

Aprendizado de Máquina para Classificar Animais

Bruna Prauchner Vargas¹

¹Escola Politécnica – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)


bruna.prauchner@acad.pucrs.br

1. Introdução

O quarto trabalho, proposto na disciplina de Fundamentos de Inteligência Artificial, consiste em escolher um *dataset* para aprendizado de máquina e fazer experimentos, utilizando 4 técnicas diferentes. O *dataset* escolhido é sobre a classificação de diversos animais, contém 17 atributos, e não tem valores faltantes.

2. Informações sobre o dataset

O *dataset* foi retirado do repositório de aprendizado de máquina da Universidade da Califórnia Irvine[1].



UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Zoo Data Set
Download: [Data Folder](#) [Data Set Description](#)

Abstract: Artificial, 7 classes of animals

Data Set Characteristics:	Multivariate	Number of Instances:	101	Area:	Life
Attribute Characteristics:	Categorical, Integer	Number of Attributes:	17	Date Donated	1990-05-15
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	217391

Figura 1. Informações sobre o dataset escolhido para o trabalho

Lista dos atributos:

1. Nome do animal;
2. Pelo;
3. Penas;
4. Ovos;
5. Leite;
6. No ar;
7. Aquático;
8. Predador;
9. Dentado;
10. Espinha dorsal;
11. Respira;
12. Venenoso;
13. Barbatanas;

14. Patas;
15. Cauda;
16. Domesticado;
17. Tamanho;
18. Tipo.

O nome do animal é único para cada instância, e os itens 14 e 8 são números e o resto dos atributos são valores booleanos.

3. Aprendizado de Máquina

O aprendizado de máquina visa melhorar a performance de um dado agente em tarefas futuras após observar o estado do mundo. Aprendizado é útil porque não é possível antecipar todas as situações que o agente vai encontrar e é difícil programar todas as mudanças possíveis.

3.1. Aprendizado Supervisionado

Utiliza um conjunto de dados contendo entrada e saída e aprende uma função que gera saídas apropriadas para os novos valores de entrada. O aprendizado é a busca no espaço de possíveis hipóteses que irá se comportar bem mesmo em dados nunca vistos, este conjunto de dados nunca visto antes é chamado de conjunto de teste.

Quando uma hipótese possui um bom desempenho em dados nunca vistos, pode-se dizer que ela generaliza bem. Ou seja, acertou dados que não estavam no conjunto de treinamento, e isso é ótimo porque caso contrário, existiria o problema de *overfitting*, quando gera-se como hipótese uma função muito complexa que se adequa ao padrão dos dados e não generaliza bem.

Após fazer a leitura do *dataset* e dividir esse conjunto de dados em 80% para treino e 20% para teste, pode-se aplicar as técnicas de aprendizado de máquina. Para esse trabalho foram feitas duas versões, a primeira usa 80% para o conjunto de treinamento, e a segunda, usa apenas 20% para treinamento e 80% para teste. Ambas versões podem ser conferidas no *Colaboratory*.

3.2. Técnicas Utilizadas

As técnicas usadas para o desenvolvimento do trabalho, serão listadas com uma breve explicação, os resultados dos experimentos estão no *Colaboratory*.

- Regressão Logística: permite a predição de valores sobre pertencer ou não a uma classe;
- *Feedforward Neural Network*: os dados se movem em apenas uma direção, para frente, dos nodos de entrada, através dos nodos escondidos (se houver) e para os nodos de saída. Não há ciclos ou *loops* na rede;
- KNN: verifica-se quais são as classes dos K vizinhos e a classe mais frequente será atribuída à classe do elemento desconhecido;
- *K-Means*: é um método de aprendizado não supervisionado. Particiona em n observações dentre k grupos, onde cada observação pertence ao grupo mais próximo da média.

3.3. Comparando as Duas Versões

Comparando as duas versões na primeira técnica que é regressão logística, nota-se que as classes 3, 5, 6 não obtiveram respostas certas na versão 2, e só a classe 3 não tem respostas corretas na versão 1. Logo, é considerado para regressão logística que a versão 1 teve um melhor desempenho.

Na técnica KNN, sabe-se que, precisão, revocação e *f1-score* quanto mais próximo de 1, melhor é. Ambas as versões ficaram com zero na classe 3; e a versão 2, obteve zero nas classes, 3, 5, 6, 7. Novamente a versão 1 teve melhor desempenho.

Usando rede neural, obtém-se todos os resultados em uma coluna só, assim, percebeu-se que não seria possível utilizar esse *dataset* com essa técnica.

Em *K-Means* a versão 1 foi um completo fracasso, não acertando nenhuma resposta. Na versão 2, a diagonal principal que contem os valores corretos só é composta por 2 classes.

3.3.1. Caso Especial: Rede Neural

Aplicando a técnica de *Feed forward neural network* no *dataset* zoo, percebe-se que na matriz de confusão, todos os valores ficam concentrados em uma coluna só, ou seja, como se todas as respostas estivessem erradas. Acredita-se que o principal motivo para isso acontecer é pelo *dataset* ser muito pequeno, apenas 5KB. Isto aconteceu em ambas as versões do trabalho, portanto, uma outra versão foi criada, usando um *dataset* muito maior, para testar essa técnica. Na versão 3 no *Colaboratory* pode-se conferir os novos resultados, o *dataset* utilizado é mais difícil de compreender, por ser de um assunto mais complexo, ele foi escolhido por não ter valores faltantes e ter tamanho adequado (1,4MB).

4. Conclusão

Por fim, depois de estudar sobre aprendizado de máquina[2], entender seus conceitos e aplicações percebeu-se que ao longo do trabalho, era preciso saber qual tipo de aprendizado utilizar dependendo de qual *dataset* seria o escolhido, e depois analisar a matriz de confusão para compreender seus resultados.

Referências

- [1] *Dataset from UCI*. <http://archive.ics.uci.edu/ml/datasets/zoo>. Accessed: 2018-06-22.
- [2] Stuart Russell, Peter Norvig e Ernest Davis. *Artificial intelligence: a modern approach*. English. 3ª ed. Upper Saddle River, NJ: Prentice Hall, 2010.