IF867 - Introdução à Aprendizagem Profunda

1ª Lista Teórica

Discente:

- Gabriel D'assumpção de Carvalho - gdc2@cin.ufpe.br

Curso:

- Ciências Atuariais - 7º Período

04/12/2024

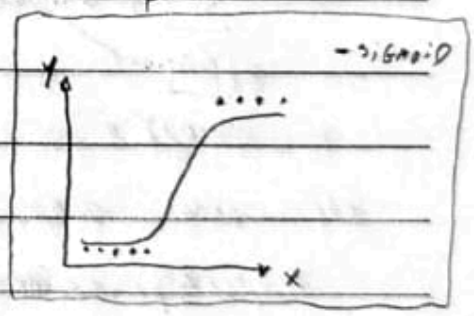
LISTA 10

1) A REGRESSÃO LOGÍSTICA E LINEAR SÃO MODELOS ESTATÍSTICOS, QUE FAZEM ESTIMATIVA DE UMA VARIÁVEL DEPENDENTE/RESPOSTA (Y) A PARTIR DE UM CONJUNTO DE VARIÁVEIS EXPLICATIVAS/INDEPENDENTES ($\tilde{X}$).

ESSES DOIS MODELOS SÃO MUITO UTILIZADOS PARA PREVISÃO DE RISCO DE INADIMPLÊNCIA, PREÇO DE IMÓVEIS, CUSTOS DE CLIENTES DE PLANOS DE SAÚDE, SE VAI CHOVER OU NÃO NO DIA.

POR MAIS QUE OS DOIS MODELOS SEJAM UTILIZADOS PARA FAZER PREVISÕES, A REGRESSÃO LOGISTICA É UTILIZADA APENAS QUANDO SE TEM A VARIÁVEL (Y) DICOTÔMICA, ASSUMINDO APENAS OS VALORES 0 (ZERO) E 1 (UM). COM BASE NISSO, ESSA REGRESSÃO VAI ESTIMAR A PROBABILIDADE DA OCORRÊNCIA DE UM EVENTO DADO AS VARIÁVEIS X E UM CONJUNTO DE PARÂMETRO β.

$$P(Y=1 \mid \tilde{X}, \tilde{\theta}) = \underbrace{\frac{1}{1 + e^{-\tilde{\beta}^T \tilde{X}}}}_{\text{SIGNOID}}$$

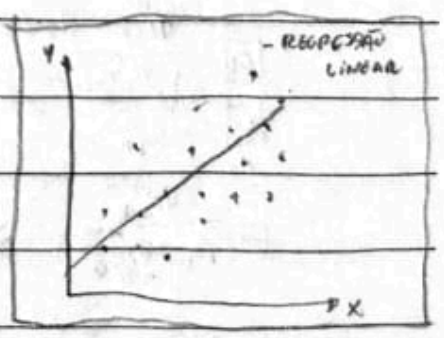


NO MODELO DE REGRESSÃO LINEAR A VARIÁVEL (Y) DEVE SER CONTÍNUA ($Y \in \mathbb{R}$). PORTANTO, ESSE MÉTODO ESTIMA PONTUALMENTE O VALOR DE (Y) A PARTIR DOS CONJUNTO X E β

$$Y = \underbrace{\tilde{\beta}^T \cdot \tilde{X}}_{\text{VETOR VARIÁVEL EXPLICATIVA}} + \varepsilon \leftarrow \text{RESÍDUO}$$

