3º Trabalho Computacional
(TIP7077 - Inteligência Computacional Aplicada)

Aluno: Abelardo Vieira Mota
Matrícula: 366598

Sobre os dados

Sobre os abalones

O data set utilizado neste trabalho chama-se Abalone e encontra-se disponível na seguinte página web https://archive.ics.uci.edu/ml/datasets/Abalone.
De acordo com o repositório do data set, os dados foram doados em 01/12/1995 e têm com origem o estudo *"The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait"*.

Os dados referem-se a exemplares de abalone, tipo de moluscos gatrópodes comestíveis, e são compostos de 9 atributos(Tabela 1).

Tabela 1 - Atributos do data set

Name	Data Type	Measurement	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	Perpendicular to length
Height	continuous	mm	With meat in shell
Whole weight	continuous	grams	Whole abalone
Shucked weight	continuous	grams	Weight of meat
Viscera weight	continuous	grams	Gut weight (after bleeding)
Shell Weight	continuous	grams	Agter being dried
Rings	integer		+1.5 gives the age in years

Uma motivação para o uso de aprendizado de máquina apresentada no repositório é a seguinte: uma das tarefas realizadas com abalones é determinar suas idades. Tal como com as árvores, a estimação da idade dos abalones pode ser feita pela contagem de aneis que se formam em seus corpos. Para tanto, realiza-se a abertura de suas conchas, um processo de coloração e então a contagem, com auxílio de microscópio, da quantidade de aneis.

Comparado com o processo de obtenção de outras características dos abalones, o processo de contagem de aneis é bastante complexo, além de a associação entre idade e quantidade de aneis ser influenciada por outros fatores ambientais. Propõe-se então a utilização de outras características dos abalones para realizar a predição de suas idades.

Informações sobre o data set

O data set consiste nos atributos apresentados na Tabela 1 para 4177 exemplares.

De acordo com o arquivo **abalone.names**, disponibilizado conjuntamente com o data set, os atributos possuem as seguintes características estatísticas(Tabela 2):

Tabela 2 - Características estatísticas dos atributos

	Length	**Diam**	**Height**	**Whole**	**Shucked**	**Viscera**	**Shell**	**Rings**
Min	0.075	0.055	0.000	0.002	0.001	0.001	0.002	1
Max	0.815	0.650	1.130	2.826	1.488	0.760	1.005	29
Mean	0.524	0.408	0.140	0.829	0.359	0.181	0.239	9.934
SD	0.120	0.099	0.042	0.490	0.222	0.110	0.139	3.224

A distribuição dos dados(Figura 1), correlações entre suas features(Figura 2) e box plot das features(Figura 3):

