

Aprendizado de máquina para reconhecimento de padrões

Carlos Augusto Chiarelli

Geral

- HYPE (assunto do momento)
- os dados não são valiosos por si só, necessário extrair informação
- Novo petróleo
- Ao resolver um problema de dados, começo da maneira mais simples e vou aumentando a complexidade

Regressão linear

- modelo simples e robusto
- mais adequada para prever valores contínuos
- quanto mais dados, melhor a estimativa
- todo MODELO em ML possui uma **função de custo**
- sempre buscamos minimizar a função de custo
- quanto menor o valor dessa função, menor o erro do modelo
- quando encontro o mínimo da função de custo, o modelo para de aprender (encerra o treinamento)

Regressão linear

- funções de custo podem ter um único mínimo ou vários mínimos
- o aprendizado do modelo é guardado em variáveis que chamamos de parâmetros ou pesos
- a taxa de aprendizado serve pra dizer se que encontrar o mínimo da função de custo rápido ou devagar
- o custo é minimizado com a **DESCIDA DO GRADIENTE**
- descida do gradiente é semelhante para todos modelos de ML

Descida do gradiente

Realiza de 3 maneiras:

- Batch : olho para todos os exemplos (linhas da tabela) para atualizar os dados
- Estocástico : uso um exemplo para fazer a atualização nos parâmetros
- mini-batch : pego um conjunto de dados para atualizar os parâmetros (**MAIS USADO**)

Regressão logística

- maneira simples e robusta de resolver uma classificação
- gera uma fronteira de decisão (para classificar)
- função de custo praticamente idêntica da regressão linear
- multiclassificação (mais de uma classe): cria-se um classificador para cada classe e escolhe o que retorna maior probabilidade (tiver mais certeza do resultado)

Regularização

- se refere a qualquer problema pra evitar overfitting (decorar dados treino)
- evitar com que o modelo decore o conjunto de dados de treino e vá mal no conjunto de dados de teste
- consiste geralmente em adicionar variáveis à função de custo impedindo que ela zere, se ela zerar então o modelo acerta todo treino (decorou) e irá mal no teste
- proponho um modelo e penalizo ele com um termo adicional (hiperparâmetro)

Métricas de teste e validação

- **Classificação**

- acurácia
- recall
- f1
- matriz de confusão

- **Regressão**

- R^2
- Erro absoluto (MAE)
- Erro quadrático médio (MSE)