TP555 - AI/ML

Lista de Exercícios #2

Regressão Linear

1. **Qual técnica de regressão linear você usaria se tivesse um conjunto de treinamento com milhares de features? Explique por quais razões você utilizaria esta técnica.**

R: A técnica de regressão linear com gradiente descendentes estocático seria á técnica mais apropriado, para um conjunto de treinamento com milhares de features. Ao contrário do algoritmo em batelada, este modelo utiliza apenas uma parte do conjunto de treinamento de forma aleatória, e essa é uma das abordagens iterativas que permite obtermos uma convergência mais rápida.

2. **Suponha que as features (i.e., atributos) do seu conjunto de treinamento tenham escalas muito diferentes. Qual técnica de regressão linear pode sofrer com isso e como? O que pode ser feito para mitigar este problema?**

R: Se as features (i.e., atributos) em um conjunto de treinamento tenham escalas muito diferentes, a técnica de regressão linear qie pode sofrer com isso é a técnica de regressão linear com gradientes descendentes e, para evitar esse problema, a variação de todos os atributos deve ser escalonada para que cada atributo contribua com a mesma importância/peso para o cálculo da distância, ou seja, do erro quadrático médio.

3. Suponha que você use o gradiente descendente em batelada e plote o erro de cada época. Se você perceber que o erro aumenta constantemente, o que provavelmente está acontecendo? Como você pode consertar isso?

R: Provavelmente pode ser um indicativo de que o passo de aprendizagem utilizado esteja demasiadamente grande, fazendo o aloritmo divergir. Para o que problema seja solucionado é necessário ajustar o passo de aprendizagem adequadamente através de uma análise do gráfico da função de custo em função do número de iterações.

4. Entre os algoritmos baseados no gradiente descendente (GD) que discutimos (batch, estocástico e mini-batch), qual deles chega mais rapidamente à vizinhança da solução ótima? Qual deles realmente converge? O que você pode fazer para que os outros também convirjam?

R: O algoritmo que chega mais rápido é o gradiente descendente estocástico.
O gradiente descendente em batelada (batch) caminha diretamente para o mínimo. Portanto, ele é o que realmente converge.
Para que os outros algoritmos também convirjam é necessário modificar o passo de aprendizagem.

5. Em sala de aula, nós discutimos 3 tipos de algoritmos baseados no gradiente descendente (batch, estocástico e mini-batch), porém, o código do mini-batch foi o único que não foi apresentado. Portanto, neste exercício eu peço que vocês
    a. Implementem o algoritmo do mini-batch
    b. Testem sua implementação com y = 2*x1 + 2*x2 + w, onde x1, x2 e w são M = 1000 valores retirados de uma distribuição aleatória Gaussiana normal padrão (i.e, com média 0 e variância igual a 1) e utilizando a função hipótese h = a1*x1 + a2*x2,
    c. Plotem a superfície de erro, a superfície de contorno com os parâmetros a1 e a2 para cada iteração do mini-batch, e o gráfico de iteração versus erro,
    d. Encontrem o valor ótimo do passo de aprendizagem (**Dica** : utilizem os gráficos da superfície de contorno com os parâmetros a1 e a2 para cada iteração do mini-batch e o gráfico de iteração versus erro para saber se aquele passo é o ótimo. Acessem os links abaixo para entender como vocês podem plotar os gráficos de contorno.),
        - https://matplotlib.org/3.1.1/gallery/images_contours_and_fields/contour_demo.html#sphx-glr-gallery-images-contours-and-fields-contour-demo-py
        - https://www.python-course.eu/matplotlib_contour_plot.php
    e. Comparem os resultados do mini-batch com os resultados obtidos com o GD em batelada (batch) e GD estocástico (**Dica** : para a comparação, usem os códigos que estão nos slides da aula e plotem os gráficos da superfície de contorno com os parâmetros a1 e a2 para cada iteração e o gráfico de iteração versus o erro para GD em batelada e estocástico).
    f. Baseando-se nos gráficos do item anterior, a que conclusões vocês podem chegar quanto ao treinamento dos 3 tipos de gradiente descendente?
6. Dada a seguinte função hipótese e assumindo o erro quadrático médio como função de erro

$$h = a0 + a1*x + a2*x^2.$$

Encontre as equações de atualização dos pesos/parâmetros para esta função. Em seguida, utilizando os vetores x e y definidos abaixo, encontre os parâmetros a0, a1 e a2

através do método da regressão de forma fechada e com gradiente descendente em batelada. y = 3 + 1.5*x + 2.3*x^2 + w,
onde x é um vetor coluna com M = 1000 valores retirados de uma distribuição aleatória uniformemente distribuída no intervalo de -5 a 5 e w é outro vetor coluna com M valores retirados de uma distribuição aleatória Gaussiana com média 0 e variância igual a 10.

a. Plote o gráfico do número de iterações versus o erro.
b. Baseado no gráfico acima, encontre o melhor valor para o passo de aprendizagem.