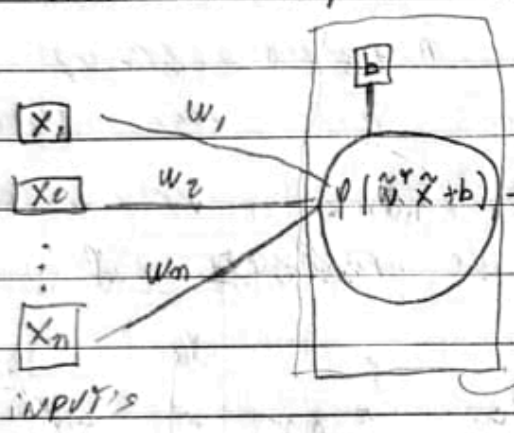
VETOR DE PARÂMETROS

VARIÁVEL RESPOSTA



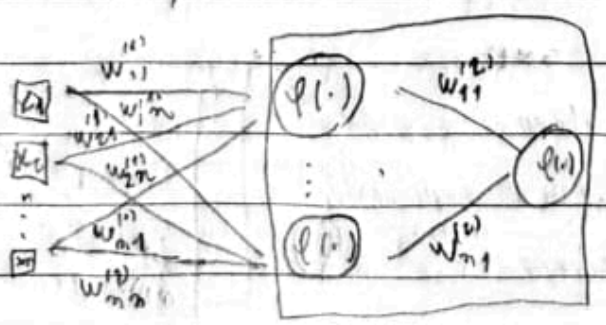
2) O MODELO PERCEPTRON FOI UMA DAS PRIMEIRAS ESTRUTURA DE REDES NEURAIS ARTIFICIAIS PROPOSTAS. ELA É INSPIRADA NO FUNCIONAMENTO DE UM ÚNICO NEURÔNIO BIOLÓGICO, ONDE RECEBE INFORMAÇÕES DE ENTRADA (INPUTS) ATRAVÉS DE "DENDRITOS", PROCESSA A INFORMAÇÃO EM SEU "NÚCLEO" POR MEIO DE UMA FUNÇÃO DE ATIVAÇÃO, E TRANSMITE O RESULTADO PROCESSADO PELA "SINAPSE" COMO SAÍDA (OUTPUT). ESSE PROCESSO É REPRESENTADO MATEMATICAMENTE COMO UMA COMBINAÇÃO LINEAR DAS ENTRADAS PONDERADAS ($w \cdot x + b$) SEGUIDA DE UMA FUNÇÃO DE ATIVAÇÃO PARA PREDIZER O VALOR DA SAÍDA

• ESTRUTURA PERCEPTRON



COMO PODE SER VISTO ACIMA, O MODELO PERCEPTRON É BASTANTE SIMPLES, SENDO LIMITADO A PROBLEMAS LINEARMENTE SEPARÁVEIS.

ALGUNS ANOS APÓS O PERCEPTRON FOI PROPOSTO O MULTI LAYER PERCEPTRON, SUA ESTRUTURA É COMPOSTA POR MAIS CAMADAS OCULTAS, AFIM DE RESOLVER PROBLEMAS NÃO LINEARES.



O TEOREMA DA APROXIMAÇÃO UNIVERSAL DIZ QUE UM MLP COM UMA CAMADA, M NEURÔNIOS E O CONJUNTO CERTO DE PESOS É POSSÍVEL MODELAR QUALQUER FUNÇÃO CONTÍNUA

DADO O CONJUNTO DOS PESOS, É IMPORTANTE ENCONTRAR OS VALORES ÓTIMOS PARA OS bias QUE É UM PARÂMETRO DE DESLOCAMENTO LINEAR, SEM O bias, TODAS AS FUNÇÕES DE ATIVAÇÃO IRIAM PASSAR PELO PONTO (0, 0), PREJUDICANDO O MODELO A APRENDER PADRÕES MAIS COMPLEXO. PORTANTO, SEM O bias O TEOREMA DA APROXIMAÇÃO UNIVERSAL NÃO SERIA APLICADO NA PRÁTICA.

3) A REGULARIZAÇÃO EM MODELOS ESTATÍSTICOS TEM A FINALIDADE DE RESTRINGIR O ESPAÇO PARAMÉTRICO, PARA ESTIMAR UM MODELO MAIS SIMPLES E COM UMA MELHOR CAPACIDADE DE GENERALIZAÇÃO, DIMINUINDO O OVERFITTING. ESSA RESTRIÇÃO É FEITA COM UMA PENALIZAÇÃO NA FUNÇÃO DE PERDA.

