1º 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.

Sobre o data set

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

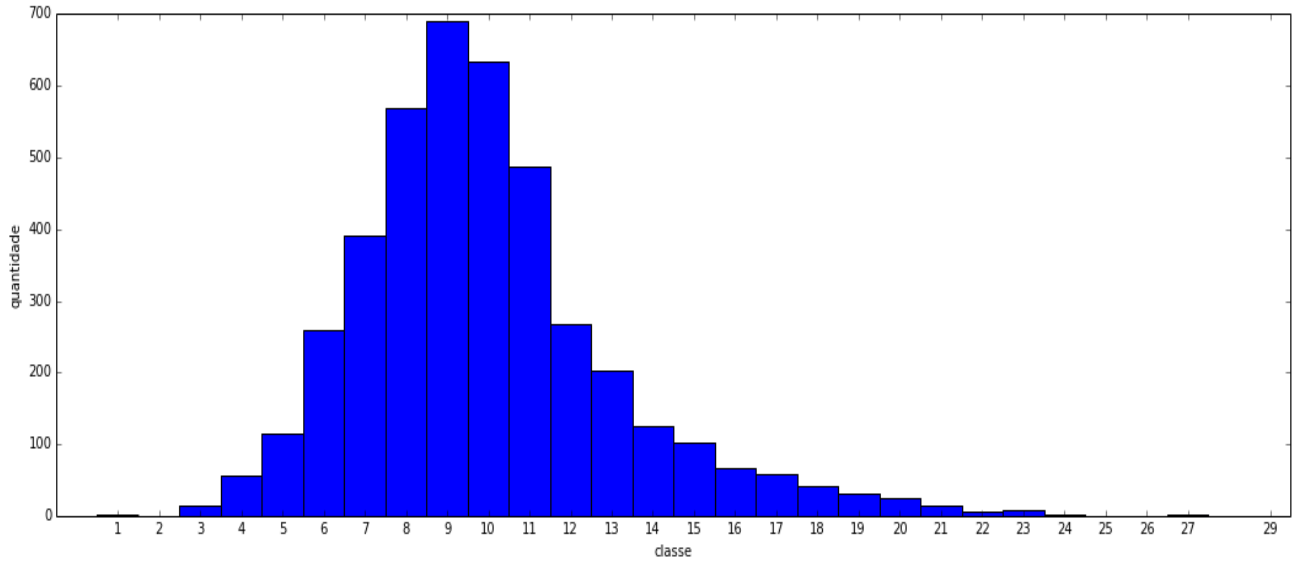
De acordo com o arquivo **abalone.names**, disponibilizado conjuntamente com o data set, os atributos possuem as seguintes características estatísticas(Tabela 2):

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
Correl	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.0

Os dados estão distribuídos por classes de acordo com a próxima figura(Figura 1):

Figura 1 - Distribuição dos dados por classes



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/

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 de outros com 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 na página http://nbviewer.ipython.org/github/abevieiramota/TrabalhosDeICA/blob/master/Trabalho%201.ipynb?create=1 e encontram-se em um repositório no github, de URL https://github.com/abevieiramota/TrabalhosDeICA

Método dos mínimos quadrados

Utilizei um modelo de regressão linear múltiplo com o método dos mínimos quadrados.
Testes foram realizados particionando-se aleatoriamente os dados entre training-set e testing-set, com proporções (80%: 20%) e (50%: 50%), com 100 rodadas e com a técnica de regularização de Tikhonov, para valores de λ iguais a 0.0, 0.001, 0.005 e 0.01.
A cada rodada foram calculados os pesos do modelo, o erro quadrático médio, a taxa de acerto de todo o testing-set e a taxa de acerto para cada classe.

Observações

Os dados do atributo *Sex*, por serem não numéricos, foram convertidos para números, utilizando o seguinte mapeamento: M=-1 F=1 I=0. Pesquisei um pouco sobre qual a influência da escolha de tais valores numéricos no treinamento do modelo, mas não encontrei material sobre o assunto.