Figura 1 - Distribuição dos dados por classes

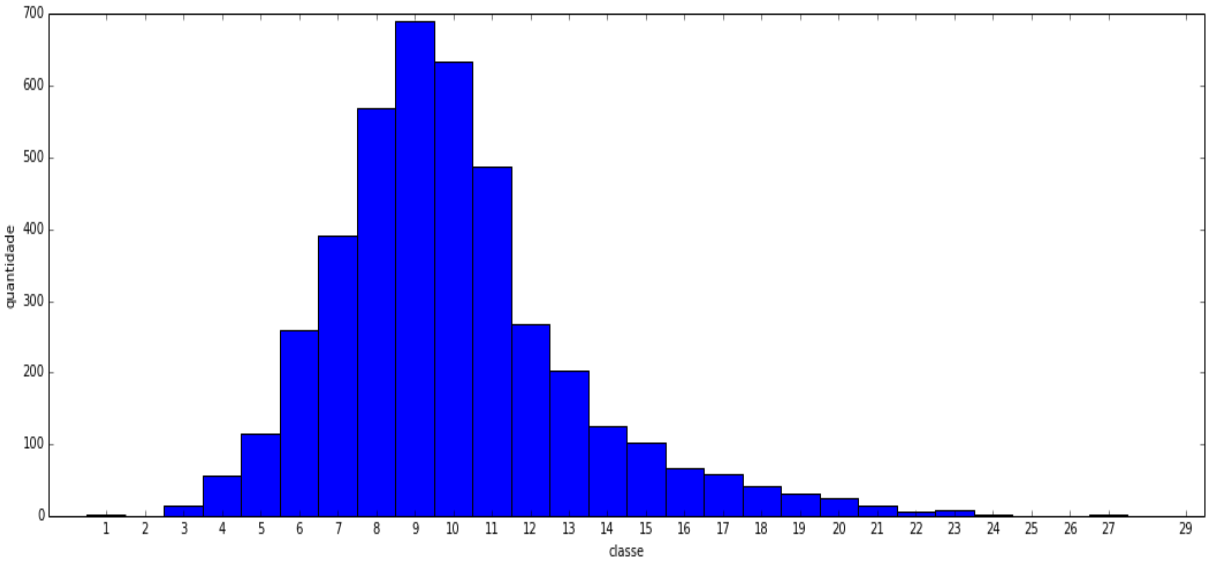


Figura 2 - Correlação