O REGULARIZADOR l_1 (LASSO) PENALIZA A FUNÇÃO DE PERDA $L(\tilde{w}, \tilde{b})$ PELO PRODUTO DA SOMA DO MÓDULO DOS PESOS $|\tilde{w}|$ POR UM HIPERPARÂMETRO (λ) QUE CONTROLA A FORÇA DE REGULARIZAÇÃO.

$$J(\tilde{w}, \tilde{b}) = L(\tilde{w}, \tilde{b}) + \hat{\Sigma}_{i=1} |w_i| \lambda$$

PELO FATO DA DERIVADA DO l_1 SER

$$\frac{G(\lambda |w|)}{Gw} = \begin{cases} \lambda, & \text{SE } w > 0 \\ -\lambda, & \text{" } w < 0 \end{cases}$$

QUANDO FOR CALCULADO O GRADIENTE DESCENDENTE O REGULARIZADOR l_1 VAI FORÇAR OS PESOS PARA A ORIGEM, PODENDO ZERAR ALGUNS

PESOS. PORTANTO, O l_1 É BASTANTE VANTAJOSO QUANDO QUEREMOS DIMINUIR A DIMENSIONALIDADE DO MODELO, OBSERVANDO AS VARIÁVEIS QUE SÃO MAIS IMPORTANTES.

JÁ O REGULARIZADOR l_2 VAI PENALIZAR $L(\vec{w}, \vec{b})$ PELO PRODUTO DA SOMA DOS QUADRADOS DOS PESOS POR λ

$$J(\vec{w}, \vec{b}) = L(\vec{w}, \vec{b}) + \sum_{i=1}^{n} w_i^2 \lambda$$

DIFERENTE DO l_1, O l_2 PUXA OS PESOS PARA A ORIGEM DE FORMA CONTÍNUA, PELO FATO DA SUA DERIVADA SER

$$\frac{\partial}{\partial w}(w^2 \lambda) = 2w\lambda$$

COM ISSO O REGULARIZADOR l_2 NÃO ZERA ABSOLUTAMENTE UM w_i, MAS PODENDO DEIXAR ELE BEM PRÓXIMO DE ZERO. ENTRETANTO ELE É UM BOM REGULARIZADOR PARA SE UTILIZAR QUANDO SE TEM TODAS AS VARIÁVEIS IMPORTANTE MAS SE PRETENDE TER UM MODELO QUE CONSIGA GENERALIZAR PARA NOVAS OBSERVAÇÕES.

COMO VISTO ANTERIORMENTE A ESCOLHA DO HIPERPARÂMETRO λ IMPACTA NA FORÇA DA REGULARIZAÇÃO, CONTROLANDO A COMPLEXIDADE DO MODELO E CAPACIDADE DE GENERALIZAÇÃO. UM λ BAIXO FAZ COM QUE O ERRO DO TREINAMENTO TENHA POUCA ALTERAÇÃO, DANDO MARGEM DO MODELO SE AJUSTAR AOS DADOS; COM UM λ ALTO O MODELO FICA FORÇADO A TER UMA TAXA DE ERRO MAIS ALTA, FAZENDO COM QUE O MODELO SE TORNE MENOS COMPLEXO E EVITANDO OVERFITTING MAS PODENDO TER UM UNDERFITTING. PORTANTO É IMPORTANTE ESCOLHER UM λ NEM TÃO ALTO E NEM TÃO BAIXO, PODENDO UTILIZAR TÉCNICAS DE VALIDAÇÃO CRUZADA PARA ENCONTRAR λ^*.

)

4) NAS PRIMEIRAS REDES ARTIFICIAIS OS CIENTISTAS ESTAVAM INTERESSADOS EM MODELAR UM NEURÔNIO QUE DISPARA E NÃO DISPARA, UTILIZANDO UMA FUNÇÃO DE LIMIARIZAÇÃO. POR CONTA DO USO DO GRADIENTE A FUNÇÃO SIGMOIDE FOI UMA ESCOLHA UTILIZADA, SENDO UMA FUNÇÃO NÃO LINEAR, CONTÍNUA, SUAVE E DIFERENCIÁVEL EM TODOS OS PONTOS.

A FUNÇÃO SIGMOIDE É UTILIZADA QUANDO QUEREMOS MODELAR PROBABILIDADE, TENDO O VALOR DE SAÍDA PERTENCENTE AO INTERVALO $[0, 1]$. POR CONTA DISSO, ELA É BASTANTE UTILIZADA PARA PROBLEMAS DE CLASSIFICAÇÃO BINÁRIA.

