0.009 [0.926125; 0.944125]	0.899409 ± 0.007 [0.892409; 0.906409]	0.540476 ± 0.076 [0.464476; 0.616476]	0.938639 ± 0.007 [0.931639; 0.945639]	Não se aplica

Tabela 5 - Média (Class) para os algoritmos pelo método Bagging.

	AD	k-NN	NB	MLP	RF
Média (Class)	0.900816 ± 0.008 [0.892816; 0.908816]	0.912449 ± 0.006 [0.906449; 0.918449]	0.822109 ± 0.010 [0.812109; 0.832109]	0.931383 ± 0.005 [0.926383; 0.936383]	Não se aplica

2.1. Os resultados foram melhores que os individuais?

*O **AD** e o **MLP** apresentaram acurácia melhor pelo método Boosting. Já o **NB** teve melhor resultado no valor individual. E o **k-NN** apresentou semelhança na acurácia para o valor individual e método Bagging.*

2.2. Qual algoritmo mais melhorou e menos melhorou com o uso de comitê? Para essa comparação, utilize apenas a análise visual, comparando os resultados dos algoritmos individuais com os fornecidos pelo Bagging e Boosting.

O algoritmo de Árvore de Decisão obteve uma melhora com o uso do comitê, enquanto que o Naive Bayes obteve um menor desempenho entre os algoritmos.

2.3. Existiu um padrão de comportamento entre os algoritmos de classificação?

Aparentemente, usar os algoritmos de classificação pode melhorar em alguns modelos, como no modelo NB, que manteve a acurácia pelo método bagging e piorou drasticamente pelo método boosting. Ao que tudo indica, os algoritmos parecem não ter um padrão de comportamento.

Com os resultados da última linha, é possível responder a pergunta (1), pois esta linha representa o melhor desempenho da estratégia com um tamanho específico, independente do tipo de classificador utilizado. Faça esta análise para as duas estratégias e responda às seguintes perguntas:

2.4. Qual foi o tamanho com melhor desempenho para o Bagging?

Os resultados foram praticamente iguais para todos os tamanhos, em que para os tamanhos 15 e 20 a alteração foi apenas após a terceira casa decimal. Considerando essa pequena diferença, o melhor tamanho foi 20.

2.5. Qual foi o tamanho com melhor desempenho para o Boosting?

O tamanho com melhor desempenho foi o 20 com uma média de 0.8349542238.

2.6. Foi possível detectar algum padrão de comportamento para estas duas estratégias de aprendizado?

Para as duas estratégias utilizadas, boosting e bagging, é possível observar que a média de precisão para os tamanhos 10, 15 e 20 são bem próximas em cada estratégia. Como foi falado acima, para o Bagging a média só mudou após a terceira casa decimal. E para o Boosting, essas médias de precisão também estão bem próximas.

Com os resultados da última coluna, é possível responder a pergunta (3), pois esta coluna representa o melhor desempenho do tipo de classificador específico, independente do tamanho do comitê utilizado. Faça esta análise para as duas estratégias e responda às seguintes perguntas:

2.7. Qual foi o tamanho com melhor desempenho para o Bagging?

O MLP obteve melhor desempenho pelo método Bagging, com tamanho 20.

2.8. Qual foi o tamanho com melhor desempenho para o Boosting?

O MLP obteve melhor desempenho pelo método Boosting, com tamanho 15.

2.9. Foi possível detectar algum padrão de comportamento para estas duas estratégias de aprendizado?

*O método bagging se saiu bem para os algoritmos KNN e NB. No geral, se mostrou um método com bom desempenho. Já para o método Boosting, observa-se uma acurácia boa para os algoritmos **AD** e MLP, porém, não apresentou um bom desempenho para o NB.*

Uma vez feitas as análises com o Bagging e o Boosting, é preciso responder as perguntas (2) e (4). Para tal, crie mais duas tabelas, uma para o Stacking Homogêneo e a outra para o Stacking

Heterogêneo. Para o Stacking homogêneo, as mesmas configurações do Bagging e Boosting são utilizadas (tamanhos 10, 15 e 20 e usando AD, k-NN, NB e MLP), usando um combinador que tenha bom desempenho para a sua base de dados (teste uns 2 ou 3 e veja qual seria o melhor). Com a tabela do Stacking Homogêneo, responda a seguinte pergunta:

Tabela 6 - Stacking Homogêneo.