Resultados

Foram executados treinamentos e testes com o modelo de regressão linear múltiplo por 100 rodadas, com os seguintes parâmetros e resultados(valores arredondados para cima)(valores máximos em vermelho)(Tabela 5):

Tabela 5 - Parâmetros e resultados dos treinamentos e testes

Proporção entre training-set e testing-set(em %)	Lambdas da regularização de Tikhonov	Taxa de acerto média	Taxa de acerto mínima	Taxa de acerto máxima	Desvio padrão da taxa de acerto
80 20	**0.0000**	0.233377	0.200000	**0.279042**	0.014579
50 50	**0.0000**	0.230024	0.190613	0.252874	0.010296
80 20	**0.0001**	0.230766	0.189222	0.271856	0.015283
50 50	**0.0001**	0.229516	0.207854	0.252395	0.009439
80 20	**0.0005**	0.234132	0.179641	0.273054	**0.016270**
50 50	**0.0005**	0.230072	0.205460	0.255268	0.009814
80 20	**0.0010**	**0.234539**	0.198802	0.275449	0.012818
50 50	**0.0010**	0.232840	**0.209770**	0.255268	0.009229

Também foram coletadas as taxas de acerto média para cada classe de abalones(Tabelas 6.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8):

Tabela 6.1 - Taxa de acerto média para cada classe. Parâmetros 80 20 0.0000

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.316709	15	NaN	22	NaN
2	NaN	9	0.558812	16	NaN	23	NaN
3	NaN	10	0.268241	17	NaN	24	NaN
4	NaN	11	0.167750	18	NaN	25	NaN
5	NaN	12	0.019197	19	NaN	26	NaN
6	0.162964	13	0.050071	20	NaN	27	NaN
7	0.309450	14	NaN	21	NaN	28	NaN

Tabela 6.2 - Taxa de acerto média para cada classe. Parâmetros 50 50 0.0000

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.324855	15	0.022727	22	NaN
2	NaN	9	0.525929	16	0.033368	23	NaN
3	NaN	10	0.257869	17	0.066345	24	NaN
4	NaN	11	0.187444	18	0.050000	25	NaN
5	0.086638	12	0.012459	19	NaN	26	NaN
6	0.181798	13	0.047419	20	NaN	27	NaN
7	0.290409	14	0.016266	21	NaN	28	NaN

Tabela 6.3 - Taxa de acerto média para cada classe. Parâmetros 80 20 0.0001

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.302923	15	NaN	22	NaN
2	NaN	9	0.565790	16	0.111111	23	NaN
3	NaN	10	0.253862	17	NaN	24	NaN
4	NaN	11	0.170759	18	NaN	25	NaN
5	NaN	12	0.020771	19	NaN	26	NaN
6	0.150554	13	0.049320	20	NaN	27	NaN
7	0.317463	14	NaN	21	NaN	28	NaN

Tabela 6.4 - Taxa de acerto média para cada classe. Parâmetros 50 50 0.0001

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.303467	15	0.020833	22	NaN
2	NaN	9	0.535969	16	0.037264	23	NaN
3	NaN	10	0.259602	17	0.040000	24	NaN
4	NaN	11	0.183883	18	NaN	25	NaN
5	0.091492	12	0.013800	19	NaN	26	NaN
6	0.169095	13	0.043274	20	NaN	27	NaN
7	0.314990	14	0.028884	21	NaN	28	NaN

Tabela 6.5 - Taxa de acerto média para cada classe. Parâmetros 80 20 0.0005

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.315788	15	NaN	22	NaN
2	NaN	9	0.555365	16	NaN	23	NaN
3	NaN	10	0.255722	17	NaN	24	NaN
4	NaN	11	0.176831	18	NaN	25	NaN
5	NaN	12	0.022155	19	NaN	26	NaN
6	0.151704	13	0.050045	20	NaN	27	NaN
7	0.319436	14	NaN	21	NaN	28	NaN

Tabela 6.6 - Taxa de acerto média para cada classe. Parâmetros 50 50 0.0005

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.285894	15	0.017857	22	NaN
2	NaN	9	0.541056	16	0.038681	23	NaN
3	NaN	10	0.262577	17	NaN	24	NaN
4	NaN	11	0.177219	18	0.055556	25	NaN
5	0.051930	12	0.010705	19	NaN	26	NaN
6	0.154468	13	0.047810	20	NaN	27	NaN
7	0.335252	14	0.022750	21	NaN	28	NaN