COM O AVANÇO DA COMPUTAÇÃO, FOI POSSÍVEL CRIAR ESTRUTURAS DE NN COM MUITAS CAMADAS OCULTAS. ENTRETANTO, AO UTILIZAR A FUNÇÃO SIGMOIDE EM MUITAS CAMADAS A REDE SE TORNA INCAPAZ DE RETROPROPAGAR O GRADIENTE DA CAMADA DE SAÍDA DE VOLTA PARA AS PRIMEIRA CAMADAS. ISSO ACONTECE PORQUE O PONTO MÁXIMO DA DERIVADA DA SIGMOIDE É IGUAL A 0,25, APROXIMANDO O GRADIENTE PARA O PONTO ZERO À CADA CAMADA OCULTA NO BACKPROPAGATION, CAUSANDO A CHAMADA DISSIPAÇÃO DO GRADIENTE.

A FUNÇÃO RELU FOI IMPLEMENTADA EM REDES NEURAIS PROFUNDAS, PARA SUPRIMIR A LIMITAÇÃO DA SIGMOIDE NO TREINAMENTO E DEVIDO AO SEU BAIXO CUSTO COMPUTACIONAL, SENDO ELA

$$\text{RELU}(x) = \max(0, x) \quad \text{e} \quad \text{RELU}'(x) = \begin{cases} 1, & \text{se } x > 0 \\ 0, & \text{se } x \leq 0 \end{cases}$$

5)

① O MODELO QUE MELHOR GENERALIZA É O TERCEIRO POIS ALÉM DE TER A 2ª MAIOR ACURÁCIA DE TREINAMENTO POSSUI O MELHOR DESEMPENHO NOS DADOS DE TESTE.

② O 1 MODELO APRESENTOU UM POSSÍVEL OVERFITTING, PELO FATO DA ACURÁCIA DO TESTE ESTÁ EM 80%, SENDO 15% A BAIXO DA ACURÁCIA DO TREINAMENTO.

③ O MODO QUE APRESENTOU UNDERFITTING FOI O SEGUNDO MODELO, POIS ELE NÃO FOI CAPAZ DE APRENDER OS PADRÕES NOS DADOS DE TREINO E TESTE, FICANDO COM UMA ACURÁCIA DE 70% E 68% RESPECTIVAMENTE.

④ EXISTEM DIVERSAS TECNICAS QUE PODEM SER UTILIZADAS, DENTRE ELAS EXISTEM OS REGULARIZADORES DA FUNÇÃO DE PERDA, DIMINUIR A PROFUNDIDADE DA REDE, EARLY STOP QUE FAZ UMA PARADA ANTECIPADA OU USAR UM ENSEMBLE LEARNING QUE CONSISTE EM TREINAR DIVERSOS MODELOS E FAZER A PREVISÃO COM BASE EM UMA MÉDIA

⑥ DIFERENTE DO OVERFITTING, PARA EVITAR O UNDERFITTING PODE TENTAR RECOLHER NOVOS DADOS E TRATAR DE UMA MANEIRA MELHOR, AUMENTAR A COMPLEXIDADE DO MODELO E AUMENTAR A EPOCH TIME.

a) QUANDO O DROPOUT É UTILIZADO NA MATRIZ DE PESOS VAI SER DESLIGADO EM MÉDIA (S x P) SINAPSES DE CADA CAMADA, SENDO (S) O NÚMERO DE SINAPSES E (P) A PROBABILIDADE DE DROP OUT, MAS DESLIGAR APENAS AS SINAPSES O MODELO AINDA PODE FICAR DEPENDENTE DE ALGUNS NEURÔNIOS, FAZENDO COM QUE ESSA TÉCNICA ADICIONE POUCO RUÍDO PARA A REDE. A UTILIZAÇÃO DE DROPOUT APÓS A FUNÇÃO DE ATIVAÇÃO É A PRÁTICA MAIS COMUM, POIS EM MÉDIA DESLIGA (M x P) NEURÔNIOS DA CAMADA, SENDO (M) O NÚMERO DE NEURÔNIOS. ISSO FAZ COM QUE O MODELO NÃO DEPENDA DE CERTOS NEURÔNIOS PARA FAZER A SUA ESTIMAÇÃO, SENDO UMA TÉCNICA QUE ADICIONA MAIS RUÍDO E MELHORA A CAPACIDADE DE GENERALIZAÇÃO DO MODELO.

