Relatório Impactos da Base de Aprendizagem

Aprendizado de Máquina - INFO7004

Prof. Luiz Eduardo S. Oliveira, Ph.D

Discente: Marc Queiroz

16 de agosto de 2020

1 Introdução

Este trabalho investiga os impactos da base de treinamento para 5 classificadores lineares:

- kNN
- Naive Bayes
- Linear Discriminant Analysis
- Logistic Regression
- Perceptron

Com uma base de treinamento de 20000 entradas e teste de 58646 entradas, separadas em 10 classes e 132 características.

Este trabalho tem como objetivo resolver 5 atividades propostas. As próximas seções vão apresentar as respostas para o trabalho proposto.

2 Atividade - I

Atividade: 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.

Após implementar os classificadores e executá-los com blocos de 1000 até 20000, com passos de 1000, chegou-se ao gráfico de desempenho da figura 1.

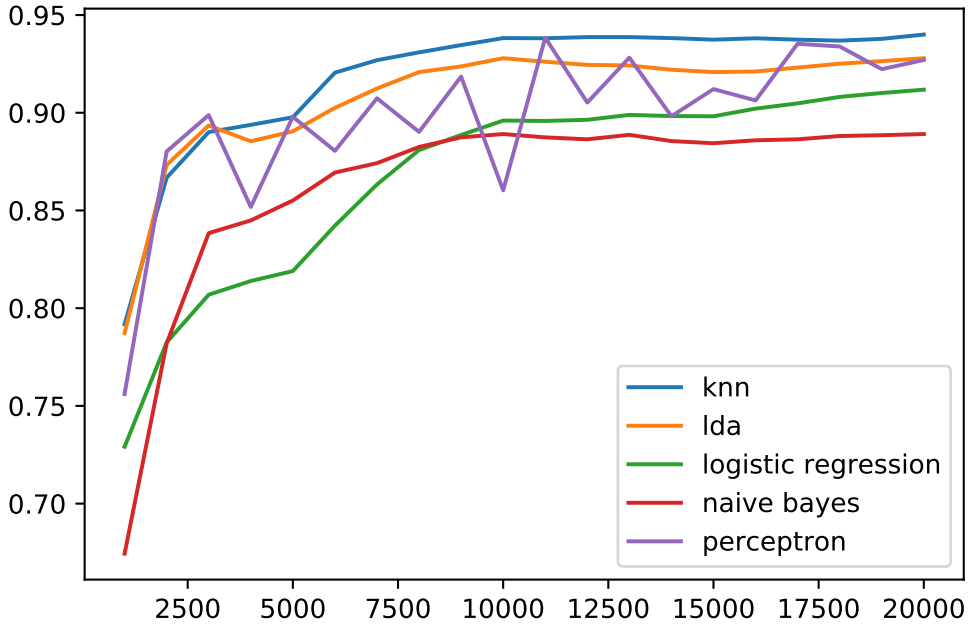

Figura 1: Gráfico de desempenho dos 5 classificadores

O tamanho da base de treinamento influencia cada classificador de maneira diferente, então alguns classificadores apresentaram resultados diferentes entre si. Apresentamos a tabela 1 para enumerar cada um dos classificadores e o tamanho do bloco no qual a base de treinamento deixa de ser relevante.

Classificador	Tamanho do bloco de treino	Acurácia
kNN	8000	0.93
LDA	8000	0.92
Logistic Regression	19000	0.91
Naive Bayes	8000	0.88
Perceptron	9000	0.91

Tabela 1: Tamanho da base de treinamento e acurácia máxima atingida

Depois de verificar os resultados é possível perceber que os classificadores kNN, LDA e Naive Bayes estabilizam seu resultado com base de treinamento com tamanho de 8000. O Perceptron apresenta um resultado serrilhado, oscilando para diferentes tamanhos, mas apresenta um resultado satisfatório com 9000 entradas. Já o classificador, Logistic Regression precisou de 19000 entradas de treinamento para chegar ao seu resultado máximo.

3 Atividade - II

Atividade: Indique qual é o classificador que tem o melhor desempenho com poucos dados = 1000 exemplos. A tabela 2 apresenta os resultados a serem anali-

sados.

Classificador	Tamanho do bloco de treino	Acurácia
kNN	1000	0.79
LDA	1000	0.78
Logistic Regression	1000	0.72
Naive Bayes	1000	0.67
Perceptron	1000	0.75

Tabela 2: Tamanho da base = 1000 e acurácia comparada

O kNN é o classificar com melhor desempenho para uma base de treinamento de 1000 entradas. Mas vale a pena ressaltar que o LDA também apresentou bom resultado.

4 Atividade - III

Atividade: Indique o classificador que tem melhor desempenho com todos os dados de treinamento. A tabela 3 apresenta os resultados a serem analisados.

