Aprendizagem Supervisionada Aula 3

Pré-processamento de Dados, RNA e Classificação

Bernard da Silva
Orientador: Rafael Parpinelli
23/11/2022

UDESC
UNIVERSIDADE DO ESTADO DE SANTA CATARINA

Aula Passada

- Tipo de dados (numéricos, categóricos);
- Inspeção visual de dados;
- Medidas estatísticas;
- Escolha de medidas;
- Exemplos de visualização de dados univariados, bivariados e multivariados;
- Limpeza de dados.

Objetivo dessa aula

- Continuação “noção de como pré-processar dados”
- Dicas para seleção de algoritmos
- Conceito RNA

Feature Engineering

Consiste em criar novas variáveis de entradas utilizando variáveis já presentes no banco de dados. São divididas em diferentes categorias:

- Decomposição / Splitting
- Cruzamento / Crossing
- Reinterpretação / Reframing
- Discretização / Binning

Feature Engineering

- Decomposição / Splitting
  - Consiste em “separar” uma variável em múltiplas partes
    - Ex: Decompõe Data = 2014-09-20T20:45:40Z em Ano = 2014, Mês = 09, Dia = 20
    - Peso = 6.280 -> Quilo = 6, Gramas = 0.280

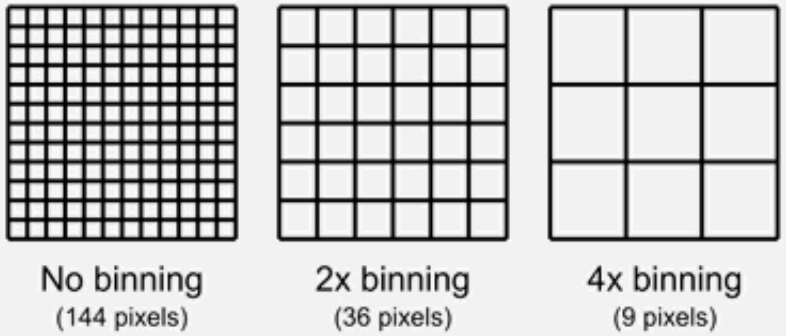
- Cruzamento / Crossing
  - Combina diferentes variáveis em uma nova variável
    - Ex: Multiplicação ou divisão de uma variável por outra variável.

  Qual motivo de se fazer Crossing?

Feature Engineering

- Reinterpretação / Reframing
  - Consiste em “reinterpretar” um dado
    - Ex: trocar unidades de peso
      - Transformar kg em g, milhas em quilômetros, etc.

- Discretização / Binning
  - Consiste em agrupar valores contínuos em categorias ou intervalos maiores



Feature Selection e Redução de Dimensionalidade

- Seleção de Features e Redução de Dimensionalidade são o processo de REDUZIR o número de variáveis de entrada para uma modelagem com os objetivos de:
  - Remover variáveis de entrada redundantes
  - Reduzir o custo computacional
  - Reduzir a complexidade do modelo
  - Aumentar a interpretabilidade do modelo
  - Aumentar a acurácia do modelo, se possível

- Seleção de Features pode ser divididos em 2 tipos principais:
  - Abordagem Wrapper
  - Abordagem de Filtros

Feature Selection

- Abordagem Wrapper

  - Treina vários modelos com diferentes subconjuntos de features e seleciona o melhor
  - Características:
    - Computacionalmente exigente;
    - O resultado é dependente do modelo (as melhores features de uma Rede Neural pode ser diferente de uma Random Forest);
    - Se tiver poucos dados, aumenta o risco de overfit,.
    - Resultados tendem a ser melhores do que abordagem de filtro.

  - Técnicas:
    - Forward Selection, Recursive Feature Elimination (RFE), Algoritmos Genéticos, etc.

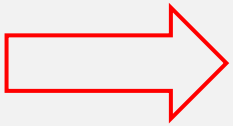
Feature Selection

- Abordagem de Filtros
  - Calcula métricas estatísticas das variáveis de entrada e elimina aquelas que forem menos interessantes;
  - Características:
    - Computacionalmente barato
    - Robusto contra overfit
  - Técnicas:
    - Análise de correlação, Teste Chi-Quadrado, ANOVA: Análise de Variância, etc

Feature Encoding

- Forma mais simples de realizar Feature Encoding;
- Consiste em substituir um dado categórico por um dado numérico.

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000



Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

- Quando usar?
  - Variáveis categóricas são ordinais
  - Número de categorias é muito grande

Qual algoritmo devo utilizar?

DEPENDE!

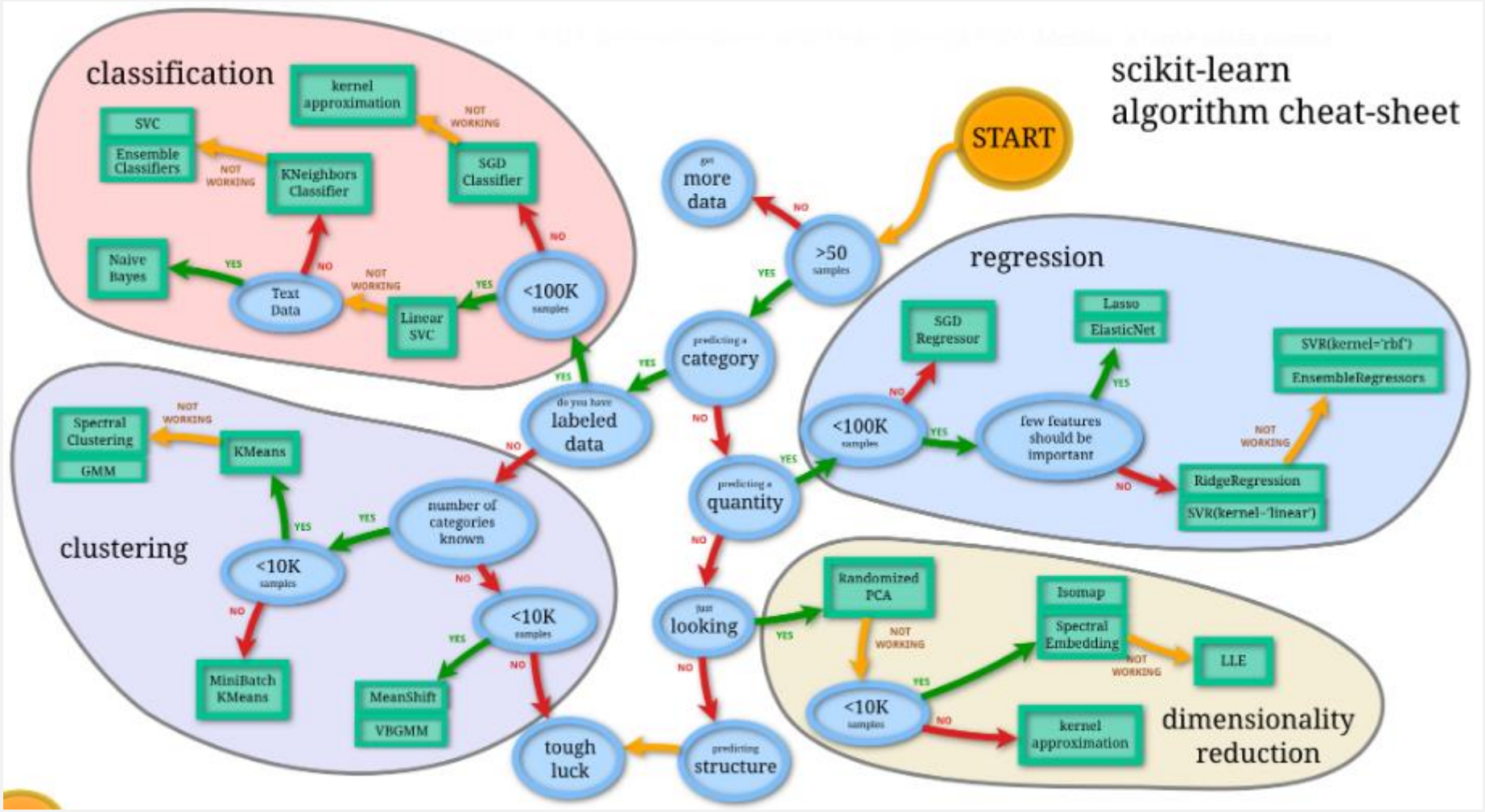
- O tamanho, a qualidade, a natureza dos dados
- O tempo computacional disponível
- A urgência da tarefa
- O que você quer fazer com os dados

Mesmo um cientista de dados experiente não pode dizer qual algoritmo terá o melhor desempenho antes de tentar algoritmos diferentes.

