Universidade Federal do Rio de Janeiro
Bacharelado em Ciência da Computação
Inteligência Artificial

Relatório do Trabalho de Machine Learning

Eduardo da Silva Barbosa - 116150432

Objetivo (Iris)

O objetivo do trabalho é construir um algoritmo que seja capaz de aprender (descobrir padrões), de modo que consiga prever os resultados de um posterior problema de forma autônoma. Para o treinamento, tal algoritmo terá como entrada informações sobre diversas flores, que possuem uma das seguintes classificações: Iris-setosa, Iris-versicolor ou Iris-virginica. Após o treinamento, tendo como entrada informações sobre as características de uma flor, o algoritmo deve ser capaz de prever a classificação da mesma.

Metodologia

Como base da metodologia foi utilizado o algoritmo PLA (Perceptron Learning Algorithm), O mesmo tem como objetivo gerar um hiperplano para tentar separar as flores em dois grupos diferentes, contudo temos três classificações de flores distintas, logo foi necessário gerar dois hiperplanos, onde o primeiro classifica se é do tipo setosa e o segundo se é do tipo versicolor, não sendo nenhuma das duas ela é do tipo virginica.

Caso não seja possível gerar um hiperplano que isole totalmente dois grupos de flores o PLA entrará em loop, para evitar que isso ocorra, foi definido um teto de 100 iterações, que nesse problema se mostrou o suficiente para se encontrar o melhor hiperplano. A cada iteração do PLA é gerado um hiperplano e o mesmo é associado ao seu índice de erros, de modo que ao final das 100 iterações é retornado o hiperplano que gerou o menor número de erros.

Vale ressaltar que 80% do dataset foi utilizado para treinamento e 20% para testar a eficiência da aprendizagem. Além disso, são feitas 100 simulações de aprendizagem e testes e ao final é feita a média de acertos e erros.

Por fim o algoritmo foi alterado para o segundo hiperplano isolar as flores de categoria virginica ao invés da versicolor, de modo que se não for classificada como setosa (pelo primeiro hiperplano) e nem como virginica (pelo segundo hiperplano) a flor é do tipo versicolor.

Resultados

```
Melhor w0 =  [-1.  -1.7 -5.1  7.1  3.4]
Melhor w1 =  [-90.   -0.8  42.1 -23.6  49.5]
Sertosa = 0 | Versicolor = 1 | Virginica = 2
Resultado final - ( Direita:Resultado da entrada de teste. | Esquerda: Gabaito )
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 2.] - [ 1.]  - Errou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 2.] - [ 2.]  - Acertou
[ 1.] - [ 2.]  - Errou
[ 2.] - [ 2.]  - Acertou
[ 1.] - [ 2.]  - Errou
[ 2.] - [ 2.]  - Acertou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
Acertos =  22
Erros =  8
100 Simulações de aprendizado e os resultados dos testes geraram as seguintes médias:
Média de acertos =  21.79
Média de erros =  8.21
```

Como pode ser observado tivemos uma taxa de aproximadamente 72,63 % de acertos (21,79/30 acertos) para o caso onde w0 representa o hiperplano que diz se é setosa ou não e w1 o hiperplano que diz se é versicolor ou não. Já alterando w1 para isolar as virginica tivemos resultados bem mais promissores como uma taxa de aproximadamente 94% de acertos, havendo casos com 100% de acerto. Segue abaixo o segundo caso.

```
Melhor w0 =  [ -1.   -1.5  -4.7  10.    4.6]
Melhor w1 =  [  63.    60.6   83.6 -114.2  -90.3]
Sertosa = 0 | Versicolor = 1 | Virginica = 2
Resultado final - ( Direita:Resultado da entrada de teste. | Esquerda: Gabaito )
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
[ 2.] - [ 2.]  - Acertou
Acertos =  30
Erros =  0
100 Simulações de aprendizado e os resultados dos testes geraram as seguintes médias:
Média de acertos =  28.23
Média de erros =  1.77
```

Conclusão

Analisando os resultados foi possível concluir que existe um hiperplano que consegue isolar perfeitamente as flores do tipo Iris-setosas fazendo com que seja sempre possível acertar essa classificação. Contudo não foi possível achar um hiperplano que conseguisse isolar 100% as Iris-versicolor e Iris-virginicas.

Também foi notado que após 100 iterações do PLA parou de haver um ganho na média de acertos e erros, assim concluindo-se que 100 iterações era o suficiente para encontrar o melhor hiperplano. Além disso, observou-se que reduzindo o número de iterações do PLA para 1 iteração, a taxa de acertos reduziram para aproximadamente 64,46%. Segue o resultado do teste com apenas uma iteração do PLA:

