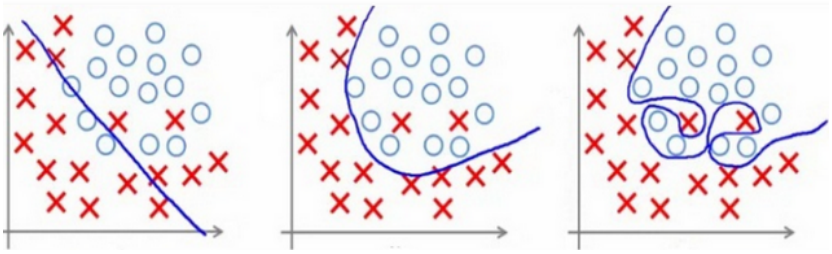
# Exercícios Deep Learning - Lista Teórica 03
Elaboração de Carolina Coimbra, Guilherme Borges e Edemir Andrade Jr.

September 6, 2019

## 1 Ajuste de modelos

**1-** Observe as figuras abaixo e procure identificar as características de cada um dos modelos que foram gerados para classificar os dados em duas categorias. Para aqueles modelos cujo ajuste é ruim, cite um problema causado por ele e proponha sugestões para solucioná-lo.



## 2 Regularização

**2-** Considere a seguinte função objetiva de mínimos quadrados regularizada L2 para uma regressão linear: $f(w) = \|\mathbf{w}'\mathbf{x} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2$. Como $\lambda$ afeta a reta estimada?

**3-** Considere uma rede neural com duas features e uma camada escondida de dois nós. Sejam $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}$ os pesos da rede:

$$W^{[1]} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} b^{[1]} = \begin{bmatrix} -2 \\ 4 \end{bmatrix} W^{[2]} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} b^{[2]} = \begin{bmatrix} -4 \end{bmatrix}$$

E os gradientes em relação a função de perda:

$$\frac{\partial J}{\partial W^{[1]}} = \begin{bmatrix} -10 & 5 \\ 1 & 2.5 \end{bmatrix} \frac{\partial J}{\partial b^{[1]}} = \begin{bmatrix} 3 \\ -3 \end{bmatrix} \frac{\partial J}{\partial W^{[2]}} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \frac{\partial J}{\partial b^{[2]}} = \begin{bmatrix} -2 \end{bmatrix}$$

**a)** Seja $\alpha = 0.1$, faça uma iteração do backpropagation nos pesos da rede. Além disso, considerando a regularização L2 com $\lambda = 0.5$, atualize os pesos considerando a regularização.

**b)** Qual diferença você nota entre os pesos com e sem regularização?

**c)** Ao final de várias iterações, após a rede convergir, o que você espera que seja a diferença entre os pesos das duas redes? Explique qual será o efeito provável da regularização no erro no conjunto de testes e porque isso ocorre.

# 3 Normalização

**4-** Explique qual é o efeito causado pela normalização das entradas, $\mathbf{x}$, no treinamento da rede.

# 4 Dropout

**5-** Suponha que você esteja treinando uma rede com dropout e a probabilidade de um nó ser mantido muda de 0.6 para 0.5.

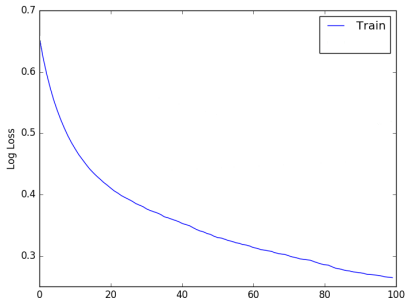
**a)** O que acontece com o efeito de regularização ao diminuir a probabilidade de um nó ser mantido na rede?

**b)** Qual o impacto dessa mudança no erro calculado com o conjunto de treino?

**c)** Durante o treinamento com dropout, a rede é modificada diversas vezes. Alguma modificação deve ser feita na rede em tempo de teste?

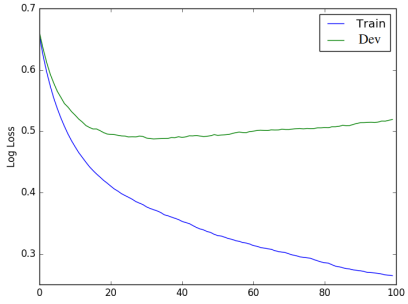
# 5 Treino, teste e validação

**6-** Imagine que você está treinando um modelo e ao plotar a curva de erro ao longo das iterações com os dados de treino você obtenha o seguinte gráfico:




**a)** Estime o número da iteração onde o modelo obteve o melhor resultado, ou seja, sua predição foi melhor.

Considere agora que, além de plotar a curva de erro ao longo das iterações com os dados de treino, a curva de erro para o conjunto de dados de validação ao longo das mesma iterações também é considerada, conforme mostrado no gráfico abaixo:

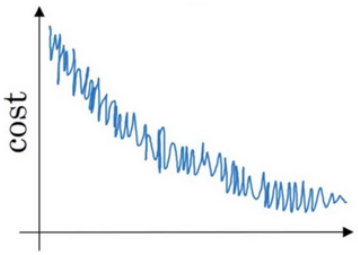



**b)** Estime o número da iteração onde o modelo obteve o melhor resultado, ou seja, onde de forma geral, sua predição foi melhor.

**c)** Explique a importância do conjunto de validação nesse caso e indique uma estratégia para obter o erro mínimo nesse conjunto.

# 6 Gradiente descendente, Mini-batch e SGD

**7-** Sobre o gradiente descendente em mini-batch, discorra sobre as vantagens e desvantagens de diferentes tamanhos para o batch. Por que o melhor tamanho de mini-batch geralmente não é 1 e nem m, mas algo intermediário?




**8-** Para cada uma das afirmativas abaixo, diga porque ela é falsa.

**a)** Treinar uma época (uma passagem pelo conjunto de treinamento) usando gradiente descendente em mini-batch é mais rápido do que treinar uma época usando descida gradiente em batch.

**b)** Uma iteração de gradiente descendente em mini-batch (computação em um único mini-batch) é mais lenta que uma iteração de gradiente descendente em batch.

**9-** Suponha que o custo J do seu algoritmo de aprendizado, plotado como uma função do número de iterações, seja representado no gráfico abaixo. A partir da análise do gráfico, explique qual deve ter sido o algoritmo utilizado: gradiente descendente em batch ou gradiente descendente em mini-batch.

# 7 Exponentially Weighted Average

**10-** Suponha que a temperatura (em graus Celsius) em Casablanca nos primeiros três dias de janeiro seja a mesma:

1º de janeiro: $\theta_1 = 10$
2 de janeiro: $\theta_2 = 10$

Digamos que você use uma média exponencialmente ponderada com $\beta = 0.5$ para acompanhar a temperatura: $v_0 = 0$, $v_t = \beta v_{t-1} + (1-\beta)\theta_t$. Se $v_2$ é o valor calculado após o dia 2 sem correção de viés, e $v_t^{corrigido} = \frac{v_t}{1-\beta^t}$ é o valor que você calcula com correção de viés. Quais são esses valores?

# 8 Notebook (Momentum, RMSprop, Adam)

**11-** Faça download do notebook S02A03 no drive e utilize o collab para completar os exercícios.