MASS

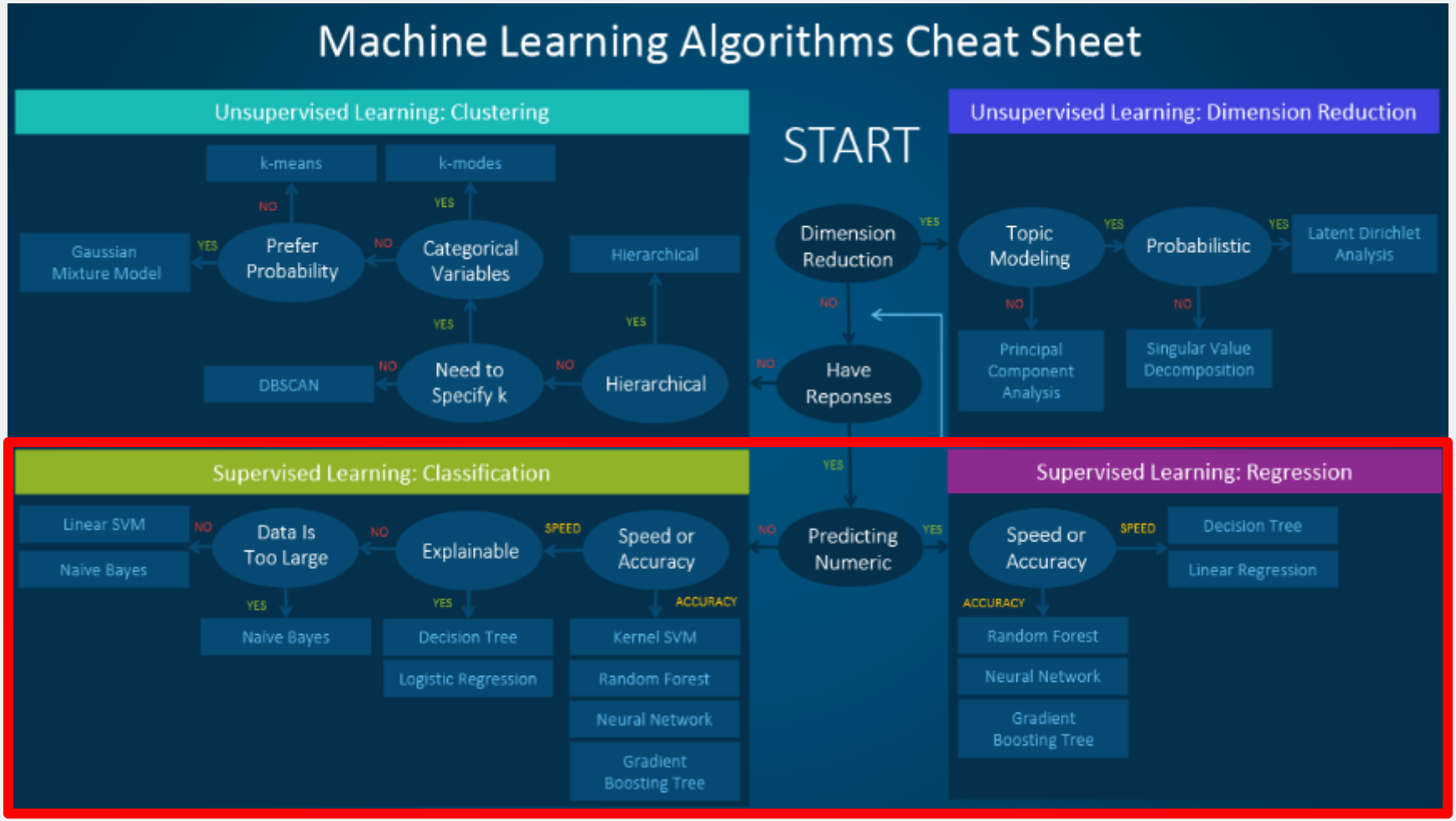
É possível fornecer algumas orientações sobre quais algoritmos tentar primeiro, dependendo de alguns fatores claros.

Folha de Consulta



UDESC

Folha de Consulta

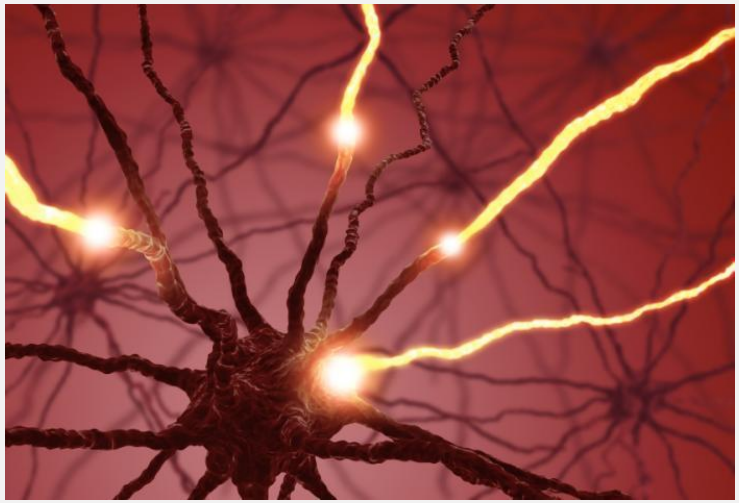


UDESC

RNA

De onde vem a inspiração da RNA?

- Inspirada na biologia
- Já era teorizado por volta de 1700 que a eletricidade era a força motriz do cérebro
- Em 1943, Warren McCulloch (neurofisiologista) e Walter Pitts (matemático) modelaram neurônios usando circuitos elétricos.
- Em 1949, Donald Hebb descobriu o efeito do reforço em neurônios.
- Em 1959 a primeira RNA foi modelada, em Stanford.



UDESC

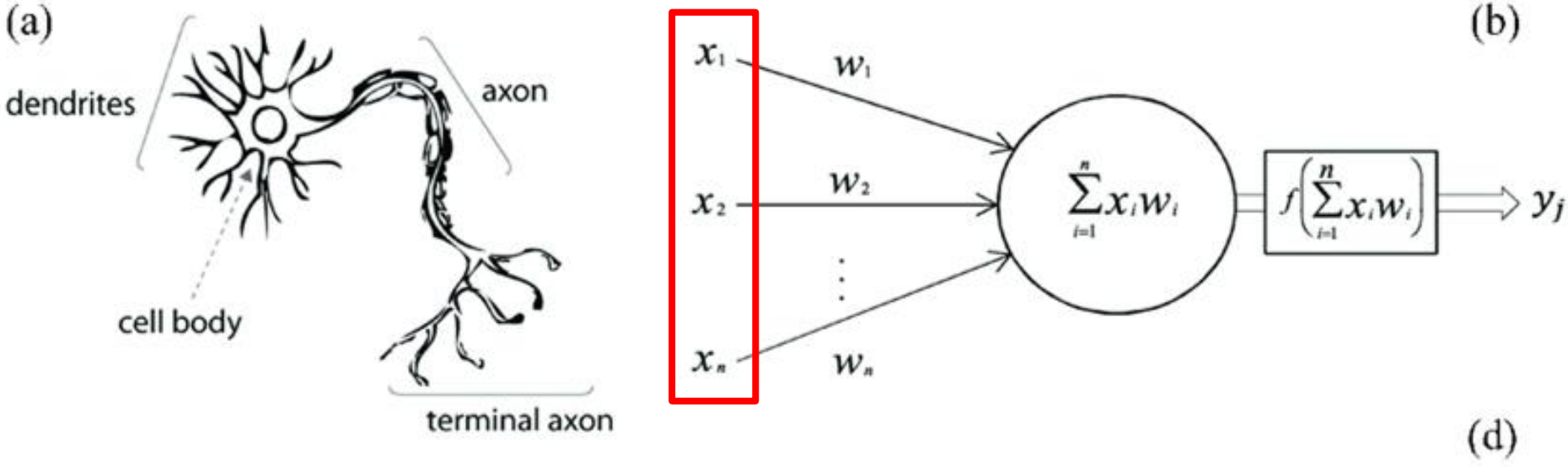
RNA

De onde vem a inspiração da RNA?

- Modelagem conexionista
- Qualquer estado mental pode ser descrito como um conjunto de ativações numéricas de unidades neurais em uma rede
- Existe uma etapa de aprendizagem (modelagem)
- A memória é criada pela modificação da conexão entre neurônios.
- É capaz de se auto organizar
- Apresenta boa flexibilidade