entre as features

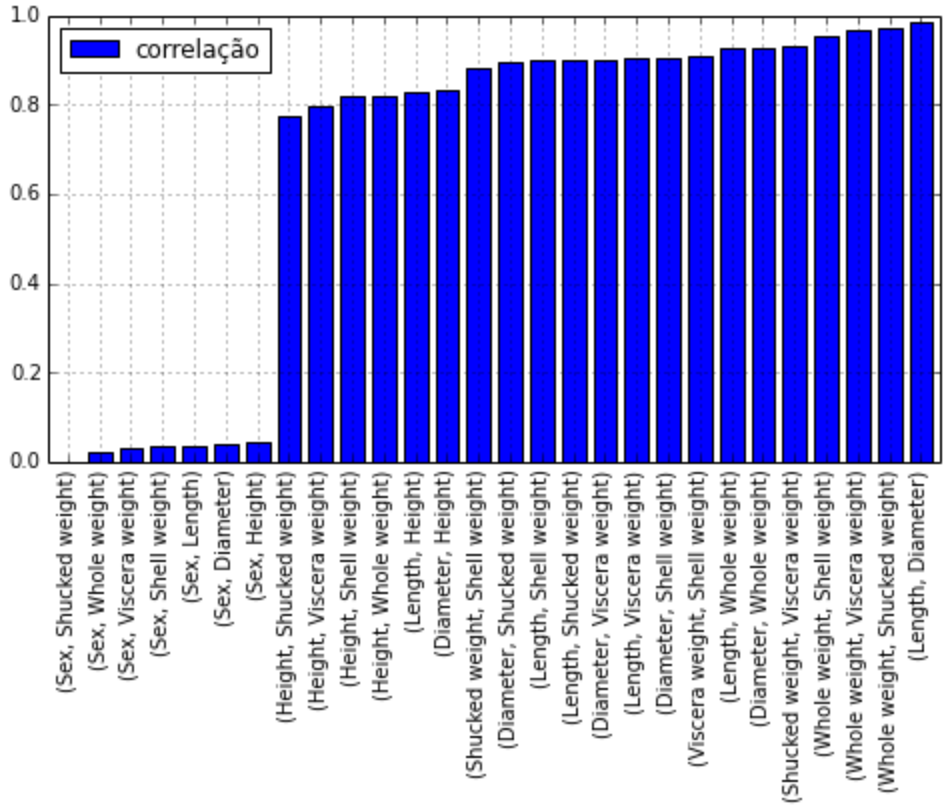
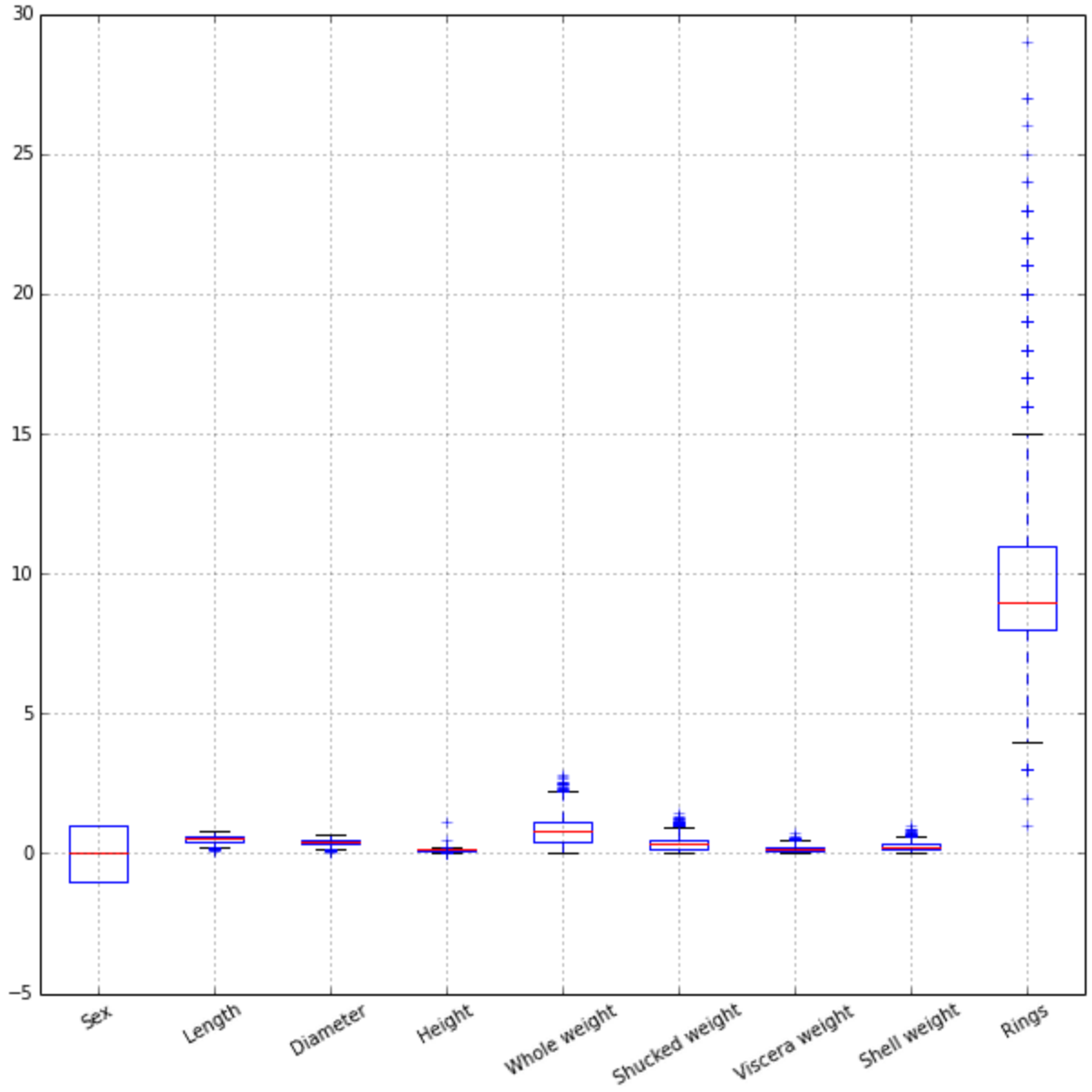


