INTRODUÇÃO

Para este laboratório foi utilizado cinco classificadores lineares diferentes:

KNN *(k-nearest neighbors):* O KNN trata-se de um classificador que analisa os vizinhos mais próximos, para classificar a qual grupo o elemento analisado pertence.

LDA *(Linear Discriminant analysis):* Método que utiliza as características mais discriminantes de uma classe para fazer a a classificação, ele faz a maximização da distância entre as classes e minimização da distância inter-classes.

Naive-Bayes: Baseado no teorema de Bayes utiliza como base para sua classificação a estatística.

Perceptron: Esse método foi inspirado pelo funcionamento de um neurônio humano, foi criado uma fórmula matemática que tenta imitar o funcionamento de um neurônio.

Regressão logística: A Regressão logística analisa o comportamento de algumas variáveis, com base nessa analise classifica as demais, isso é feito através da utilização de uma função matemática aplicada sobre a regressão linear.

OBJETIVOS

Conforme relatado no moodle o objetivo desse laboratório é analisar os classificadores citados na introdução, KNN, LDA, Naive-Bayes, Peceptron e Regressão logística. Consideramos a base de treinamento 20000 exemplos e teste (58646 exemplos), as quais contem 10 classes balanceadas e 132 características, exatamente conforme fornecido.

O objetivo do laboratório é analisar a base de dados fornecida, comparando os desempenhos como tempo de execução, acurácia para identificar assim o mais rápido e também o que tem melhor desempenho, além de verificar ainda qual o tamanho que a base de treinamento deixa de ser relevante, e por fim analisar as matrizes de confusão para verificar os erros de cada classificador.

METODOLOGIA

Os materiais utilizados foram o Google colab, scikit learn e o libre office calc, com base nas aulas e laboratório anterior foi criado um algoritmo. Com isso foi possível efetuar s análises conforme orientado no documento – impactos da base de aprendizagem.

Conforme solicitado no documento o algoritmo alimentou, de 1000 em 1000 amostras até um total de 20000, utilizando a base de treinamento. Para treino e classificação foi adotado a forma *default* de cada classificador da biblioteca scikit learn.

O algoritmo me retorna as matrizes confusão para cada iteração, portando 100 matrizes e os resultados da acurácia, f1store, tempo de treinamento para cada quantidade de amostras, todos os dados foram colocados em uma planilha do libre office para edição das conclusões.