A modelagem de um neurônio

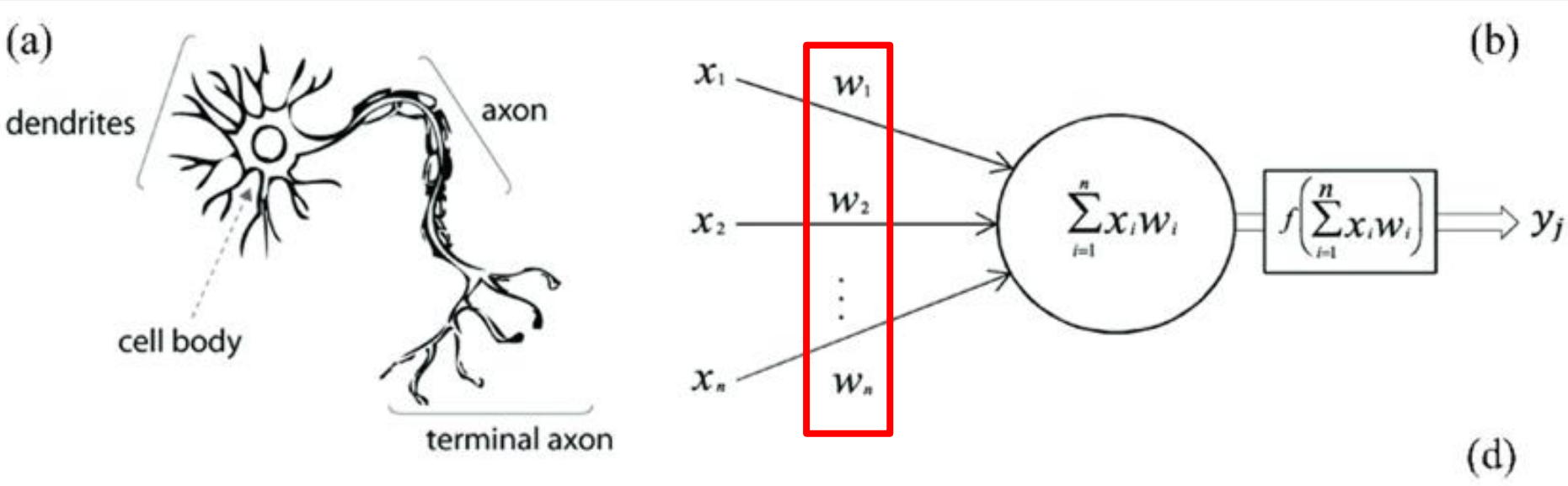
- Um neurônio tem N entradas



UDESC

A modelagem de um neurônio

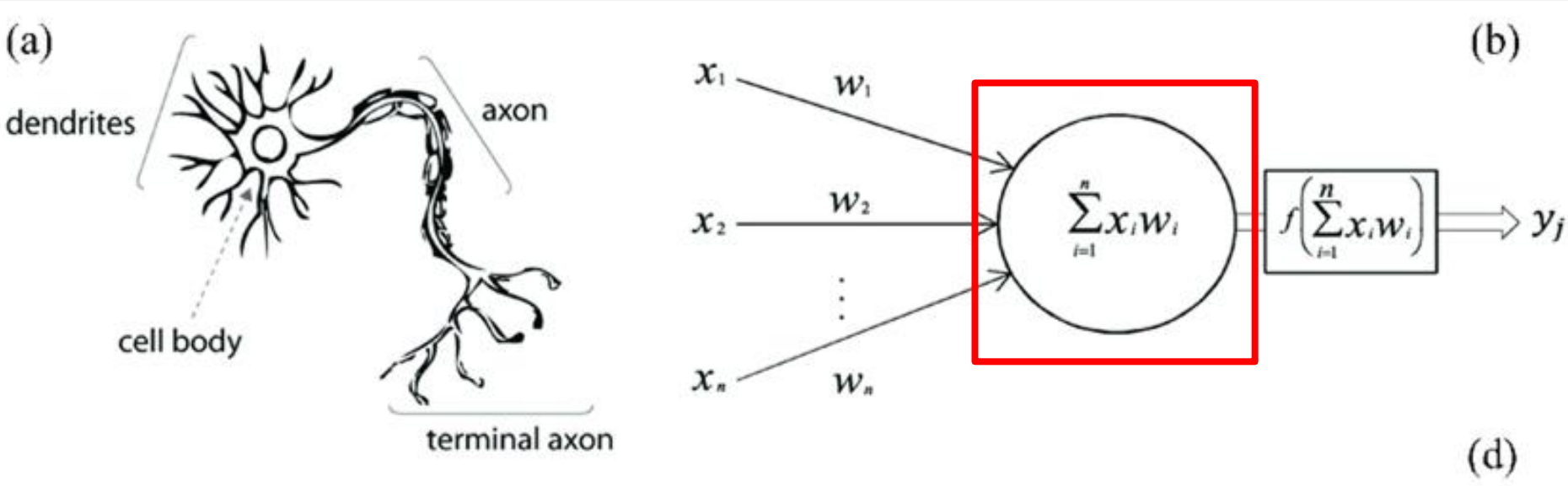
- Um neurônio tem N entradas
- Cada entrada tem um peso (weight)



UDESC

A modelagem de um neurônio

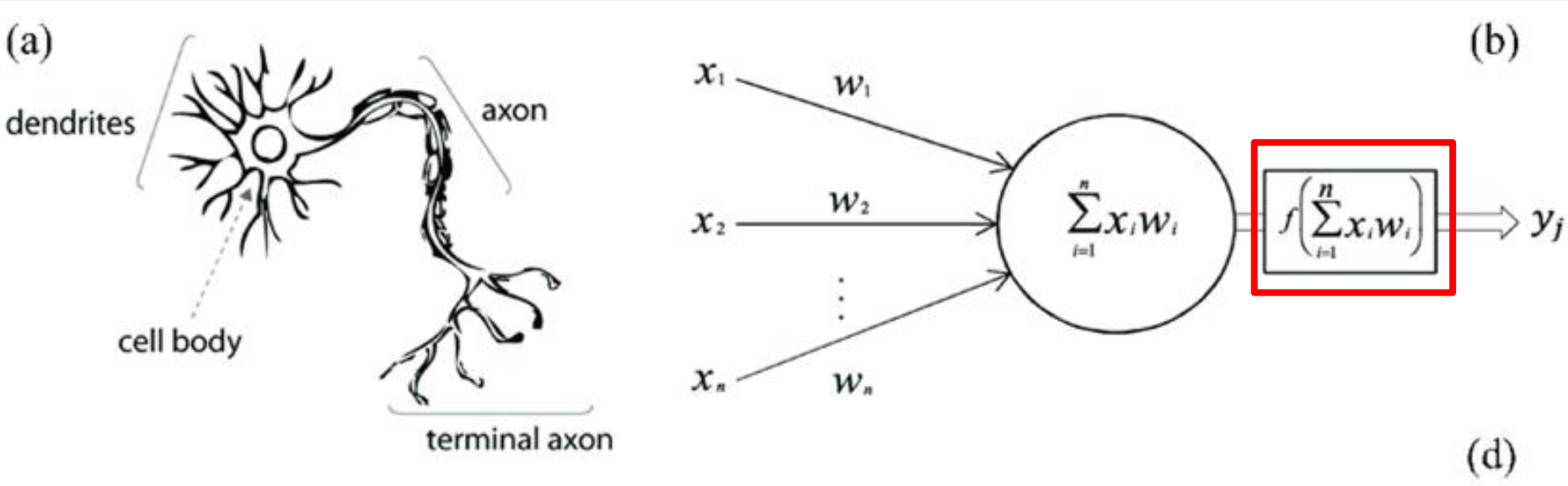
- Um neurônio tem N entradas
- Cada entrada tem um peso (weight)
- O neurônio soma a multiplicação das entradas com os pesos



UDESC

A modelagem de um neurônio

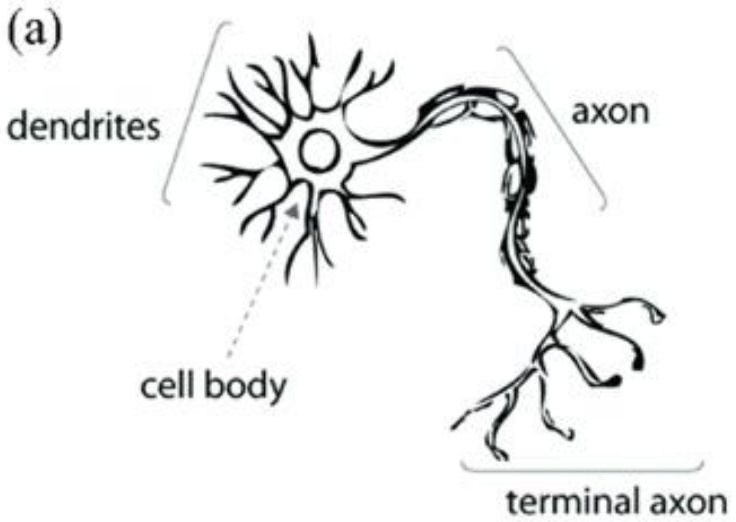
- Um neurônio tem N entradas
- Cada entrada tem um peso (weight)
- O neurônio soma a multiplicação das entradas com os pesos
- Esse somatório é aplicado em uma função (função de ativação)

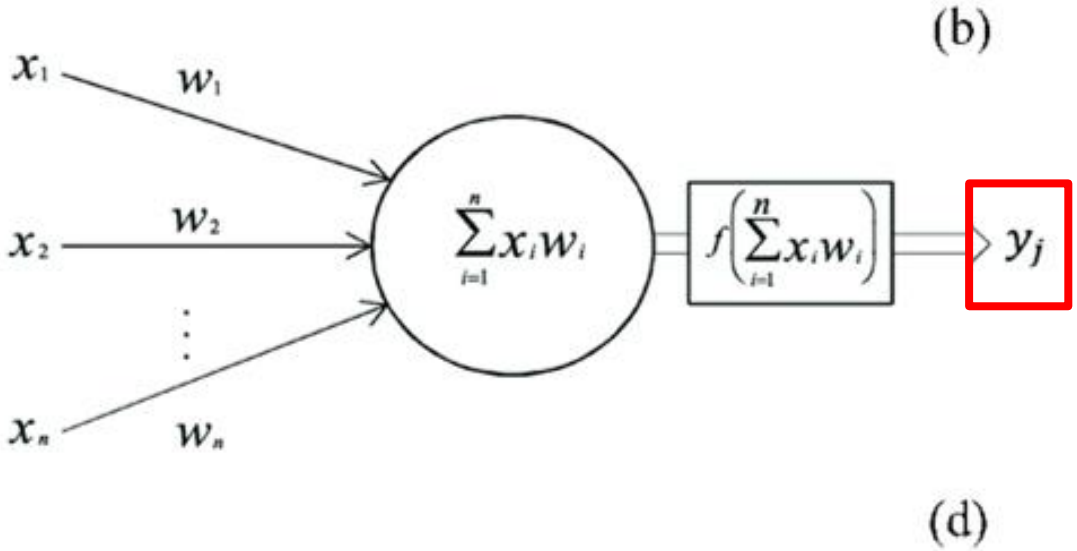


UDESC

A modelagem de um neurônio

- Um neurônio tem N entradas
- Cada entrada tem um peso (weight)
- O neurônio soma a multiplicação das entradas com os pesos
- Esse somatório é aplicado em uma função (função de ativação)
- O neurônio gera uma saída