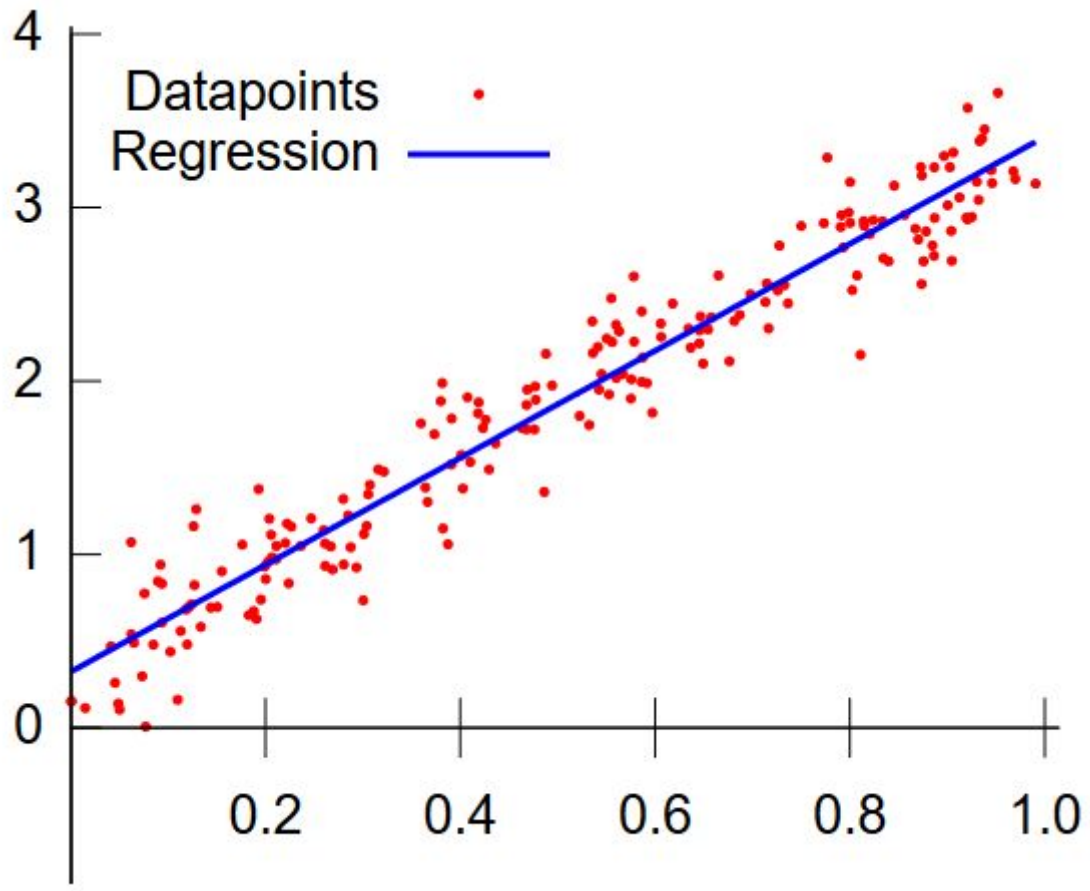
```
Melhor w0 =  [-1.  -1.2 -5.2  7.5  3.3]
Melhor w1 =  [ 0.  -1.3  6.4 -4.6 -0.7]
Sertosa = 0 | Versicolor = 1 | Virginica = 2
Resultado final - ( Direita:Resultado da entrada de teste. | Esquerda: Gabaito )
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 0.] - [ 0.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 1.]  - Acertou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
[ 1.] - [ 2.]  - Errou
Acertos =  17
Erros =  13
100 Simulações de aprendizado e os resultados dos testes geraram as seguintes médias:
Média de acertos =  19.34
Média de erros =  10.66
```

Objetivo (Salário)

Nesse segundo caso, para o treinamento, o algoritmo terá como entrada informações sobre diversos funcionários, onde tais informações influencia diretamente o salário dos mesmos. Após o treinamento, o algoritmo deve ser capaz de prever aproximadamente o salário de um determinado funcionário, baseado nas características.

Metodologia

Como base da metodologia foi utilizada a regressão linear, usando a abordagem dos mínimos quadrados, que tem como objetivo descobrir um hiperplano que aproxima da melhor forma os resultados. O hiperplano encontrado, equilibra da melhor forma o erro entre os todos os pontos. Abaixo segue um exemplo:



Vale ressaltar que nesse problema o dataset foi ajustado, onde no lugar das categorias(male, famale, assistant, associate, full, masters, doctorate) foram utilizados números para representar. O dataset teve 80% destinado para aprendizagem e 20% para os testes. Além disso, foi calculado o erro absoluto de cada resultado e o erro geral.

Resultados

Resultado	Valor ideal	Erro absoluto
37828.8529029	26775.0	11053.8529029
31290.2293531	33696.0	2405.77064685
33929.6840968	28516.0	5413.68409681
29423.2485289	24900.0	4523.24852887
33750.6215003	25748.0	8002.62150034
27144.0492316	29342.0	2197.95076842
28611.8783939	20690.0	7921.87839389
13459.1853258	17095.0	3635.81467415
14212.2079829	15350.0	1137.7920171
10532.2652883	16244.0	5711.73471167
13591.0114561	20300.0	6708.98854389

Erro dentro da amostra: 36501854.9657

OBS: A coluna erro representa o erro individual de cada entrada do bd de teste. O "Erro dentro da amostra" representa $\frac{1}{N}\sum_{n=1}^{N}(h(x_n)-y_n)^2$.

Conclusão

Analisando os resultados percebe-se que o método não se mostrou tão eficiente, havendo erro de mais de 8000, o que é um erro significativo. Contudo os valores resultantes não se mostraram extremamente divergentes do ideal.