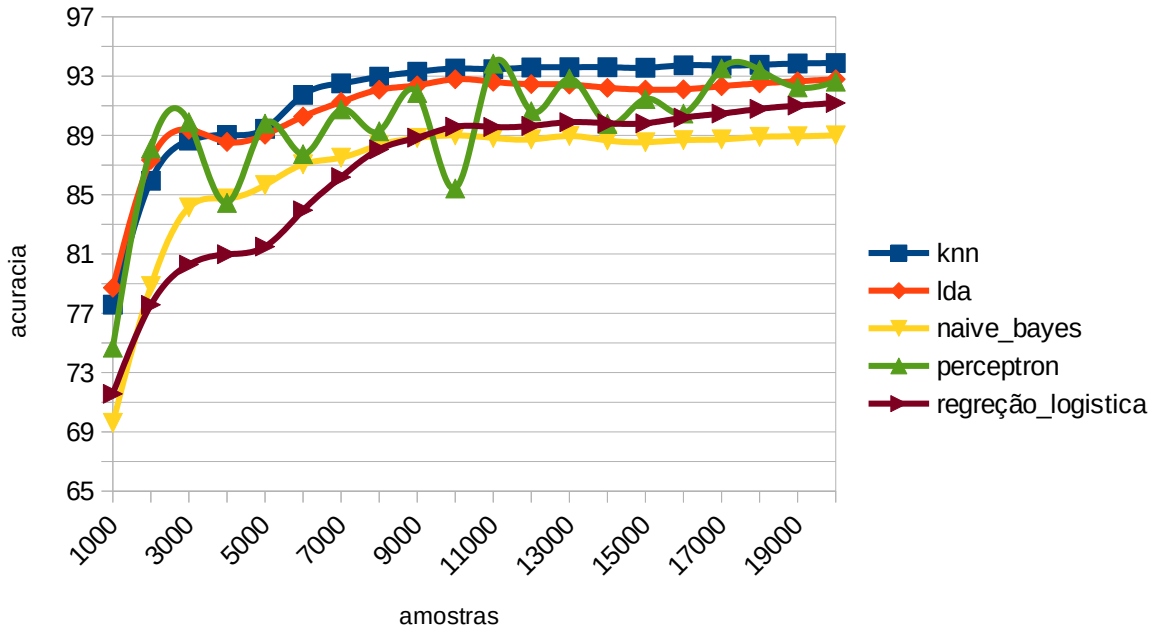
RESULTADOS E DISCUSSÃO

Caso 1: Compare o desempenho desses classificadores em função da disponibilidade de base de treinamento. Alimente os classificadores com blocos de 1000 exemplos e plote num gráfico o desempenho na base de testes. Analise em qual ponto o tamanho da base de treinamento deixa de ser relevante.

Conforme orientado foi incrementado de 1000 em 1000 e gerado o gráfico para fazer a comparação da acurácia em cada passo.

Analisando o gráfico abaixo, podemos confirmar que a partir de 10000 amostras todos os classificadores com exceção ao perceptron, a acurácia tende a se tornar linear, a melhora da acurácia aumenta com pouca relevância.

Gráfico 1:



Caso 2: Indique qual é o classificador que tem o melhor desempenho com poucos dados < 1000 exemplos.

Conforme tabela abaixo o classificador com melhor desempenho abaixo de 1000 amostras trata-se do LDA, com acurácia superior a 78%, podemos observar que o tempo de treinamento do KNN já está elevado, o que forçou a executar em três partes devido queda do google colab, até 12000 amostras na primeira, até 16000 na segunda e por fim até 20000.

Tabela 1: (1000 amostras)

Classificador	Tempo de Treinamento	Tempo de Testes	Acurácia
knn	17,4286	0,0104	0,7758
LDA	0,0732	0,0339	0,7873
naive_bayes	0,533	0,0042	0,6961
perceptron	0,0358	0,0423	0,7469
regressao_logistica	0,0357	0,1381	0,7157

Caso 3: Indique o classificador que tem melhor desempenho com todos os dados.

O KNN trata-se do classificador com melhor desempenho porém com maior tempo de treinamento.

Tabela 2:

Classificador	Amostras	Tempo de Treinamento	Tempo de Testes	Acurácia
knn	20000	177,7656	0,3994	0,9388
lda	20000	0,0687	0,525	0,9279
naive_bayes	20000	0,5202	0,0424	0,89
perceptron	20000	0,0378	0,8588	0,9263
regressao_logistica	20000	0,0502	5,9194	0,9119

A Tabela abaixo mostra ainda os 10 piores tempo de teste:

Tabela 3:

Classificador	Amostras	Tempo de Treinamento	Tempo de Testes	Acuracia	F1Score
regressao_logistica	11000	0,0363	2,9044	0,8957	0,8957
regressao_logistica	12000	0,0351	2,9354	0,8964	0,8964
regressao_logistica	13000	0,0343	3,139	0,8989	0,8989
regressao_logistica	15000	0,0366	3,673	0,8982	0,8982
regressao_logistica	14000	0,0465	3,7078	0,8982	0,8982
regressao_logistica	16000	0,0349	4,4664	0,902	0,9021
regressao_logistica	17000	0,0391	4,6667	0,9047	0,9048
regressao_logistica	18000	0,0394	5,0241	0,9079	0,908
regressao_logistica	19000	0,0389	5,2109	0,9101	0,9101
regressao_logistica	20000	0,0502	5,9194	0,9119	0,9119

Caso 4: Indique o classificador mais rápido para classificar os 58k exemplos de teste.

Dos 10 tempos de classificadores mais rápidos 9 deles pertencem Naive-Bayes baseando-se na quantidade de amostras.

Tabela 4:

Classificador	**Amostras**	**Tempo de Testes**	**Acuracia**	**F1Score**
naive_bayes	2000	0,0037	0,7884	0,782
naive_bayes	1000	0,0042	0,6961	0,6744
naive_bayes	3000	0,0068	0,8415	0,8384
naive_bayes	4000	0,0098	0,8475	0,8449
knn	1000	0,0104	0,7758	0,782
naive_bayes	5000	0,0109	0,8565	0,8551
naive_bayes	7000	0,0128	0,875	0,8742
naive_bayes	6000	0,0131	0,8705	0,8694
naive_bayes	8000	0,016	0,8835	0,8825
naive_bayes	10000	0,0206	0,89	0,8891

Quando olhamos para as tabelas 1 e 2, podemos observar a classificação dos tempos de teste para cada classificador com a mesma quantidade de amostra para treinar. O Naive-Bayes se destaca em todos os casos.