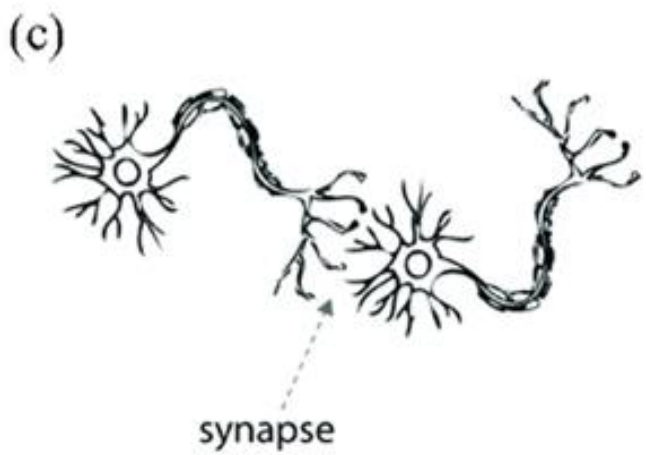


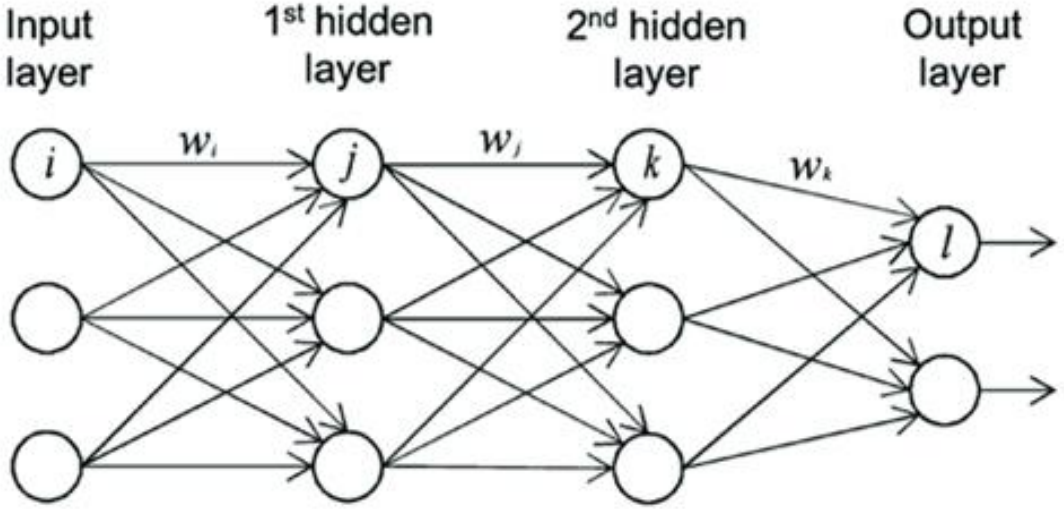


UDESC

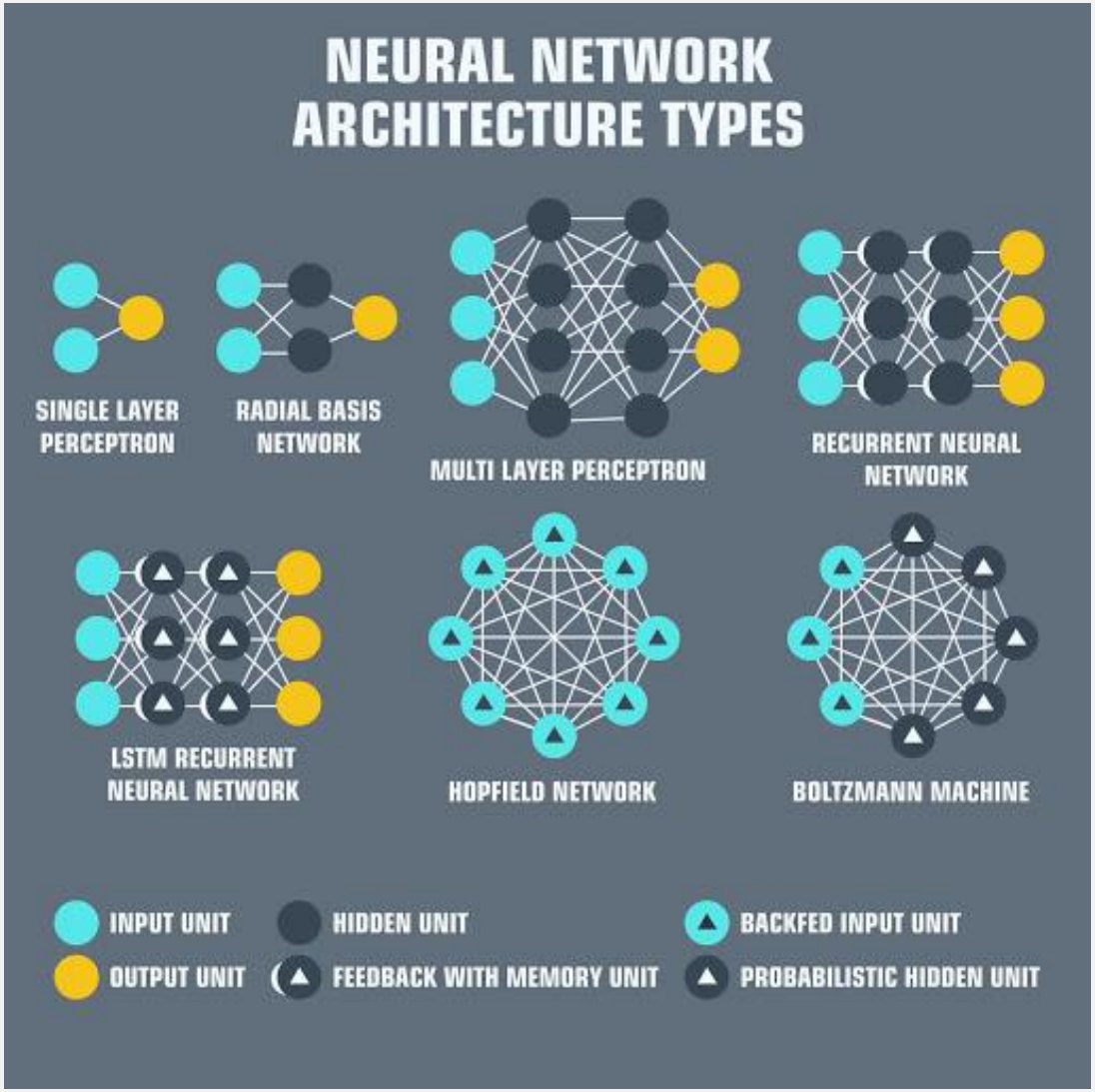
A modelagem de uma rede

- Cada neurônio recebe as saídas dos neurônios da camada anterior
- Cada neurônio envia a sua saída para os neurônios da próxima camada
- Através de um algoritmo de otimização/aprendizagem, os pesos dos neurônios são ajustados.



