Aprendizagem Automática - Lista 1

Regressão Linear

Israel de Castro Vidal – 370019

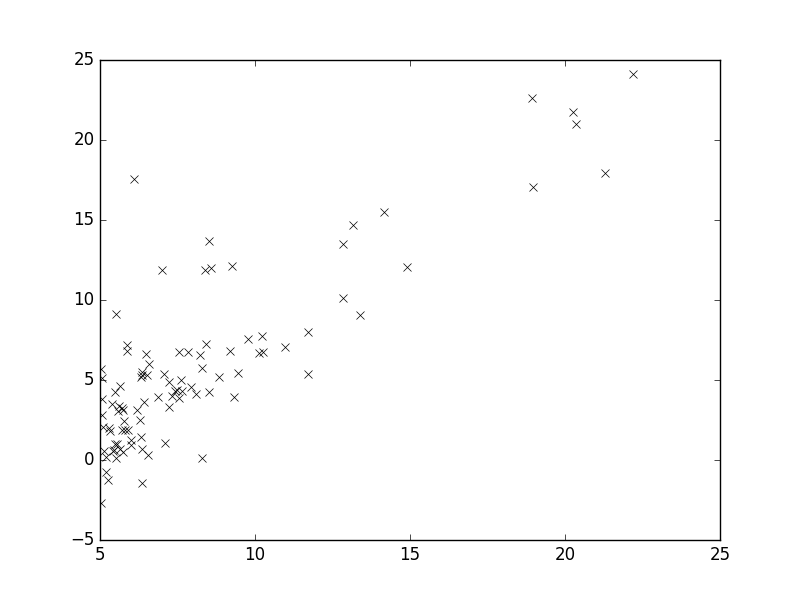
Deverá ser enviado ao professor, um arquivo texto contendo os gráficos, resultados e comentários requeridos em cada item.

1. **Regressão Linear Univariada**

- Carregue os dados contidos no arquivo ex1data1.txt.

O arquivo contem 97 linhas e 2 colunas de dados. Cada coluna se refere a uma variável. Neste problema, deve-se desenvolver um modelo onde a variável da coluna 2 será estimada a partir da variável da coluna 1.

**Apresentar**: Figura com os dados



Dataset1

**Comentários**: Um modelo de regressão linear parece ser adequado para os dados em questão? Comente.

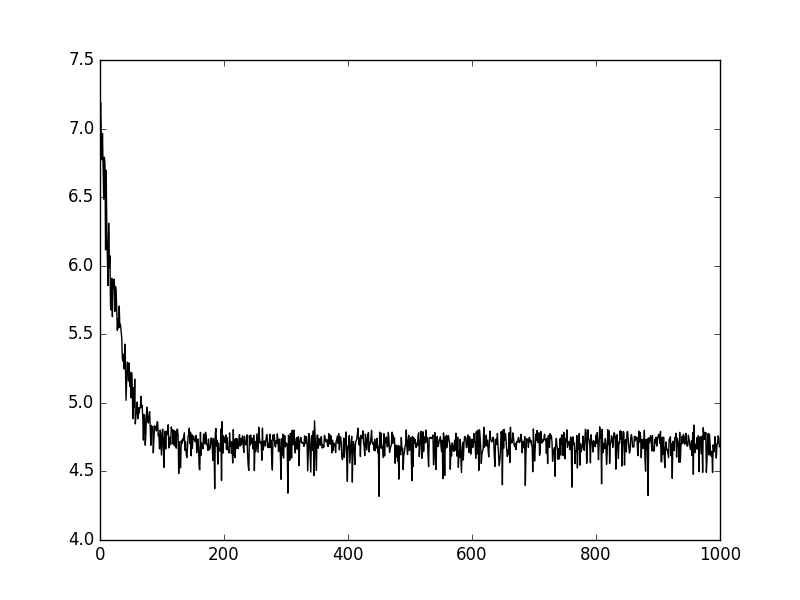
*Sim, um modelo de regressão linear parece adequado, pois, pelo gráfico anterior, parece existir uma reta que aproxima a função original minimizando o erro.*

- Implemente o algoritmo do gradiente descendente estocástico para encontrar os coeficientes da regressão.