Caso 5: Analise as matrizes de confusão. Os erros são os mesmos para todos os classificadores quando todos eles utilizam toda a base de teste?

No Knn os erros ocorrem principalmente com os números 5 e 3, com o LDA apesar de ter uma quantidade mais significativa com os números 5 e 3 acontece erros com significância em outros números também. O Naive-Bayes comete mais confusões entre 7 e 3, 1 e 7. O perceptron comete maiores confusões com os números 8 e 0, 8 e 3. A regressão logística comete erros com mais frequência nos númros 5 e 3, 1 e 3.

Podemos observar que o 1 pode ser confundido em todos os classificadores com praticamente todos os demais números, podemos ainda observar que existe grande semelhança entre todos os classificadores.

Matriz de confusão KNN:

5472	3	1	15	6	2	26	2	32	1
0	6105	175	119	56	6	35	66	34	59
12	11	5607	165	3	1	16	51	20	2
4	1	25	5646	2	51	1	53	20	16
12	11	13	3	5305	9	132	24	11	202
9	3	9	489	4	4842	41	16	83	43
31	10	4	2	3	44	5724	0	40	0
1	25	41	119	54	1	0	5773	7	76
36	24	42	114	32	38	50	27	5165	167
16	9	17	107	78	9	9	131	34	5403

Matriz de confusão LDA:

5358	10	11	15	19	0	47	17	80	3
0	6027	222	85	9	22	38	199	31	22
22	41	5605	12	1	0	4	175	27	1
1	12	29	5470	1	19	1	247	23	16
20	71	42	0	5208	0	86	5	29	261
9	11	6	314	4	5015	50	24	67	39

77	49	37	15	56	36	5460	0	125	3
0	58	47	6	58	1	0	5882	22	23
80	59	38	5	51	29	54	57	4961	361
34	31	9	91	69	7	16	98	29	5429

Matriz de confusão Naive-Bayes:

5220	1	11	32	2	1	41	0	251	1
1	5184	583	238	86	22	85	340	80	36
9	24	5289	447	4	1	8	52	53	1
2	1	212	5390	1	33	0	127	31	22
14	2	44	12	5273	0	32	44	90	211
9	6	29	103	31	4958	46	2	169	186
78	7	89	8	15	90	5286	0	285	0
1	47	175	426	21	1	1	5323	60	42
175	5	53	182	23	7	38	13	5112	87
25	5	62	151	221	4	0	55	184	5106

Matriz de confusão Perceptron:

5532	1	0	6	0	1	18	1	1	0
14	6114	46	217	14	176	27	43	2	2
88	32	5548	137	2	0	16	62	3	0
5	3	12	5698	0	60	1	28	2	10
116	13	46	17	5172	7	108	39	5	199
21	5	4	129	3	5318	40	1	6	12
129	8	5	4	5	57	5648	0	2	0
2	42	51	157	31	4	0	5796	1	13
329	39	45	457	35	225	185	20	4211	149
89	36	26	115	106	25	3	83	3	5327

Matriz de confusão Regressão Logística:

5381	5	16	12	15	4	69	6	51	1
1	5595	116	269	200	74	179	74	78	69
22	18	5585	89	12	1	33	82	45	1
4	3	37	5597	16	39	1	74	20	28

35	8	30	1	5315	2	104	41	9	177
6	12	23	497	78	4728	50	22	73	50
87	26	0	1	20	96	5517	0	111	0
0	41	40	121	165	2	0	5600	17	111
83	43	47	59	85	46	53	58	5000	221
55	22	8	143	251	0	4	150	19	5161

CONCLUSÃO

Foi possível notarmos grande semelhança entre os classificadores lineares, também foi possível ter uma visão geral do funcionamento de cada uma para conhecimento e aplicação.

Com a analise detalhada das matrizes de confusão foi possível ainda ter uma visão de onde acontecem os erros, além da analise do tempo de execução dos treinamentos e teste, o que nos permite dimensionar custos e visualizar o que seria mais vantajoso.