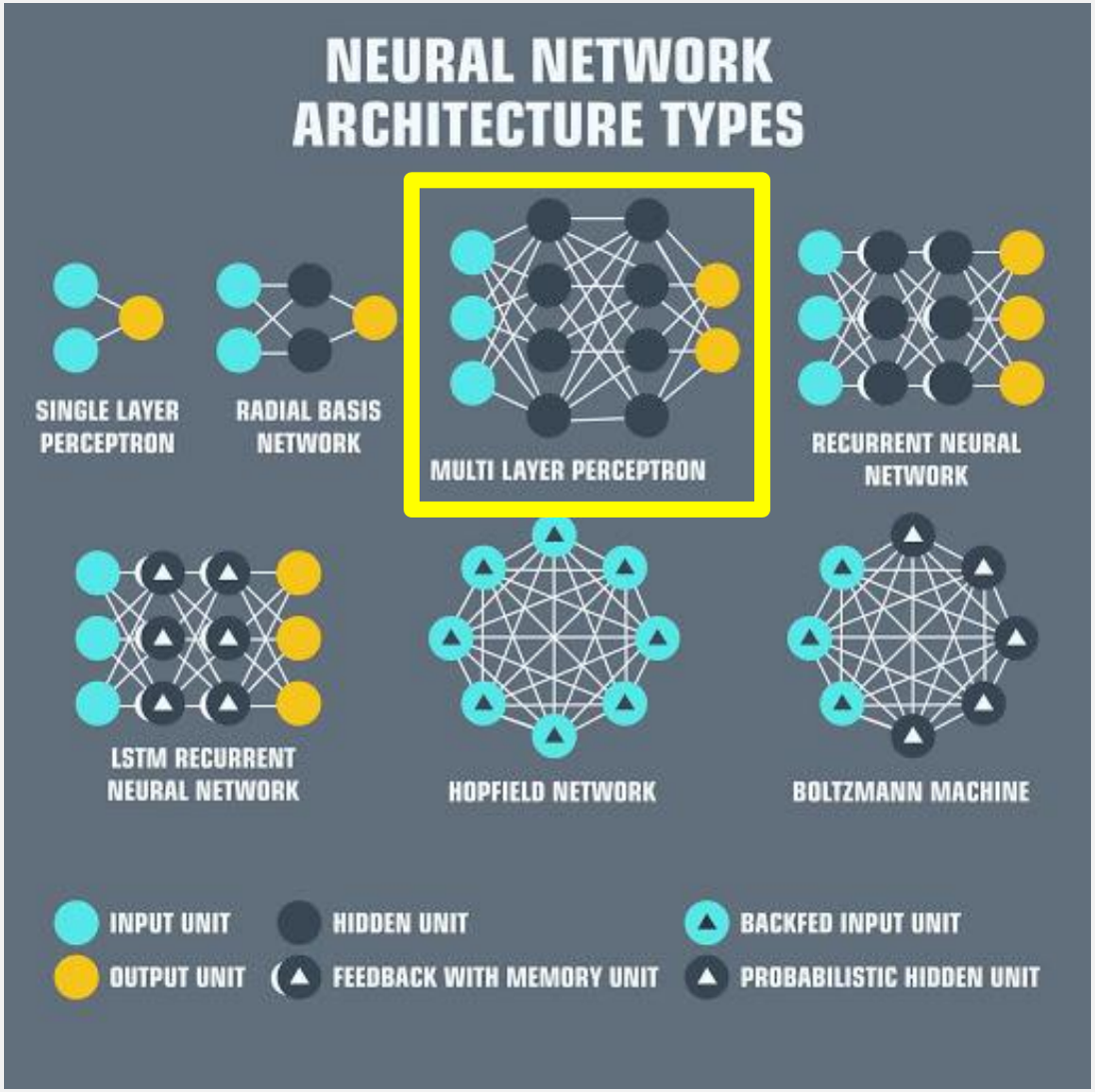


Tipos de redes neurais



UDESC

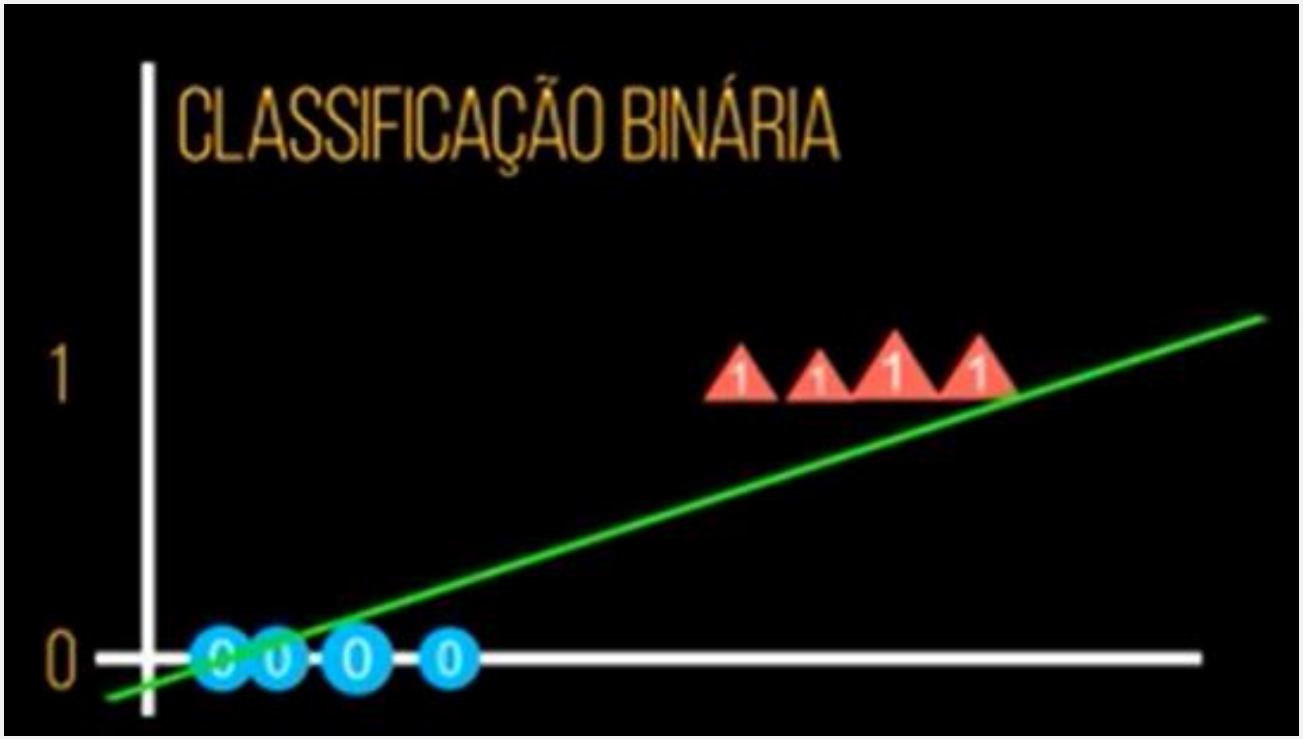
Tipos de redes neurais



UDESC

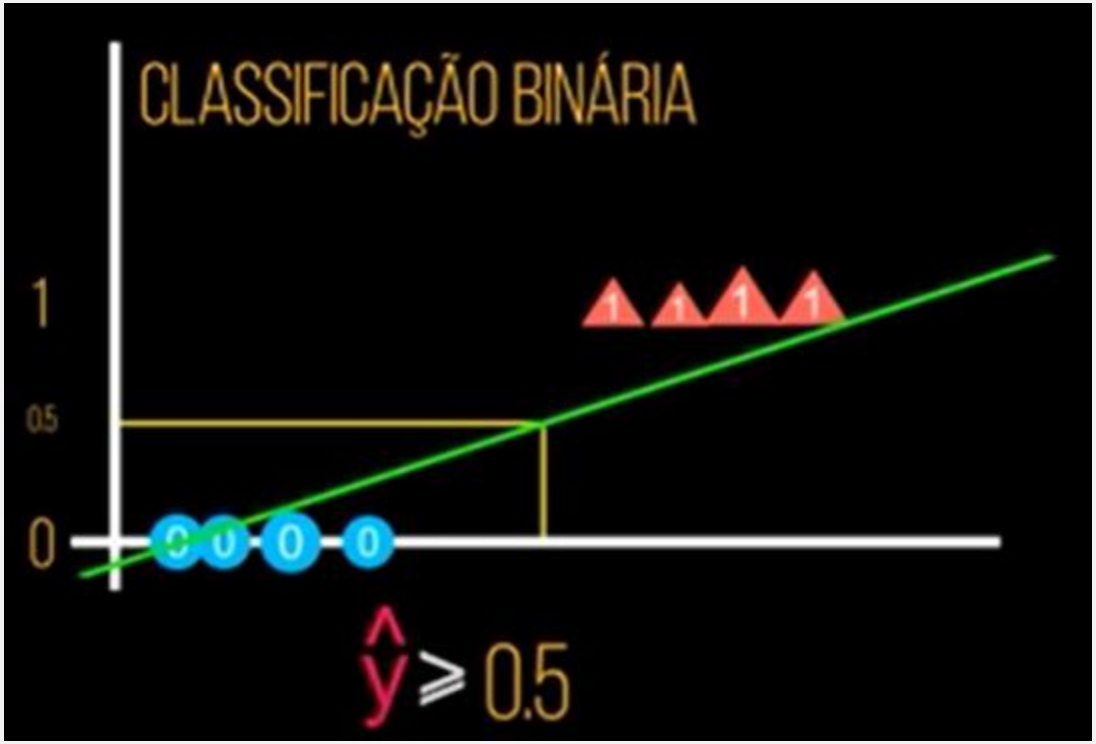
Classificação - Logistic Regression

- Como determinar uma função que determina a probabilidade de um dado?



UDESC

Classificação - Logistic Regression

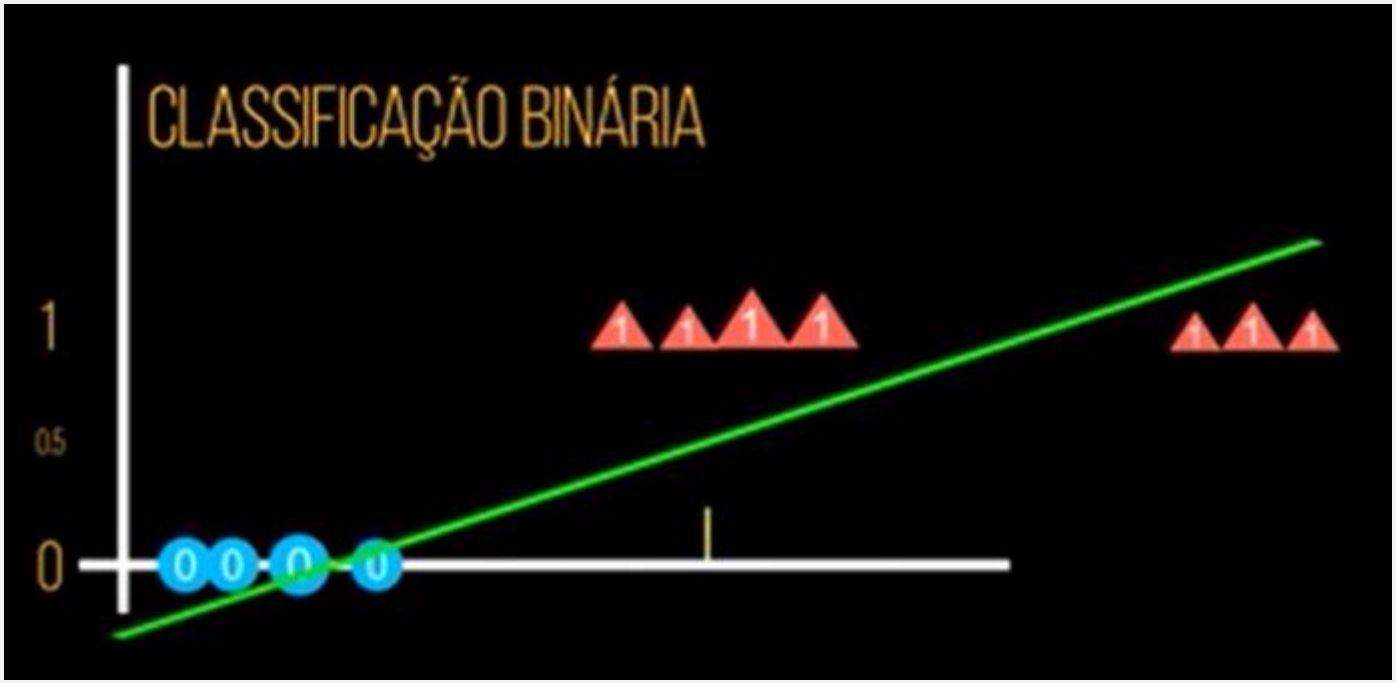


Onde está o problema?

