Relatorio da Atividade #1

Barbara Caroline Benato
RA 192865
barbarabenato@gmail.com

Breno Leite
RA 192863
brenolleite@gmail.com

I. Introdução

A primeira tarefa da disciplina visava explorar a técnica de regressão linear, a fim de encontrar o melhor modelo possível para um determinado problema evitando o overfitting do modelo, ou seja, que o modelo seja super treinado para o conjunto de dados disponível de tal forma que não seja capaz de predizer para outros dados. Para tal tarefa, optou-se por predizer o ano de lançamento de de música através de características específicas de áudio. Assim, vários modelos foram analisados utilizando uma validação cruzada para melhor análise dos dados. Uma abordagem de regressão linear utilizando a Equação Normal foi comparada com a utilizando o Gradiente Descendente.

II. Atividades

1. Perform Linear Regression (LR) as the baseline (first solution) and devise LR-based alternative (more powerful) solutions. 2. Use the specified training/test data for providing your results and avoid overfitting. 3. Devise and test more complex models. 4. Plot the cost function vs. number of iterations in the training set and analyze the model complexity. What are the conclusions? What are the actions after such analyses? 5. Use different Gradient Descent (GD) learning rates when optimizing. Compare the GD-based solutions with Normal Equations if possible (perhaps you should try with smaller sample sizes for this task). What are the conclusions?

III. Materiais e Métodos

base: subset of the Million Song Dataset (descrever) explicar cada um deles e parametros utilizados sklearn: - pre processing: scale, normalize, pca - metodos: LinearRegression e SGDRegressor

- base: subset of the Million Song Dataset (descrever) - treino validação (validação cruzada) e teste - metricas - graficos: learning_curve e plot

IV. Experiments and Discussion

Como a solução foi construída a partir de um processo gradativo, optou-se por apresentar o desenvolvimento de tal processo.

(Apresentar os dois graficos pra cada processo abaixo)

¿¿ LinearRegression
— Pre processamento
- Modelo sem nenhuma alteraç ao: resultado
- Modelo com scale
- Modelo com normalize
- Modelo scale + normalize
— Feature Selection
- Modelo com normalize + pca
- Modelo buscando features: escolhe primeiras 10, primeiras 12, primeiras 13, primeiras 50, ultimas 78, ultimas 80
—Definir modelo
—Analisar overfitting do modelo
— validaç ao cruzada: 3 / 5 / 10 folds
- iteraç oes: 50

¿¿ SGDRegressor
— Pre processamento
- Modelo sem nenhuma alteração: resultado
- Modelo com scale
- Modelo com normalize
- Modelo scale + normalize
— Feature Selection
- Modelo com normalize + pca
- Modelo buscando features: escolhe primeiras 10, primeiras 12, primeiras 13, primeiras 50, ultimas 78, ultimas 80
— Definir modelo
— Analisar overfitting do modelo
— validação cruzada: 3 / 5 / 10 folds
- iterações: 50
valores de alpha: 0.1/0.01/0.001/0.0001 + (?)

V. Conclusions and Future Work

The main conclusions of the work as well as some future directions for other people interested in continuing this work.

References

[1] Christopher M. Bishop. “Pattern Recognition and Machine Learning”. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.