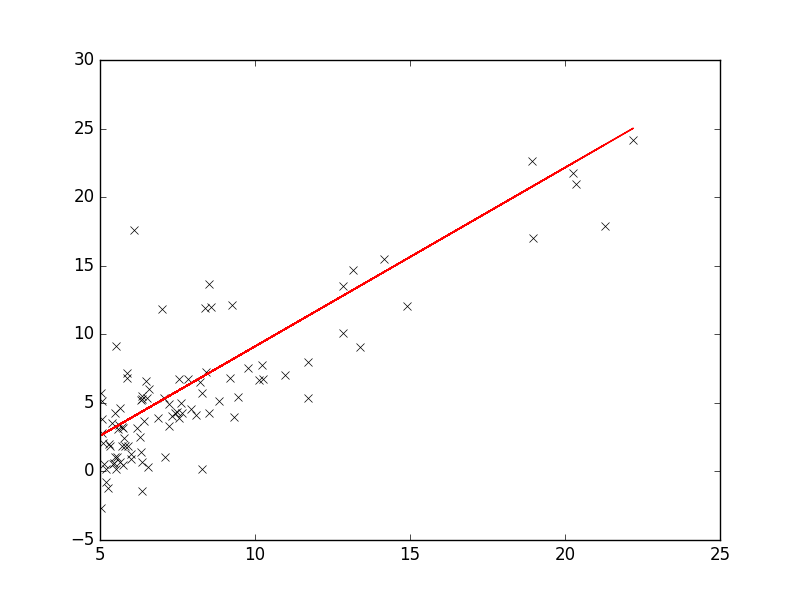
Para este algoritmo utilize α = 0.001 e utilize 1000 épocas de treinamento. Para cada época de treinamento, calcule o erro quadrático médio (EQM) . Plote o gráfico “épocas x EQM”

**Apresentar**: Valor final dos coeficientes e gráfico épocas x EQM.

*W0 = -3.917517058471562, W1 = 1.303534229171897*

**

*EQM x Épocas*



*Dataset1 com a reta gerada pela regressão linear*

**Comentários**: Através do gráfico “épocas x EQM” é possível verificar que o algoritmo está “aprendendo” ? Comente.

*Sim, através do gráfico epoch x mse podemos observar que o algoritmo está aprendendo, pois há uma queda considerável no erro ao longo das épocas. No entanto, também é possível observar que grande parte do aprendizado ocorre antes da época 200, melhorando pouquíssimo após isso.*

1. **Regressão Linear Múltipla**

- Carregue os dados contidos no arquivo ex1data2.txt.

O arquivo contem 47 linhas e 3 colunas de dados. Cada coluna se refere a uma variável. Neste problema, deve-se desenvolver um modelo onde a variável da coluna 3 será estimada a partir das variáveis das coluna 1 e 2. Os dados apresentados referem-se a um problema de estimação do preço de casas. As variáveis 1 e 2 são a área da casa e o número de quartos, respectivamente. A variável 3 é o preço do imóvel.

- Implemente o algoritmo do gradiente descendente estocástico para encontrar os coeficientes da regressão.

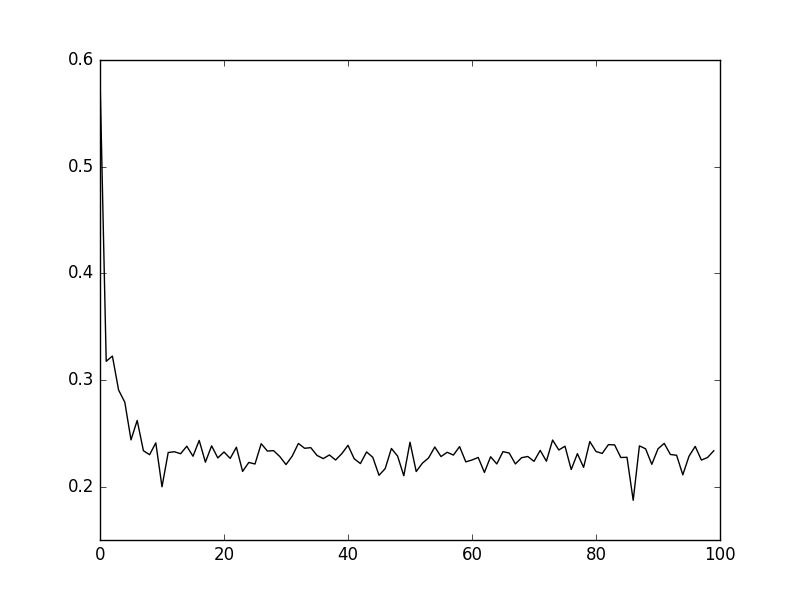
Para este algoritmo utilize α = 0.01 e utilize 100 épocas de treinamento. Para cada época de treinamento, calcule o erro quadrático médio (EQM) . Plote o gráfico “épocas x EQM”