Figura 3 - Boxplot das features



Discussão sobre o data set

O dataset Abalone possui um conjunto de classes para as quais há poucos exemplos, como a **1**, **2**, **25** e **26**.
Tal situação pode configurar um problema, chamado por *desbalanceamento de classes*, dado que a quantidade de exemplos dessas classes pode ser insuficiente para que o modelo utilizado generalize bem para elas.
Algumas das soluções para tal problema são:

- up-sampling: replicação de exemplos de classes com poucos exemplos
- down-sampling: remoção de exemplos de classes com muitos exemplos

Material e métodos

Ambiente Python

Para o desenvolvimento deste trabalho, foi utilizada a linguagem Python, versão 2.7, e um conjunto de bibliotecas e ferramentas utilizadas em computação científica, descritas a seguir(Tabela 3).

Tabela 3 - Bibliotecas e ferramentas utilizadas

Nome	Descrição	Referência
Numpy	Biblioteca Python para manipulação de vetores.	http://www.numpy.org/
Matplotlib	Biblioteca Python para plotagem de gráficos.	http://matplotlib.org/
Pandas	Biblioteca Python para análise de dados.	http://pandas.pydata.org/
IPython	Console interativo para execução de códigos Python.	http://ipython.org/
Pybrain	Biblioteca Python para redes neurais artificiais.	http://pybrain.org/

Escolhi esse ambiente para o desenvolvimento do trabalho pois já possuo certa experiência e pretendo aprimorar meus conhecimentos nele.

Existem na internet diversos textos comparando esse ambiente Python com outros de mesma finalidade, como o Matlab e o Octave. A seguir apresento algumas diferenças entre o ambiente Python e o Matlab(Tabela 4).

Tabela 4 - Algumas diferenças entre o ambiente Python e o Matlab

Característica	Ambiente Python	Matlab	Observação
Licença	GPL-compatible	Proprietário	Por ser proprietário, o Matlab não permite que o usuário veja e possa testar a implementação de seus códigos, além de não permitir contribuição direta da comunidade de usuários.
Custo	Gratuito	Pago	

Curva de aprendizagem	Curta	Curta	Apesar de as curvas de aprendizagem dos dois ambientes serem curtas, para computacão científica, a curva do Matlab pode ser mais curta, visto que essa é sua finalidade.
Performance	Média	Média	De acordo com http://wiki.scipy.org/PerformancePython, a performance do Python(com numpy) e do Matlab para um conjunto de testes foi similar. Outras linguagens obtiveram performance superior, como o C++, com performance aproximadamente 10 vezes melhor.

Os códigos utilizados no trabalho podem ser visualizados nas páginas:

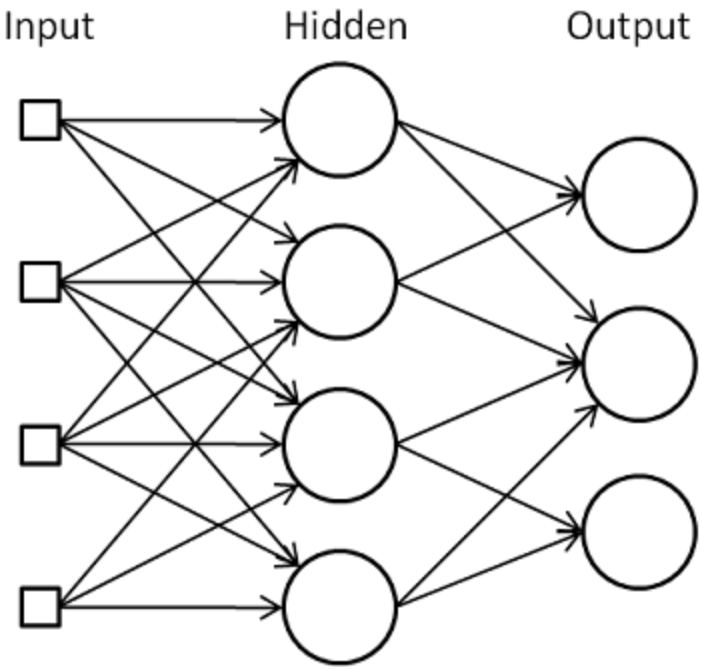
1. Análise dos dados
2. Treinamento do modelo

Rede MLP

O modelo utilizado neste trabalho foi a rede Multilayer Perceptron(MLP), que é uma rede neural artificial composta por três camadas:

1. **input**: camada cujos neuronios recebem como entrada o input da rede
2. **hidden**: camada cujos neuronios recebem como entrada o output da camada input
3. **output**: camada cujos neuronios recebem como entrada o output da camada hidden e retornam o output da rede

A seguinte imagem exemplifica o formato de uma rede MLP:



Neste trabalho, utilizei a seguinte codificação:

c(y**)** = vetor de 0's a menos da (y-1)-ésima posição
Por exemplo, caso y = 17, **c(**y**) = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]**
↳16-ésima posição

Portanto, a rede MLP gerada possui 29 neurônios na camada output, número de classes do data set Abalone.

Foi utilizada uma taxa de aprendizado com valor igual a 0.01.