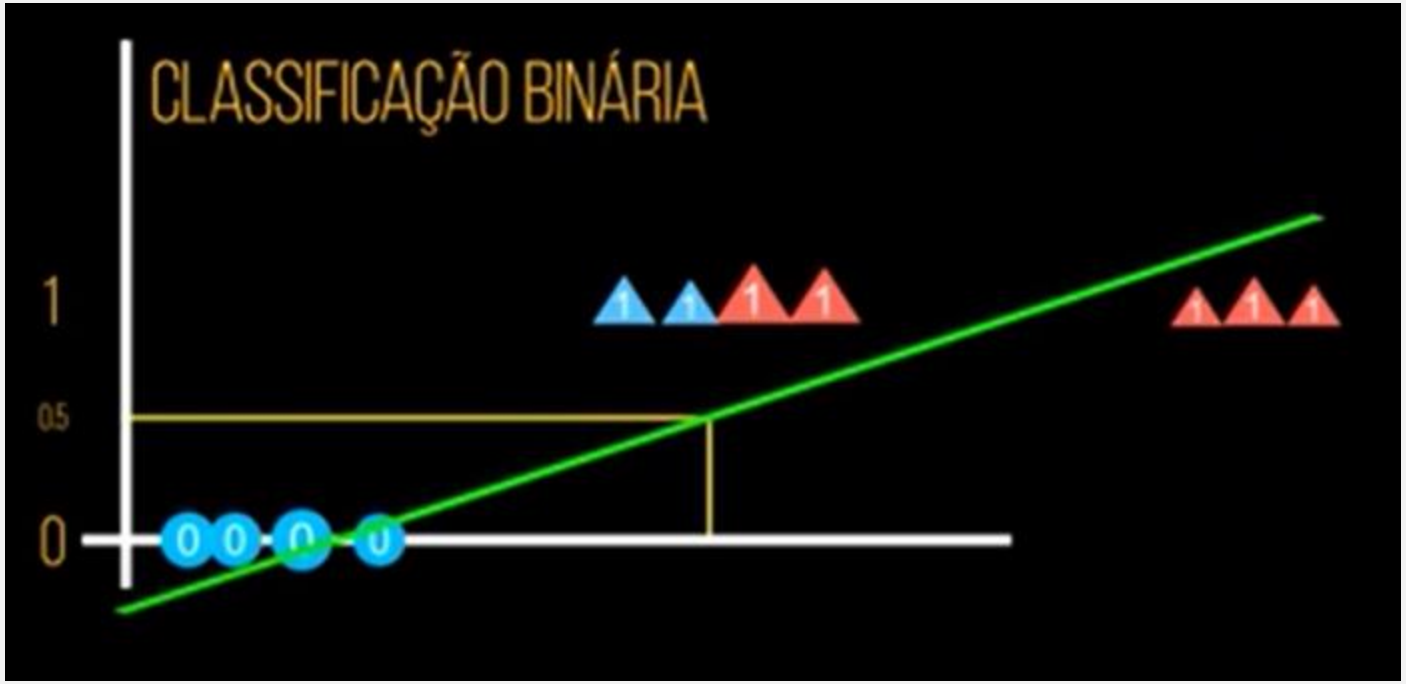
UDESC

Classificação - Logistic Regression

O que aconteceria?



Classificação - Logistic Regression



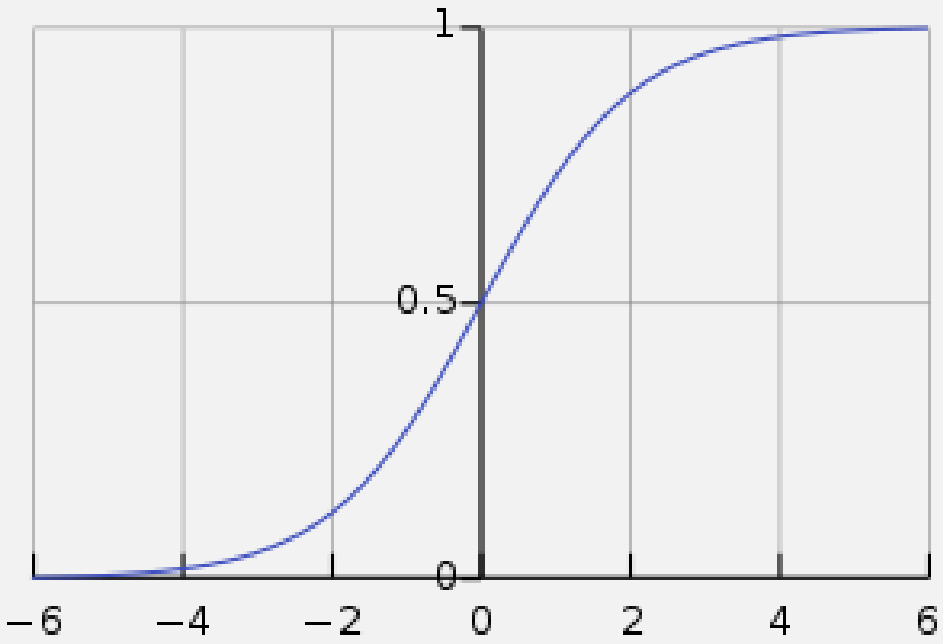
O que pode ser feito para melhorar a classificação?

Classificação - Logistic Regression

Função de ativação (Sigmóide)

$$g(z) = \frac{1}{1 + e^{-z}}$$

$$z = w^T x$$



UDESC

Classificação - Logistic Regression

Sigmóide (Python)

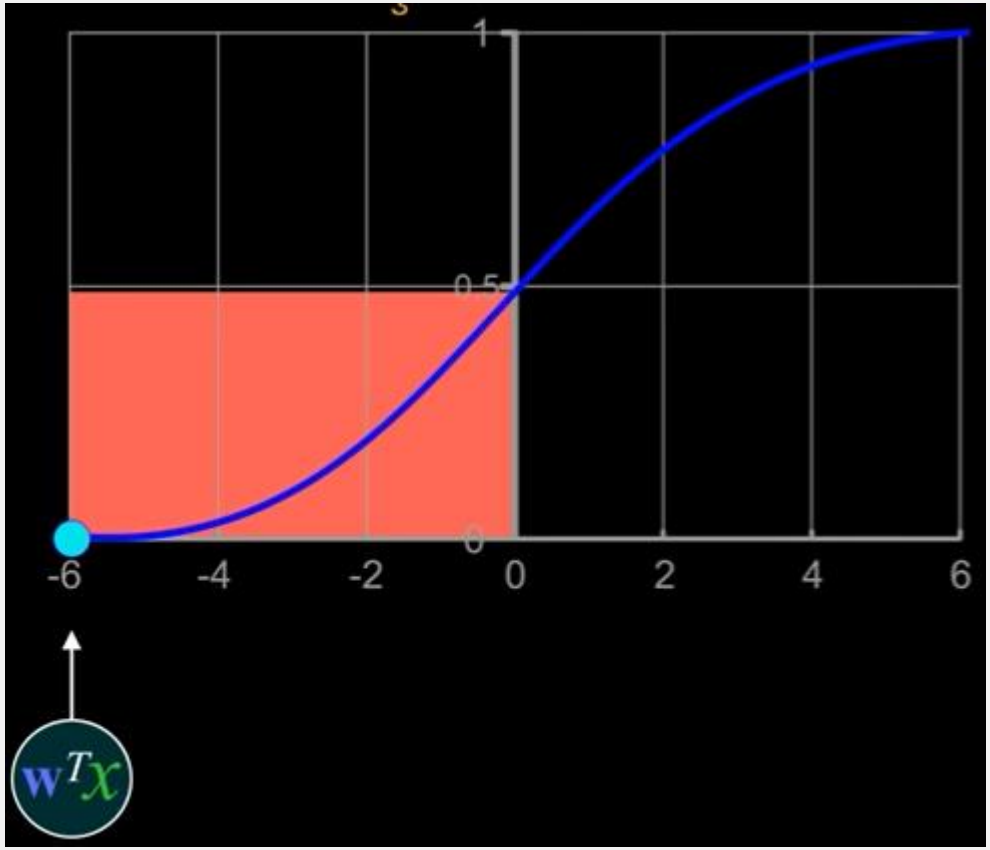
$$g(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

Sigmoid retornará a probabilidade de cada classe

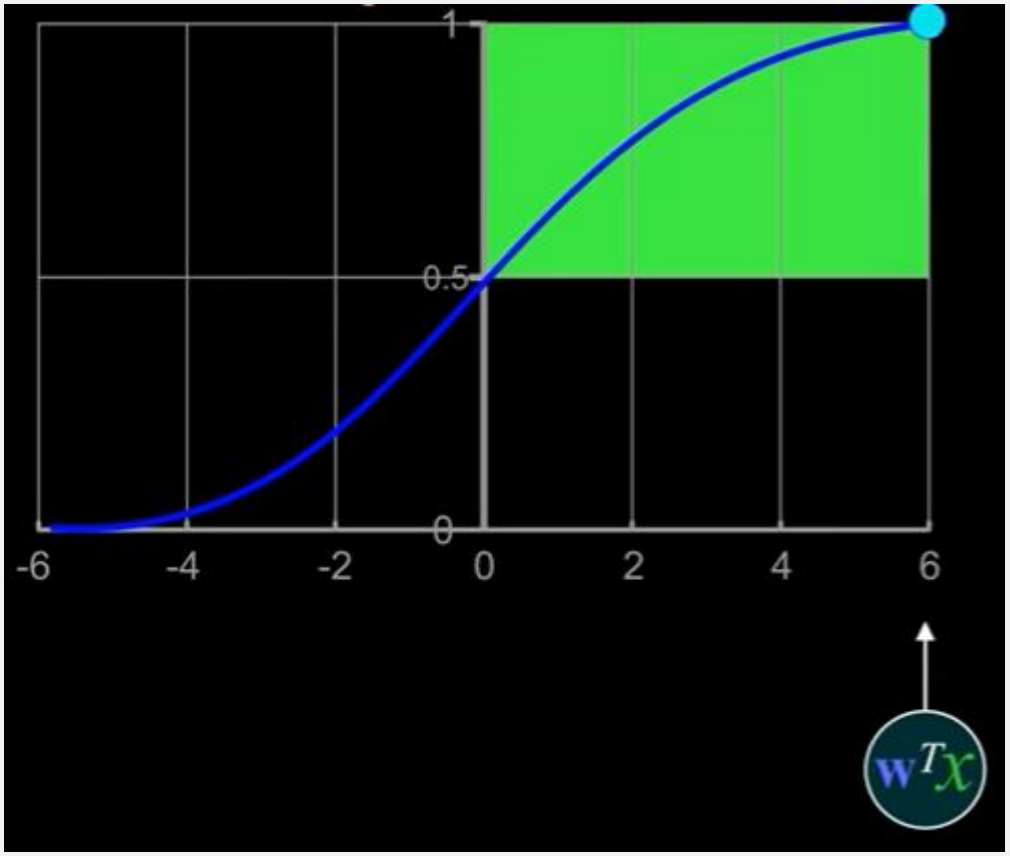
```
def sigmoid(z):
    return 1 / (1 + np.exp(-z))

sigmoid(X @ w.T)
```