b) O GRADIENT EXPLOSIVE É UM PROBLEMA QUE SE DÁ QUANDO OS PARÂMETROS DA REDE RECEBE GRANDES ATUALIZAÇÕES NO BACK-PROPAGATION FAZENDO COM QUE OS PARÂMETROS SEMPRE FIQUEM COM VALORES ALTOS E NÃO VÃO PARA O MÍNIMO. ISSO PODE ACONTECER QUANDO OS VALORES INICIAIS DOS PARÂMETROS SÃO ALTOS OU A DERIVADA DA FUNÇÃO DE ATIVAÇÃO SEJA MAIOR QUE UM, FAZENDO COM QUE O GRADIENTE SEMPRE AUMENTE A CADA CAMADA OCULTA.

O VANISH GRADIENT É O INVERSO DO EXPLOSIVE, QUE SERIA A ATUALIZAÇÃO INSIGNIFICANTE DOS PARÂMETROS NA ETAPA DE BACKPROPAGATION. ESSE PROBLEMA PODE ACONTECER SE A DERIVADA DA FUNÇÃO DE ATIVAÇÃO TIVER GRANDES PONTOS DE SATURAÇÃO, COMO A FUNÇÃO SIGMOIDE.

QUANDO ESCOLHEMOS UM GRANDE VALOR PARA A TAXA DE APRENDIZADO O AJUSTE DOS PARÂMETROS SEMPRE VÃO DAR UM SAÍ

TO EM TORNO DO MÍNIMO LOCAL OU GLOBAL, EVITANDO A CONVERGÊNCIA. ESSA SITUAÇÃO É DENOMINADA DE OVER SHOOTING.

7) QUANDO TEMOS UM PROBLEMA DE REGRESSÃO ESTAMOS INTERESSADOS EM FAZER PREVISÕES PARA VALORES CONTÍNUOS, PARA FAZER BOAS PREVISÕES DEVEMOS MINIMIZAR O NOSSO ERRO QUE SERIA A DISTÂNCIA DO VALOR PREVISTO ($\hat{Y}$) PARA O VALOR OBSERVADO (Y), COM ISSO PODEMOS USAR AS SEGUINTES FUNÇÕES DE PERDA:

$$MSE = n^{-1} \cdot \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 \quad \text{OU} \quad MAE = n^{-1} \cdot \sum_{i=1}^{n} |\hat{Y}_i - Y_i|$$

COMO PODEMOS VER A LINA, A FUNÇÃO MSE ELEVA A DISTÂNCIA DE ($\hat{Y}_i$) E (Y_i) AO QUADRADO, FAZENDO COM QUE ERROS PEQUENOS SEJAM MAIS INSIGNIFICANTES MAS PENALIZA MAIS OS ERROS GRANDES. SE TIVERMOS GRANDES QUANTIDADES DE OUTLIERS É INTERESSANTE UTILIZAR A FUNÇÃO MAE POR SER MAIS ROBUSTA, PELO FATO DE CALCULAR A DISTÂNCIA ENTRE ($\hat{Y}_i$) E (Y_i) DE FORMA LINEAR.

QUANDO TEMOS PROBLEMAS DE CLASSIFICAÇÃO FICAMOS INTERESSADOS EM MODULAR A PROBABILIDADE DE CADA CLASSE (Y_i) DADO UM CONJUNTO DE CARACTERÍSTICAS X E UM CONJUNTO DE PARÂMETRO θ. UMA DAS FUNÇÕES DE PERDA MAIS UTILIZADA PARA CLASSIFICAÇÃO É A CROSS-ENTROPY, QUE É DADA POR

$$L_{CE} = n^{-1} \cdot \sum_{i=1}^{n} [Y_i \cdot \log(\hat{Y}_i) - (1-Y_i) \cdot \log(1-\hat{Y}_i)]$$

SENDO $\hat{Y}_i = P(Y_i | X; \theta)$

8)

① O GRADIENTE DESCENDENTE É UM ALGORITMO DE OTIMIZAÇÃO UTILIZADO PARA MINIMIZAR A FUNÇÃO DE PERDA, SENDO UTILIZADO NA FAZE DE TREINAMENTO DO MODELO. ELE FUNCIONA AJUSTANDO ITERATIVAMENTE OS PARÂMETROS (θ) NA DIREÇÃO OPOSTA DO GRADIENTE DA FUNÇÃO DE PERDA EM RELAÇÃO AOS PARÂMETROS.