Classificador	Tamanho do bloco de treino	Acurácia
kNN	20000	0.94
LDA	20000	0.92
Logistic Regression	20000	0.91
Naive Bayes	20000	0.88
Perceptron	20000	0.75

Tabela 3: Tamanho da base = 20000 e acurácia comparada

O melhor desempenho é dado pelo classificar kNN atingindo uma acurácia de 94% para uma base de treinamento de 20000 entradas.

5 Atividade - IV

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

Para responder essa atividade, o tempo de todos os testes foram capturados. Agrupando os classificadores e calculando o tempo médio para todos os testes realizados, pode-se apresentar a tabela 4.

Classificador	Tamanho do bloco de teste	Tempo médio em (ms)
kNN	58646	73.85
LDA	58646	0.02
Logistic Regression	58646	1.12
Naive Bayes	58646	0.56
Perceptron	58646	0.01

Tabela 4: Tamanho do teste 58646, comparada com o tempo médio de cada classificador

O classificador mais rápido é o Perceptron com 0.01 ms de tempo médio, em segundo lugar o LDA com 0.02 ms.

6 Atividade - V

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

Abaixo são apresentados as matrizes confusões para os 5 classificadores. As matrizes confusão apresentam a quantidade de acertos e erros por classe e também a sua porcentagem, que serão muito úteis para análise.

Como visto na seção 4, **Atividade - III**, apontou-se que o melhor classificador é o kNN. Utilizando sua matriz confusão, figura 2, pode-se observar que a sua maior confusão foi entre as classes 3 e 5, com um erro de 6.8% ou 377 erros.

Continuando com a análise, pode-se escolher o classificador LDA, que ficou em segundo lugar na **Atividade - III**. Utilizando sua matriz confusão, figura 3, pode-se observar que sua maior confusão foi entre as classe 8 e 9, com um erro de 6.3% ou 361 erros.

É interessante observar que os dois classificadores, embora tenham acurácias parecidas, apresentem dificuldades em predizer classes diferentes.

Conclui-se que os erros não são os mesmos para todos os classificadores.