Classificação - Logistic Regression



Classificação - Logistic Regression



Classificação - Logistic Regression

Função Erro

EM REGRESSÃO LINEAR

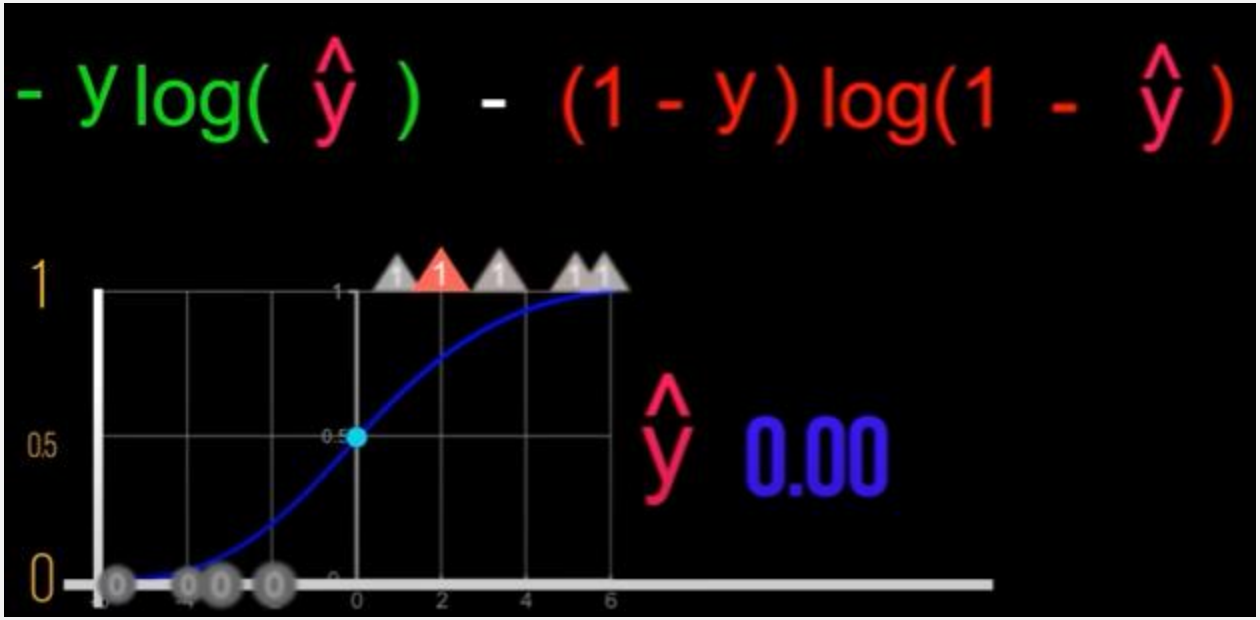
$$(\hat{y} - y)^2$$

EM REGRESSÃO LOGISTICA

$$-y \log(\hat{y}) - (1-y) \log(1 - \hat{y})$$

Classificação - Logistic Regression

Função Erro



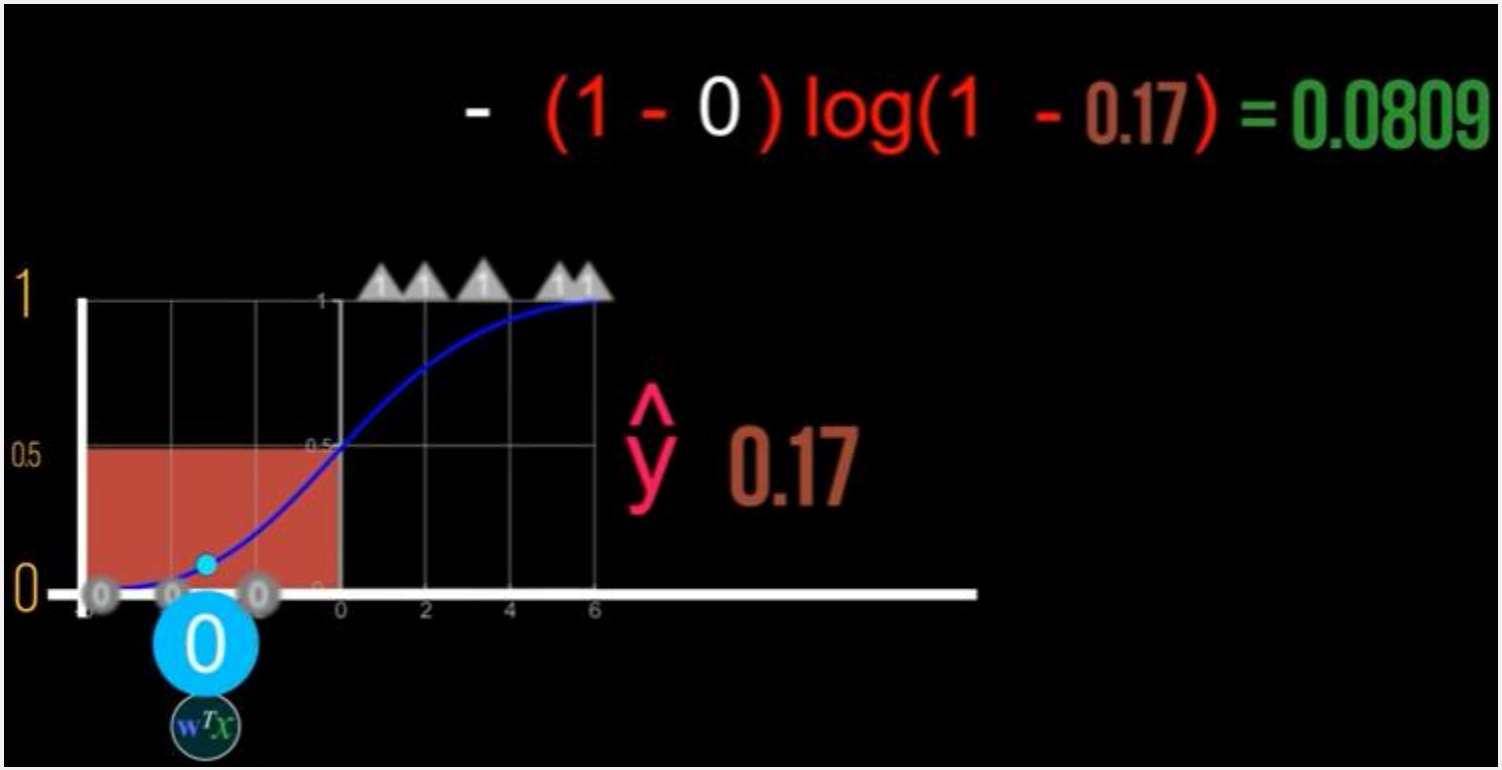
UDESC

Classificação - Logistic Regression

Função Erro

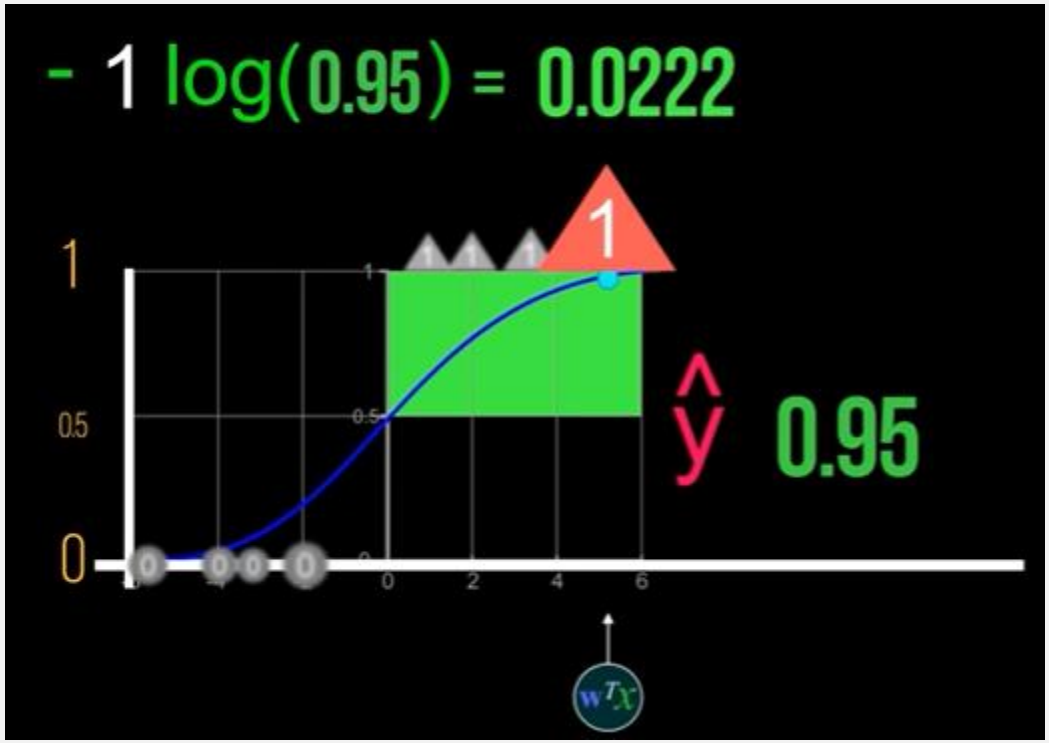
Classificação - Logistic Regression

Função Erro



Classificação - Logistic Regression

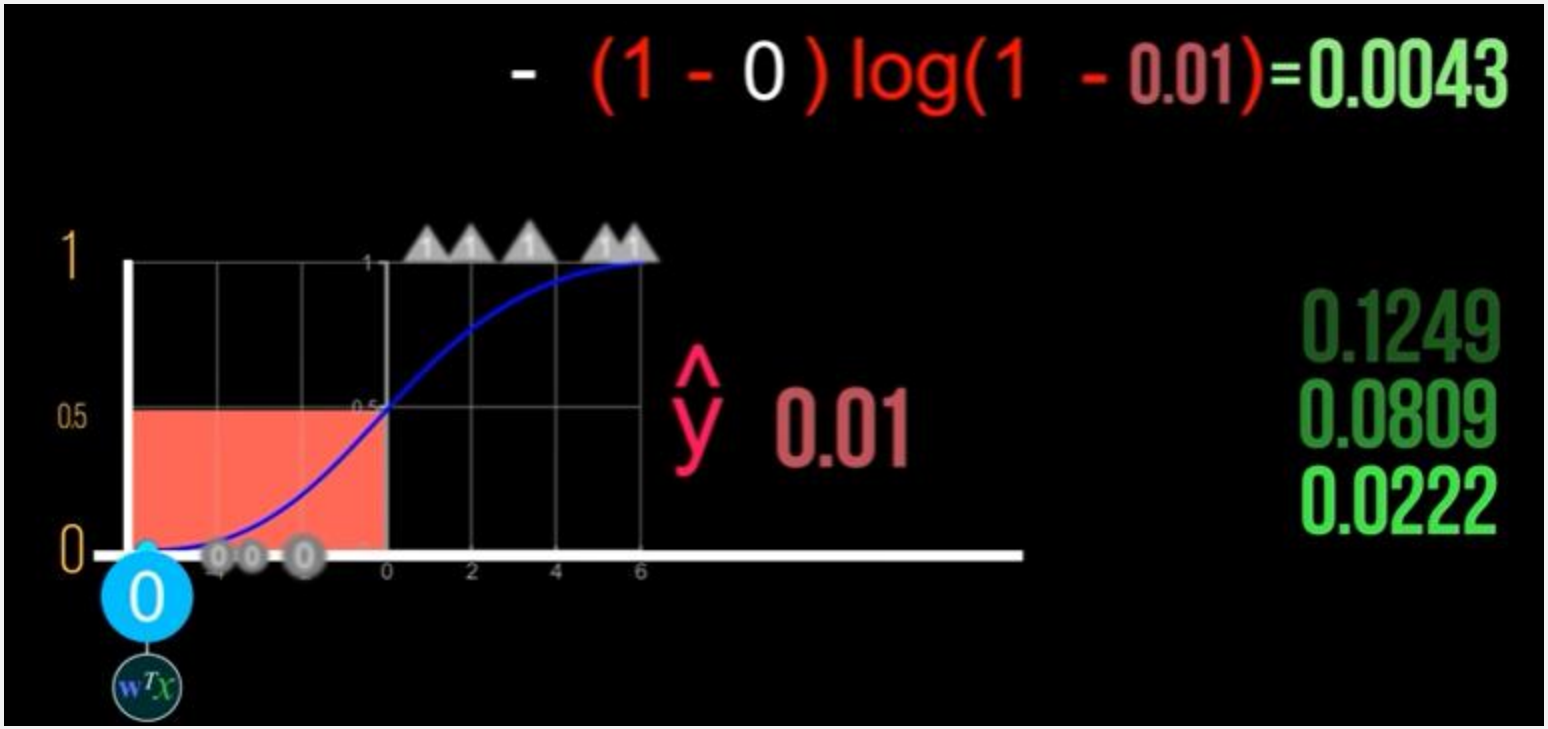
Função Erro



UDESC

Classificação - Logistic Regression

Função Erro



Classificação - Logistic Regression

Função Erro

EM REGRESSÃO LOGÍSTICA

$$\frac{1}{m}\sum_{i=1}^{m} -y \log(\hat{y}) - (1-y) \log(1 - \hat{y})$$

```
def cost(w, X, y):
    w = np.matrix(w)
    X = np.matrix(X)
    y = np.matrix(y)

    m = len(X)

    parte1 = np.multiply(-y, np.log(sigmoid(X @ w.T)))
    parte2 = np.multiply((1 - y), np.log(1 - sigmoid(X @ w.T)))