Estratégia	10	15	20	Média (Class)
AD	0.894898 ± 0.010	0.896054 ± 0.007	0.893265 ± 0.012	0.894739 ± 0.010
k-NN	0.625510 ± 0.011	0.623537 ± 0.012	0.627755 ± 0.012	0.625601 ± 0.012
NB	0.824694 ± 0.009	0.825782 ± 0.010	0.815238 ± 0.011	0.821905 ± 0.010
MLP	0.541633 ± 0.012	0.542313 ± 0.011	0.541633 ± 0.013	0.541859 ± 0.012
RF	Não se aplica	Não se aplica	Não se aplica	Não se aplica
Média (TAM)	0.721684 ± 0.011	0.721922 ± 0.010	0.719473 ± 0.012	

Prec = Precisão e DP = Desvio Padrão

2.10. Usar o Stacking Homogêneo foi melhor que o Boosting ou o Bagging?

Não, os métodos Boosting e Bagging se saíram melhor de forma geral.

2.11. Foi possível detectar algum padrão de comportamento para estas três técnicas, Stacking, Bagging e Boosting?

*A partir das três técnicas, tivemos resultados melhores em alguns modelos, e muito ruins em outros. O **MLP** e **k-NN**, por exemplo, só tiveram boa precisão pelo método Boosting e Bagging.*

Para criar a tabela do Stacking heterogêneo, utilize os mesmos tamanhos das configurações homogêneas (10, 15 e 20 classificadores). Para os algoritmos de classificação, escolha três classificadores base, de acordo com o desempenho dos classificadores individuais. Por exemplo, se na sua análise comparativa, o melhor classificador foi MLP, seguido por: k-NN, AD e NB, então escolha o MLP (Método A), k-NN (Método B) e a AD (Método C), pois estes são os três classificadores com o primeiro, segundo e terceiro melhor desempenho. Então, utilize quatro configurações:

A. 50% do Método A e 50% do Método B

B. 50% do Método A e 50% do Método C

C. 50% do Método B e 50% do Método C

D. 33% do Método A, 33% do Método B e 33% do Método C

De acordo com os resultados individuais obtidos no Checkpoint 2:

- Método A: MLP = 0.9285714
- Método B: KNN = 0.9144898
- Método C: AD = 0.8979592

Tabela 7 - Stacking Heterogêneo.

Estratégia	10	15	20	Média (Class)
Configuração A	0.560476 ± 0.076	0.555714 ± 0.085	0.510068 ± 0.069	0.510068 ± 0.077
Configuração B	0.763197 ± 0.190	0.844694 ± 0.135	0.848571 ± 0.124	0.818821 ± 0.150
Configuração C	0.895374 ± 0.010	0.897143 ± 0.009	0.89619 ± 0.012	0.896236 ± 0.010
Configuração D	0.85483 ± 0.116	0.892109 ± 0.008	0.851633 ± 0.123	0.86619 ± 0.082
Média (TAM)	0.768469 ± 0.098	0.797415 ± 0.059	0.776616 ± 0.082	

Prec = Precisão e DP = Desvio Padrão

Daí, uma tabela similar a Tabela 1 pode ser criada para o Stacking heterogêneo. Com a última linha da tabela do Stacking heterogêneo, responda a pergunta (2), respondendo os mesmos questionamentos do Bagging (perguntas 2.4 a 2.6). Com os resultados da última coluna, é possível responder a pergunta (4), (itens 2.7 a 2.9).

Uma vez respondidas as perguntas (1), (2), (3) e (4), é preciso responder a pergunta (5). Para responder a esta pergunta, analise as tabelas dos Stackings Homogêneo e Heterogêneo e responda às seguintes perguntas:

2.12. Qual foi a estrutura que apresentou melhor precisão, homogêneo ou heterogêneo?

*Stacking heterogêneo. Por esse Stacking, o resultado mais baixo foi a configuração A. Já no Stacking homogêneo, o modelo **k-NN** e **MLP** tiveram precisão baixa.*

2.13. É possível detectar algum padrão de comportamento para as duas estruturas de stacking?

O método heterogêneo teve uma precisão boa na maioria dos casos, o homogêneo obteve resultados distintos entre os métodos.