**Apresentar**: Valor final dos coeficientes e gráfico épocas x EQM.

*W0 = 0.86476024993741019*

*W1 = 1.3966730190128032*

*W2 = -0.10865948958961531*



*EQM x Épocas*

**Comentários**: Através do gráfico “épocas x EQM” é possível verificar que o algoritmo está “aprendendo” ? Comente.

*Sim, novamente podemos perceber o aprendizado através do gráfico pela diminuição do erro quadrático médio ao longo das épocas. A oscilação que ocorre no gráfico, isso é, alguns aumentos do erro ao longo do gráfico, ocorre por conta da permutação dos dados a cada época.*

- Encontre os coeficientes da regressão utilizando o método dos mínimos quadrados.

**Apresentar**: Valor final dos coeficientes

*W0 = 0.8959791, W1 = 1.39210674, W2 = -0.08738019*

**Comentários**: Os valores obtidos pelos dois métodos são iguais? Comente.

*Não são exatamente iguais, mas são bastante próximos. A maior diferença ocorreu no valor de W2.*

1. **Regularização**

- Carregue os dados contidos no arquivo ex1data3.txt.

O arquivo contem 47 linhas e 6 colunas de dados. Cada coluna se refere a uma variável. Neste problema, deve-se desenvolver um modelo onde a variável da coluna 6 será estimada a partir das demais variáveis. Os dados apresentados referem-se a um problema de estimação do preço de casas. As variáveis 1, 2, 3, 4 e 5 são características dos imóveis e serão utilizadas como entrada do problema de regressão. A variável 6 é o preço do imóvel.

- Divida o conjunto de dados entre treino e teste. Para este problema, os primeiros 30 dados serão utilizados para treino e o restante será usado para teste.

- Encontre os coeficientes da regressão utilizando o método dos mínimos quadrados regularizado para os seguintes valores de λ = [0 1 2 3 4 5]. Utilize o conjunto de treinamento.

**Apresentar**: Valores finais dos coeficientes

*lambda 0: W = [0.79298919, 2.05650083, 0.12143939, -1.16137913, 0.06272863, -0.39110215]*

*lambda 1: W = [0.21837683, 1.3391608, 0.22506765, 0.18015951, 0.04805731, -0.24707489]*

*lambda 2: W = [0.25816829, 1.16265734, 0.23054745, 0.24910813, 0.05703977, -0.11526088]*

*lambda 3: W = [0.31875043, 1.04932735, 0.23303931, 0.25988061, 0.06804293, -0.03117398]*

*lambda 4: W = [0.38166515, 0.96594175, 0.23441173, 0.25770665, 0.07853842, 0.0254692]*

*lambda 5: W = [0.4438308, 0.90053465, 0.23490892, 0.25167386, 0.08786604, 0.06533369]*

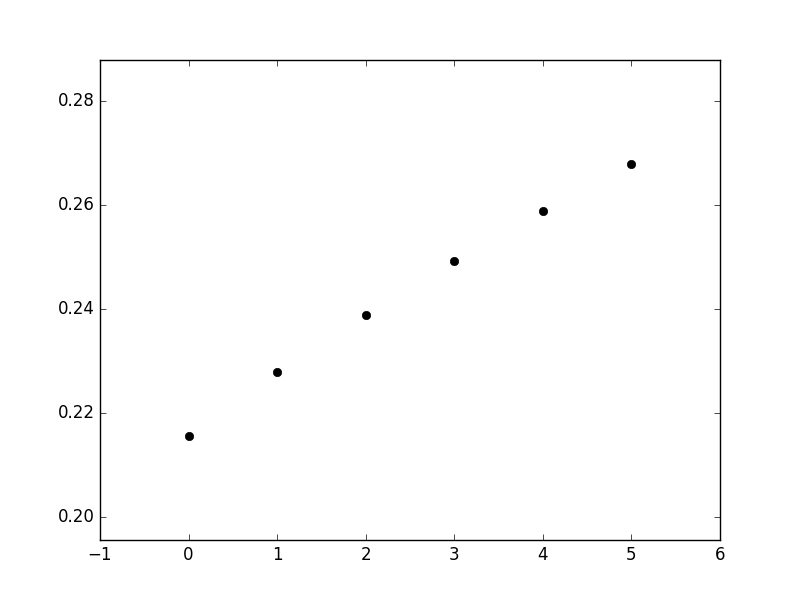
**Comentários:** Quais variáveis parecem ser menos relevantes para a regressão?