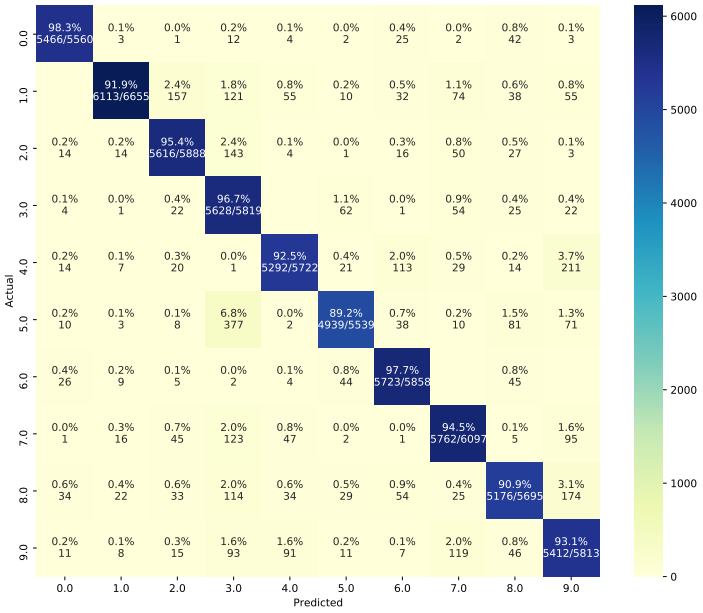

Figura 2: Matriz de confusão do classificador kNN

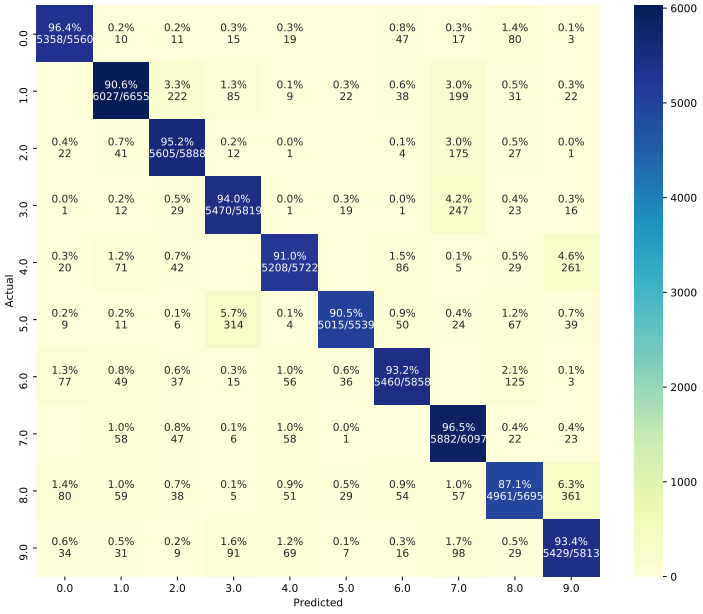

Figura 3: Matriz de confusão do classificador LDA

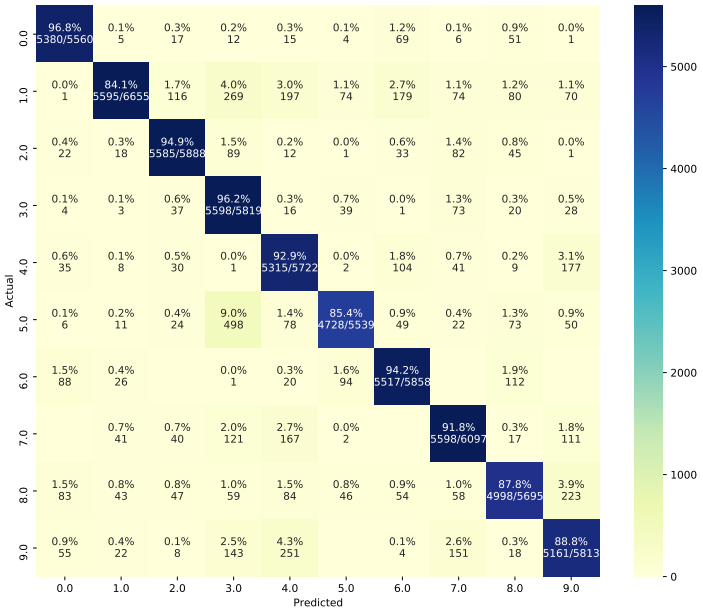

Figura 4: Matriz de confusão do classificador Logistic Regression

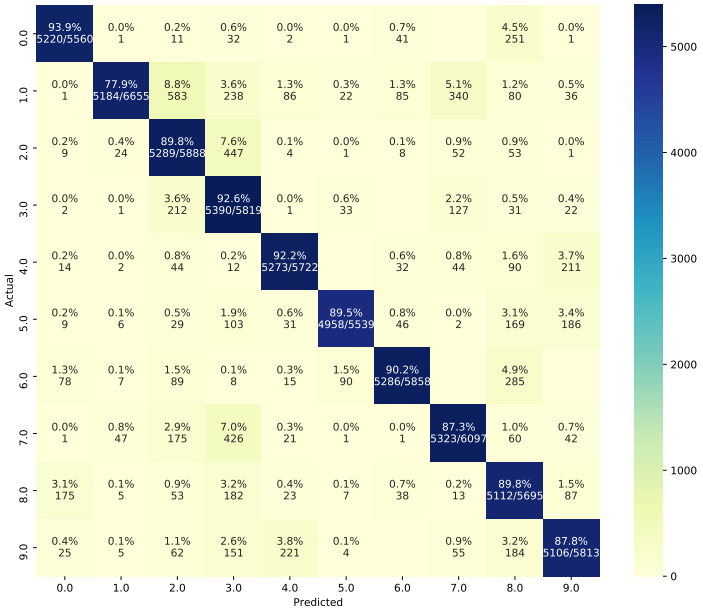

Figura 5: Matriz de confusão do classificador Naive Bayes

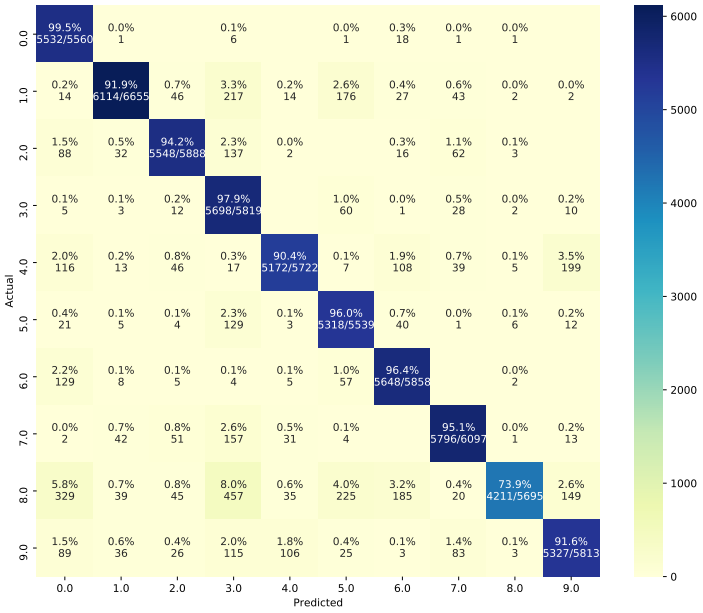

Figura 6: Matriz de confusão do classificador Perceptron