3 . Testes Estatísticos

Avaliar todos os resultados obtidos para esse e os checkpoints antigos. Para cada parte os seguintes testes devem ser executados:

- Teste de Friedman: Para tal, teremos que usar todos os resultados de todos os métodos analisados. Por exemplo, para o checkpoint 2, podemos analisar o desempenho dos 4 métodos, supondo que cada método é definido como uma amostra. Desta forma, o teste irá definir se as amostras vêm da mesma população (não rejeitar a hipótese nula) ou de populações diferentes (rejeitar a hipótese nula).
- Teste de Wilcoxon: Para os casos onde as amostras são de populações diferentes, este teste pareado vai ser aplicado para decidir qual método é diferente do outro.

Para avaliar os checkpoints, primeiro é necessário estabelecer nossas hipóteses e nível de significância. Nossa hipótese nula (H_0) afirma que não há diferença no desempenho entre os métodos (populações iguais). Já a nossa hipótese alternativa (H_1) afirma que há diferença no desempenho entre os grupos (populações diferentes). Quanto ao nível de significância, estabelecemos o $\alpha = 0.05$.

Para o Checkpoint 2 (resultados supervisionados) temos os métodos Árvore de Decisão (AD), k-NN, Naive Bayes (NB) e MLP. Ao realizar o teste de Friedman, obteve-se uma estatística igual a 30 e um p-valor de 0.0000014, ou seja, a um nível de significância de 0.05 pode-se afirmar que existem evidências estatísticas suficientes para rejeitar a hipótese nula. Isto significa que existe diferença no desempenho entre os grupos.

Para analisar detalhadamente essa diferença entre os grupos, utilizamos o teste de Nemenyi. Os resultados são apresentados na Tabela 8.

Tabela 8 - Teste de *Nemenyi* para checkpoint 2.

Método	***p-valor***
AD e k-NN	*0.30713*
AD e NB	*0.30713*
AD e MLP	*0.00299*
k-NN e NB	*0.00299*

k-NN e MLP	*0.30713*
NB e MLP	*0.00100*

Observa-se que para os métodos AD e MLP, k-NN e NB e NB e MLP temos p-valor < 0.05, ou seja, para esses métodos dois a dois existe diferença no desempenho e foram estes métodos que influenciaram no resultado do p-valor para o teste de Friedman.

Para o Checkpoint 3 (resultados não supervisionados) temos os métodos K-means, Hierárquico Aglomerativo e Expectation Maximization. Ao realizar o teste de Friedman, obteve-se uma estatística de 22.211 e um p-valor igual a 0.00002, ou seja, a um nível de significância de 0.05 pode-se afirmar que existem evidências estatísticas suficientes para rejeitar a hipótese nula. Isto significa que existe diferença no desempenho entre os grupos.

Para analisar detalhadamente essa diferença entre os grupos, utilizamos o teste de Nemenyi, que analisa os métodos par a par. Os resultados podem ser observados na Tabela 9.

Tabela 9 - Teste de *Nemenyi* para checkpoint 3.

Método	***p-valor***
K-means e Hierárquico Aglomerativo	*0.025585*
K-means e Expectation Maximization	*0.001000*
Hierárquico Aglomerativo e Expectation Maximization	*0.087974*

Portanto, temos para os métodos K-means e Hierárquico Aglomerativo e K-means e Expectation Maximization p-valor < 0.05. Ou seja, para esses métodos dois a dois existe diferença no desempenho e foram estes que influenciaram no resultado do p-valor para o teste de Friedman. Já para o Hierárquico Aglomerativo e Expectation Maximization o p-valor é > 0.05, ou seja, não existe diferença de desempenho entre estes dois métodos.

Para o checkpoint 4 (abordagem de ensembles) os métodos utilizados, assim como, os testes de Friedman são apresentados na Tabela 10.

Tabela 10 - **Teste de Friedman para o checkpoint 4.**

Método	***p-valor***
Boosting class 10	*0.000649*
Boosting class 15	*0.002733*
Boosting class 20	*0.002482*
Bagging class 10	*0.00042*

Bagging class 15	*0.000616*
Bagging class 20	*0.000990*
Stacking homogêneo class 10	*0.000883*
Stacking homogêneo class 15	*0.000354*
Stacking homogêneo class 20	*0.000981*
Stacking heterogêneo class 10	*0.03971*
Stacking heterogêneo class 15	*0.0987*
Stacking heterogêneo class 20	*0.09845*

Como podemos observar nos valores destacados (vermelho), com p-valor > 0.05, os métodos Stacking heterogêneo, com classificação 15 e 20, não rejeitam a hipótese nula, ou seja, apresentam desempenhos iguais (população igual). Já os outros métodos tiveram desempenhos diferentes, para verificar quais configurações são diferentes entre si, usaremos o teste de Nemenyi.