    somatorio = np.sum(parte1 - parte2)

    return  somatorio/m
```

UDESC

Classificação - Logistic Regression

Gradiente

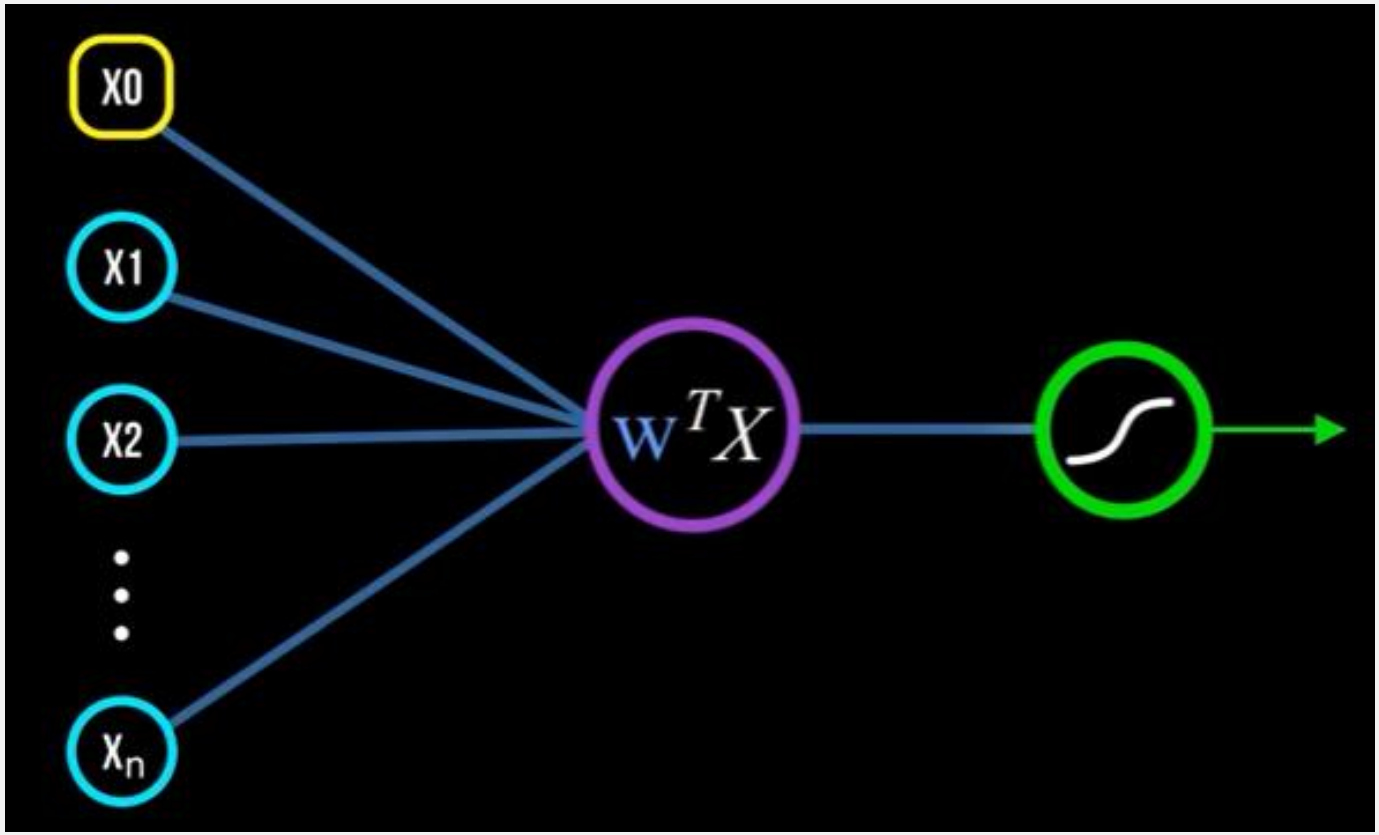
Para cada j em $\mathbf{w}$ {

$$w_J := w_J - \alpha \frac{1}{m} \sum_{i=1}^{m} (\hat{y} - y^{(i)}) x_J^{(i)}$$

}

$$\hat{y} = g(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

Classificação - Logistic Regression



UDESC

Trabalho Final (parte 1)

Análise Exploratória de Dados
Data de entrega do relatório final: 07/12

Dúvidas e envio do trabalho:

e-mail: bernarddss62gmail.com

UDESC

Referências

- https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

- https://playground.tensorflow.org/#activation=tanh&batchSize=14&dataset=spiral&regDataset=reg-gauss&learningRate=0.03&regularizationRate=0&noise=20&networkShape=6,3,1&seed=0.74989&showTestData=false&discretize=false&percTrainData=50&x=true&y=true&xTimesY=false&xSquared=false&ySquared=false&cosX=false&sinX=false&cosY=false&sinY=false&collectStats=false&problem=regression&initZero=false&hideText=false