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L(\theta)}{\partial \theta}$$

- θ_t: PARÂMETRO NA EPOCH T
- θ_{t+1}: " " " T+1
- η: TAXA DE APRENDIZADO
- $L(\theta)$: FUNÇÃO DE PERDA

ESSE ALGORITMO É BASTANTE UTILIZADO POIS É O MÉTODO MAIS EFICIENTE PARA ENCONTRAR OS θ^*, SENDO CAPAZ DE SER UTILIZADOS EM MODELOS DE REGRESSÕES E REDES NEURAIS E TEM A CAPACIDADE DE FUNCIONAR PARA GRANDES QUANTIDADES DE DADOS.

②

- BATCH: ESSE É UM MÉTODO QUE CONSISTE EM UTILIZAR TODAS AS OBSERVAÇÕES DO CONJUNTO DE TREINAMENTO PARA CALCULAR A MÉDIA DA FUNÇÃO DE PERDA PARA DEPOIS FAZER A ATUALIZAÇÃO DOS PARÂMETROS. PORTANTO, A SUA CONVERGÊNCIA PARA O MÍNIMO É SUAVE MAS SO É RECOMENDADA PARA UMA QUANTIDADE PEQUENA DE OBSERVAÇÕES, CASO CONTRARIO LEVARIA MUITAS EPOCAS ATÉ SUA CONVERGÊNCIA.

• STOCHASTIC: É UM MÉTODO QUE UTILIZA APENAS UMA OBSERVAÇÃO ALEATÓRIA DA BASE DE TREINAMENTO, FAZENDO COM QUE OS PARÂMETROS SEJAM FREQUENTEMENTE MODIFICADO. ENTRETANTO, A SUA NATUREZA ESTOCÁSTICA FAZ COM QUE ELE POSSA ESTAGNAR EM MÍNIMOS LOCAIS MAS OSCILA EM TORNO DO MÍNIMO GLOBAL, SENDO MAIS INSTÁVEL, O QUE PODE DIFICULTAR A CONVERGÊNCIA.

• MINI-BATCH: É O MÉTODO MAIS COMUM, POIS É UMA UNIÇÃO DO BATCH COM O STOCHASTIC. ELE FUNCIONA SELECIONANDO UM AMOSTRA DE TAMANHO K DA BASE DE TREINAMENTO, FAZENDO COM QUE NÃO DEMORE TANTO POR PARA ATUALIZAR OS PARÂMETROS E TEM POSSIBILIDADE DE FUGIR DE MÍNIMOS LOCAIS.

q)

① A TAXA DE APRENDIZADO É UM HIPERPARÂMETRO QUE CONTROLA O TAMANHO DO PASSO QUE O ALGORITO DE OTIMIZAÇÃO DÁ EM DIREÇÃO AO MÍNIMO DA FUNÇÃO DE CUSTO. ELE CONTROLA A VELOCIDADE DA CONVERGÊNCIA, AFETANDO A ESTABILIDADE E EFICIÊNCIA DURANTE O TREINAMENTO.

② UMA TAXA DE APRENDIZADO MUITO ALTA VAI FAZER COM QUE OS PARÂMETROS DE GRANDES SALTOS, DE ESSA TAXA POR CONSTANTE AO LONGO DO TREINAMENTO PODE SER QUE O MODELO NÃO CONVERGIA PARA O MÍNIMO, FICANDO SEMPRE EM TORNO DELE.

AO CONTRÁRIO, QUANDO SE TEM UMA TAXA MUITO BAIXA

O MODELO DEMORA MUITAS ÉPOCAS PARA CONVERGIR, PODENDO FICAR PRESO EM MÍNIMOS LOCAIS.