Tabela 6.7 - Taxa de acerto média para cada classe. Parâmetros 80 20 0.0010

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.320880	15	NaN	22	NaN
2	NaN	9	0.556177	16	NaN	23	NaN
3	NaN	10	0.256546	17	NaN	24	NaN
4	NaN	11	0.176059	18	NaN	25	NaN
5	NaN	12	0.022750	19	NaN	26	NaN
6	0.166359	13	0.049073	20	NaN	27	NaN
7	0.315539	14	NaN	21	NaN	28	NaN

Tabela 6.8 - Taxa de acerto média para cada classe. Parâmetros 50 50 0.0050

Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média	Classe	Taxa de acerto média
1	NaN	8	0.335377	15	NaN	22	NaN
2	NaN	9	0.518703	16	0.035203	23	NaN
3	NaN	10	0.266613	17	NaN	24	NaN
4	NaN	11	0.184016	18	NaN	25	NaN
5	0.107577	12	0.012007	19	NaN	26	NaN
6	0.180594	13	0.043151	20	NaN	27	NaN
7	0.297703	14	0.024265	21	NaN	28	NaN

Discussão

Sobre a taxa de acerto total

A melhor taxa de acerto foi 0.279042, obtida com os parâmetros 80% 20% e 0.0000. Observo que a taxa de acerto obtida com esse modelo foi bastante baixa, mas ainda é maior que uma classificação aleatória, que possui taxa de acerto, de acordo com a probabilidade, de 0.035714286.

Sobre a taxa de acerto por classe

Um resultado que chama a atenção é o da taxa média de acerto por classes para algumas classes, que obtiveram resultado NaN(Not a Number), indicando que a classe não foi encontrada no testing-set para alguma rodada. Ao notar isso, veio-me a questão: no cálculo do valor médio das taxas de acerto por classe, que valor devo atribuir a uma classe numa rodada em que ela não apareça no testing-set? No experimento feito, atribuí o valor NaN que tem o efeito de, no cálculo da média com outros números, resultar em um NaN. Deveria ter atribuído uma taxa de acerto igual a 0.0? Um problema ao se usar a abordagem do NaN é que uma ocorrência de NaN torna a média igual a NaN, perdendo-se toda possível taxa de acerto diferente de NaN. Um problema com ao se usar a abordagem do 0.0 é que ele irá diminuir significativamente a média das taxas de acerto.

Vale notar também que para muitas classes, a quantidade de exemplares é pouco significativa, como é o caso das classes 1, 2, 25, 26 e 29, que possuem cada apenas um exemplar.

Sobre se é possível melhorar

Uma dúvida que surgiu durante o desenvolvimento do trabalho foi sobre como verificar se o modelo adotado, no caso o linear múltiplo, é apropriado ou não para os dados em estudo. Quais métricas devem ser analisadas? Taxa de acerto? Erro quadrático médio? Existem heurísticas para ajudar a resolver tal questão? Existem processos bem definidos para desenvolvimento de

modelos de aprendizado, com fases bem definidas? Essa questão surgiu principalmente por, aparentemente, o modelo linear múltiplo não ser um bom modelo para os dados em estudo, fazendo surgir diversas vezes indícios de que “algo está errado”.

Na página repositório dos dados, http://archive.ics.uci.edu/ml/datasets/Abalone, na lista de papers que citam o data set Abalone, constam vários papers com, aparentemente, modelos diferentes do linear múltiplo, como:

- Speeding Up Fuzzy Clustering with Neural Network Techniques
- CLOUDS: A Decision Tree Classifier for Large Datasets
- Complete Cross-Validation for Nearest Neighbor Classifiers

Conclusão

A taxa de acerto obtida no experimento mostrou-se abaixo da esperada por mim, motivando o teste com o modelo linear múltiplo com parâmetros diferentes dos utilizados neste trabalho, a replicação de dados de classes que possuem poucos exemplares, ou a utilização de outros modelos.
Os dados utilizados também não parecem apropriados para o uso de tal modelo ou o uso da seleção aleatória dos conjuntos training-set e testing-set, visto que, para algumas classes, a quantidade de exemplos é bastante pequena.