*Claramente a variável x1 é a mais relevante. Pode-se observar isso ao perceber que para todos os casos de lambda temos um peso maior que 0.9 para essa variável.*

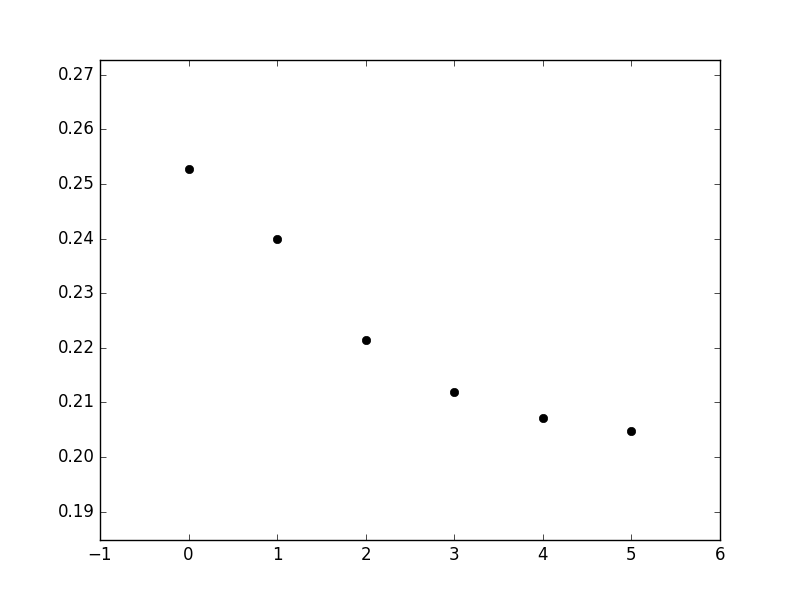
*As variáveis* ***x4*** *e* ***x5*** *são as* ***menos relevantes****, pois são as que sempre tem os pesos menores em todas os casos de lambda.*

- Encontrar o valor do EQM para os dados de treinamento e de teste para cada um dos valores de λ.

**Apresentar**: Dois gráficos. EQM x λ no conjunto de treinamento e EQM x λ no conjunto de teste



EQM x λ(treinamento)



EQM x λ(teste)

**Comentários**: Como os valores dos coeficientes variam com λ ? Explique o motivo.

Comente o crescimento/decrescimento dos erros presente nas figuras EQM x λ

*Pode-se perceber que, a medida que aumentamos o lambda, o erro quadrático médio aumenta para o conjunto de treinamento, e diminui para o conjunto de teste.*

*Isso ocorre pois, ao aumentarmos o lambda, estamos dando um peso maior aos* ***Ws*** *que ao erro na nossa função objetivo, de modo que o modelo tenderá a diminuir os coeficientes utilizados para cada variável, em especial para as variáveis que são menos relevantes para o modelo.*

*Portanto, podemos perceber um aumento no erro quadrático médio no conjunto de treinamento quando aumentamos o lambda pois estamos mais preocupados em otimizar a utilização dos coeficientes que em otimizar o erro.*

*O comportamento contrário, isso é, a diminuição do erro a medida que aumentamos o lambda, pode ser observado no conjunto de teste pois, quando damos coeficientes menores para as variáveis que menos influenciam no modelo(aumentando o lambda no treinamento), estamos buscando diminuir o overfiting e deixar o modelo mais genérico, o que podemos observar que ocorreu para esse conjunto de dados.*