7. Neste exercício você vai utilizar o arquivo ***training.csv*** onde a primeira coluna são os valores de x (feature) e a segunda de y (label). Baixe o arquivo do endereço: training.csv. Após, leia o conteúdo do arquivo, ou seja, os vetores x e y, com os seguintes comandos:

```
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('training.csv', header=None)

x = df[0].to_numpy()
y = df[1].to_numpy()

fig = plt.figure(figsize=(10,10))
plt.plot(x, y, 'b.')
```

Em seguida, utilize o algoritmo do ***gradiente descendente em batelada*** para encontrar os parâmetros de cada uma das seguintes funções hipóteses.

a. h = a0 + a1*x
b. h = a0 + a1*x + a2*x^2
c. h = a0 + a1*x + a2*x^2 + a3*x^3
d. h = a0 + a1*x + a2*x^2 + a3*x^3 + a4*x^4

Para cada uma das funções hipótese acima faça o seguinte:

a. Encontre os valores ótimos dos parâmetros através do método de forma fechada, i.e., equação normal, ou também conhecida como método dos mínimos quadrados.
b. Encontre as equações de atualização dos parâmetros assumindo o erro quadrático médio como função de erro.
c. Encontre o valor ótimo do passo de aprendizagem.
d. Plote um gráfico que mostre x vs. y e x vs. h, ou seja, um gráfico comparando os dados originais com a estimativa (i.e., hipótese) da função que gerou y.
e. Plote um gráfico com do número de iterações versus o erro.

Em seguida responda às seguintes perguntas

A. Qual das funções hipótese acima aproxima melhor a função alvo (target), ou seja, qual produz o menor erro ao final do treinamento?

B. Dado que você encontrou os parâmetros que otimizam cada uma das funções hipótese acima (ou seja, você agora tem um modelo treinado que pode predizer o resultado para novos exemplos), use os dados contidos no arquivo predicting.csv e calcule o erro quadrático médio para cada um dos modelos (i.e., função hipótese). Qual função hipótese resulta no menor erro quadrático médio? O que você consegue concluir a respeito deste resultado?

8. Neste exercício você irá aplicar escalonamento de features aos dados de treinamento e teste. Dada a seguinte função objetivo y = x1 + x2,
onde x1 é um vetor coluna com M amostras retiradas de uma distribuição Gaussiana com média 0 e desvio padrão unitário e x2 é um vetor coluna com M amostras retiradas de uma distribuição Gaussiana com média 10 e desvio padrão igual a 10. Gere dois conjuntos de dados, com M = 1000 amostras cada. Um conjunto será utilizado para treinamento e o outro para teste, ou seja, validação do modelo treinado. Utilize o gradiente descendente em batelada com a seguinte função hipótese h = a1*x1 + a2*x2,
com a1 e a2 iniciais iguais a -20 e -20, respectivamente. Para todos os casos abaixo, treine os modelos com o mesmo número máximo de iterações, por exemplo, 2000 iterações e um critério de parada que faça o algoritmo parar quando o erro entre duas épocas consecutivas for menor do que 0.001, ou seja, o algoritmo irá parar se o erro for menor do 0.001 ou se atingir o número máximo de iterações. Pede-se
   a. Sem aplicar nenhum escalonamento de features aos exemplos de treinamento, plote a superfície de erro, a superfície de contorno com os parâmetros a1 e a2 encontrados durante as iterações (ou seja, o histórico de valores que o algoritmo encontra durante o treinamento do modelo) e o gráfico de erro quadrático médio versus o número de iterações para os conjuntos de treinamento e teste. **OBS.1** : Não se esqueça de encontrar o valor ótimo para o passo de aprendizagem. **OBS.2**: Não se esqueça de encontrar o valor ótimo dos pesos/parâmetros e plotá-los no gráfico de contorno com o histórico dos pesos.
   b. Aplique a normalização min-máx às features de treinamento e teste, plote a superfície de erro, a superfície de contorno com os parâmetros a1 e a2 encontrados durante as iterações e o gráfico de erro quadrático médio versus o número de iterações para os conjuntos de treinamento e teste. **OBS.1** : Não se esqueça de encontrar o valor ótimo para o passo de aprendizagem. **OBS.2** : Não se esqueça que o conjunto de testes é normalizado com os valores mín-máx encontrados para o conjunto de treinamento. **OBS.3** : Não se esqueça de encontrar o valor ótimo dos pesos/parâmetros e plotá-los no gráfico de contorno com o histórico dos pesos.
   c. Aplique a padronização às features de treinamento e teste, plote a superfície de erro, a superfície de contorno com os parâmetros a1 e a2 encontrados durante as iterações e o gráfico de erro quadrático médio versus o número de iterações para os conjuntos de treinamento e teste. **OBS.1** : Não se esqueça de encontrar o valor ótimo para o passo de aprendizagem. **OBS.2** : Não se esqueça que o conjunto de

testes é padronizado com os valores de padronização encontrados para o conjunto de treinamento. **OBS.3** : Não se esqueça de encontrar o valor ótimo dos pesos/parâmetros e plotá-los no gráfico de contorno com o histórico dos pesos.

d. Repita os itens b e c aplicando desta vez a normalização min-máx e a padronização, respectivamente, também aos targets/rótulos (ou seja, os valores de y).
e. Baseado nos resultados anteriores o que você pode concluir a respeito do escalonamento de features? (**Dica** : Comente a respeito das formas das superfícies de erro, dos números de iterações necessárias para se alcançar o ponto ótimo, isso se ele é alcançado, da diferença entre o erro quadrático médio obtido para o conjunto de treinamento e o obtido para o conjunto de testes (são similares ou diferentes), da diferença entre os valores do erro quadrático médio para os 3 casos acima, i.e., sem escalonamento e com os 2 tipos de escalonamento com e sem escalonamento dos labels (qual resulta no menor erro? Escalonar os labels traz algum benefício? Como ficam as superfícies de erro quando se escalona os labels?), e o que mais você achar interessante comentar. Quanto mais detalhada sua análise dos resultados, melhor será sua avaliação neste exercício.)