Foram realizados 100 treinamentos e coleta de dados, com 80% dos dados escolhidos aleatoriamente para treinamento e os demais para teste.

A quantidade de neurônios na camada oculta foi determinada pela regra de Kolmogorov, dando um total de 17 neurônios.

Observações

Os dados do atributo *Sex*, por serem não numéricos, foram convertidos para números, utilizando o seguinte mapeamento: M=-1.0 F=1.0 I=0.0.

O tempo de execução dos 100 treinamentos realizados foi de 3 horas e 11 minutos.
As informações estatísticas sobre as taxas de acertos para o testing set foram:

Tabela 5 - estatísticas da taxa de acerto para o testing set

média	0.789138 ~ 78%
desvio padrão	0.018965 ~ 1.8%
mínimo	0.745215 ~ 74%
máximo	0.855263 ~ 85%
mediana	0.787081 ~ 78%

Figura 2 - histograma da taxa de acerto para o testing set

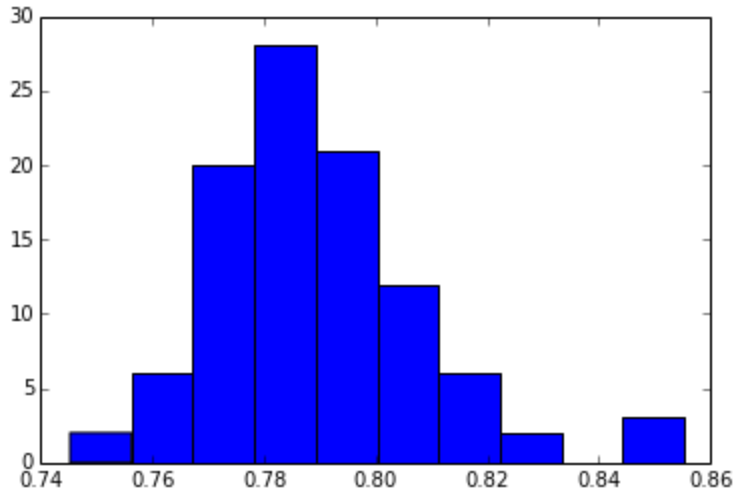
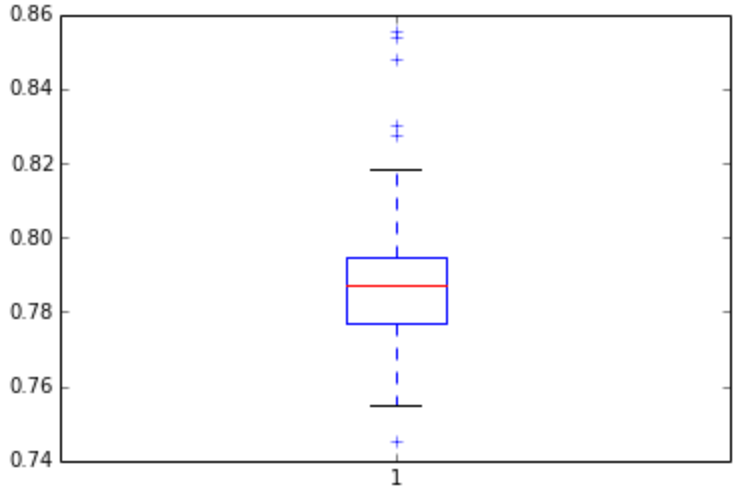


Figura 3 - boxplot da taxa de acerto para o testing set



Comparação entre MLP, Perceptron Simples e Classificador linear dos mínimos quadrados

Tabela 6 - Comparativo de taxas

modelo	mínimo	média	máximo
MLP	74%	78%	85%
Perceptron simples	17%	22%	25%
Classificador linear dos minimos quadrados	20%	23%	28%

Discussão

Sobre a taxa de acerto

A utilização da rede MLP foi a que trouxe melhor taxa de acerto dentre os três modelos até então utilizados.
Uma explicação é que a rede MLP, com a quantidade de neurônios utilizados, consegue descrever dados com uma complexidade maior que a que os outros modelos conseguem descrever.