Nas Tabelas 11 a 20, apresentadas abaixo, são mostradas para cada método o p-valor encontrado utilizando o teste de Nemenyi para os modelos par a par. Caso o p-valor seja menor que 0.05, afirmamos que esses dois métodos apresentam desempenhos diferentes e que foram estes métodos que influenciaram no resultado do p-valor para o teste de Friedman. Caso obtenha-se um valor > 0.05, afirmamos que os dois métodos apresentam desempenhos iguais. A codificação utilizada nas tabelas a seguir é a seguinte: ***0 - AD, 1 - KNN, 2 - NB e 3 - MLP.***

Tabela 11 - Teste de Nemenyi para Boosting class 10

	0	1	2	3
0	1.000000	0.109694	0.00100	0.896267
1	0.109694	1.000000	0.30713	0.017062
2	0.001000	0.307130	1.00000	0.001000
3	0.896267	0.017062	0.00100	1.000000

Tabela 12 - Teste de Nemenyi para Boosting class 15

	0	1	2	3
0	1.000000	0.109694	0.00100	0.896267
1	0.109694	1.000000	0.30713	0.017062
2	0.001000	0.307130	1.00000	0.001000
3	0.896267	0.017062	0.00100	1.000000

Tabela 13 - Teste de Nemenyi para Boosting class 20

	0	1	2	3
0	1.000000	0.013021	0.00100	0.799047
1	0.013021	1.000000	0.30713	0.132924
2	0.001000	0.307130	1.00000	0.001000
3	0.799047	0.132924	0.00100	1.000000

Tabela 14 - Teste de Nemenyi para Bagging class 10

	0	1	2	3
0	1.000000	0.009860	0.00100	0.701825
1	0.009860	1.000000	0.30713	0.160247
2	0.001000	0.307130	1.00000	0.001000
3	0.701825	0.160247	0.00100	1.000000

Tabela 15 - Teste de Nemenyi para Bagging class 15

	0	1	2	3
0	1.000000	0.017062	0.00100	0.896267
1	0.017062	1.000000	0.30713	0.109694
2	0.001000	0.307130	1.00000	0.001000
3	0.896267	0.109694	0.00100	1.000000

Tabela 16 - Teste de Nemenyi para Bagging class 20

	0	1	2	3
0	1.000000	0.005517	0.001000	0.225871
1	0.005517	1.000000	0.225871	0.507386
2	0.001000	0.225871	1.000000	0.005517
3	0.225871	0.507386	0.005517	1.000000

Tabela 17 - Teste de Nemenyi para Stacking Homogêneo class 10

	0	1	2	3
0	1.00000	0.00299	0.30713	0.00100
1	0.00299	1.00000	0.30713	0.30713
2	0.30713	0.30713	1.00000	0.00299
3	0.00100	0.30713	0.00299	1.00000

Tabela 18 - Teste de Nemenyi para Stacking Homogêneo class 15

	0	1	2	3
0	1.00000	0.00299	0.30713	0.00100
1	0.00299	1.00000	0.30713	0.30713
2	0.30713	0.30713	1.00000	0.00299
3	0.00100	0.30713	0.00299	1.00000

Tabela 19 - Teste de Nemenyi para Stacking Homogêneo class 20

	0	1	2	3
0	1.00000	0.00299	0.30713	0.00100
1	0.00299	1.00000	0.30713	0.30713
2	0.30713	0.30713	1.00000	0.00299
3	0.00100	0.30713	0.00299	1.00000

Tabela 20 - Teste de Nemenyi para Stacking Heterogêneo class 10

	0	1	2	3
0	1.000000	0.046280	0.005517	0.072567
1	0.046280	1.000000	0.896267	0.900000
2	0.005517	0.896267	1.000000	0.799047
3	0.072567	0.900000	0.799047	1.000000