⑨ AS ESTRATÉGIAS ADAPTATIVAS AJUSTAM AUTOMATICAMENTE A TAXA DE APRENDIZADO COM BASE NO GRADIENTE, TORNANDO O TREINAMENTO MAIS ESTÁVEL E EFICIENTE.

UM DOS OTIMIZADORES MAIS UTILIZADO É O ADAGRAD, SENDO ELE

$$SDG \Rightarrow \theta_{T+1} = \theta_T - \eta \cdot (\sqrt{G_T + \varepsilon})^{-1} \cdot \frac{\partial L(\theta)}{\partial \theta}$$

- θ_T: PARÂMETRO NO TEMPO T
- θ_{T+1}: " " " T+1
- η: TAXA DE APRENDIZADO
- G_T: SOMA DOS QUADRADOS DOS GRADIENTES
- $\frac{\partial L(\theta)}{\partial \theta}$: GRADIENTE DA FUNÇÃO DE PERDA EM RELAÇÃO A θ

A TAXA DE APRENDIZADO É AJUSTADO PARA CADA PARÂMETRO DE ACORDO COM O GRADIENTE DA FUNÇÃO DE PERDA EM RELAÇÃO A θ. PORTANTO SE UM PARÂMETRO TEM GRADIENTES GRANDES A TAXA DIMINUI, CASO CONTRÁRIO, A TAXA AUMENTA.

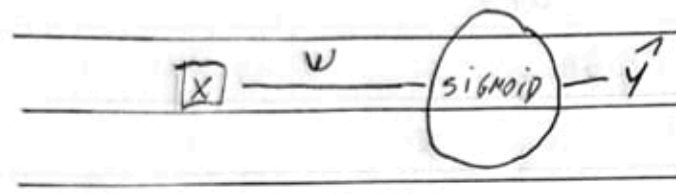
10) O BACKPROPAGATION É UM ALGORITMO DE OTIMIZAÇÃO QUE UTILIZA O GRADIENTE DESCENDENTE E A REGRA DA CADEIA PARA AJUSTAR OS VALORES DOS PARÂMETROS DA ULTIMA CAMADA ATÉ A PRIMEIRA.

O FORWARD É O PRIMEIRO PASSO NO CICLO DE TREINAMENTO. FUNCIONANDO COM O RECEBIMENTO DE UM CONJUNTO X DE VARIÁVEIS, SENDO PASSADO PARA FRENTE O RESULTADO DA FUNÇÃO $\varphi(W \cdot X + b)$ ATÉ A ULTIMA CAMADA, QUE FAZ A ESTIMAÇÃO

Depois do forward pass é feito o backward, que consiste em utilizar o backpropagation para ajustar os valores dos parâmetros da última camada até a primeira.

X Questão bonus

· Estrutura da Rede:



como pode ser observado a rede proposta é um perceptron que recebe apenas 1 features e sua função de ativação é uma sigmoid, isso sugere que o problema a ser resolvido é de classificação binária.

①

$z = 0,1 \cdot 3 = 1,2$

$\hat{y} = (1+e^{-1,2})^{-1} \approx 0,2315$

$E \approx (0,9 - 0,2315)^2 \approx 0,4469$

②

· $\frac{\partial}{\partial \hat{y}}(\hat{y}-y)^2 = 2(\hat{y}-y)$

· $\frac{\partial}{\partial z}(1+e^{-z})^{-1} = e^{-z}(1+e^{-z})^{-2} = \frac{1}{(1+e^{-z})} \cdot \frac{e^{-z}}{(1+e)^{-z}} = \frac{1}{(1+e^{-z})} \cdot \left[1 - \frac{1}{(1+e^{-z})}\right] = \hat{y}\cdot(1-\hat{y})$

· $\frac{\partial}{\partial w}(w \cdot x) = x$

③ $\frac{\partial E}{\partial w} = 2(\hat{y}-y)\cdot\hat{y}\cdot(1-\hat{y})\cdot x$

④

$w_{t+1} = 0,4 - 0,1 \cdot 2 \cdot (0,2315 - 0,9) \cdot 0,2315 \cdot (1 - 0,2315) \cdot 3 = 0,4714$

$\frac{\partial E}{\partial w} = -0,7136$