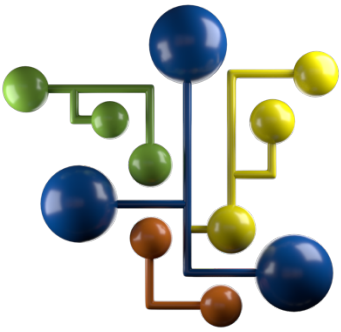


Data Science Academy

www.datascienceacademy.com.br

Deep Learning II

Pruning

O Pruning é um processo que torna as redes neurais mais eficientes. Ao contrário dos algoritmos de treinamento, o Pruning não aumenta o erro de treinamento da rede neural. O principal objetivo do Pruning é diminuir a quantidade de processamento necessário para usar a rede neural. Além disso, o Pruning às vezes pode ter um efeito "regularizante" ao remover a complexidade da rede neural. Essa regularização às vezes pode diminuir a quantidade de overfitting na rede neural. Esta diminuição pode ajudar a rede neural a ter um desempenho melhor quando apresentada a novos conjuntos de dados.

O Pruning funciona ao analisar as conexões da rede neural. O algoritmo de Pruning procura conexões individuais e neurônios que podem ser removidos da rede neural para fazê-la operar de forma mais eficiente. Ao podar conexões desnecessárias, a rede neural pode para executar mais rápido e, possivelmente, diminuir o overfitting. Podemos podar tanto as conexões quanto os neurônios.

Pruning de Conexões

O Pruning de conexões é a atividade principal para a maioria dos algoritmos de Pruning. O programa analisa as conexões individuais entre os neurônios para determinar quais conexões têm o menor impacto na eficácia da rede neural. As conexões não são a única coisa que o programa pode podar. Analisar as conexões podadas revelará que o programa também pode podar neurônios individuais.

Pruning de Neurônios

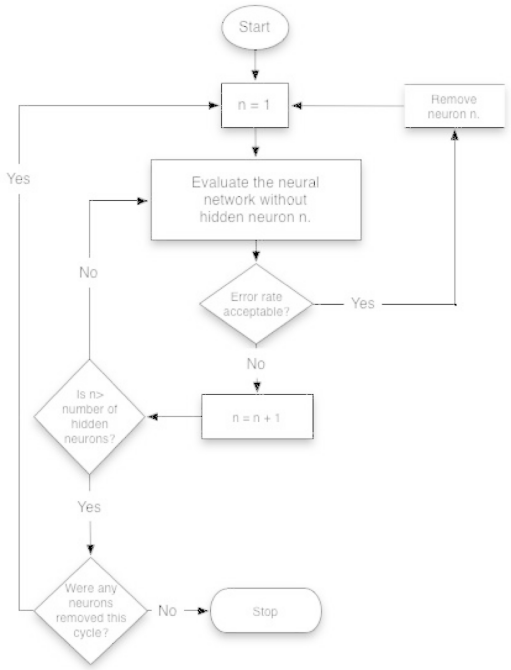
O Pruning centra-se principalmente nas conexões entre os neurônios individuais da rede neural. No entanto, para podar neurônios individuais, devemos examinar as conexões entre cada neurônio e os outros neurônios. Se um neurônio particular é cercado inteiramente por conexões fracas, não há motivo para manter esse neurônio. Se aplicarmos os critérios discutidos na seção anterior, os neurônios que não possuem conexões são o resultado final porque o programa cortou todas as conexões do neurônio. Então, o programa pode podar esse tipo de neurônio.

Melhorando ou Degradando a Performance

É possível que a poda de uma rede neural melhore seu desempenho. Qualquer modificação na matriz de peso de uma rede neural terá sempre algum impacto na precisão dos reconhecimentos feitos pela rede neural. Uma conexão que tem pouco ou nenhum impacto na rede neural pode realmente degradar a precisão com que a rede neural reconhece os padrões. A remoção desta conexão fraca pode melhorar a saída geral do modelo. Infelizmente, a poda também pode diminuir a eficácia da rede neural. Assim, você sempre deve analisar a eficácia da rede neural antes e depois da poda. Uma vez que a eficiência é o principal benefício da poda, você deve ter o cuidado de avaliar se uma melhoria no tempo de processamento vale uma diminuição da eficácia da rede neural. Avaliaremos a eficácia global da rede neural antes e depois da poda na sequência deste capítulo!

Algoritmo de Pruning

Agora vamos analisar exatamente como a poda ocorre. A poda funciona ao examinar as matrizes de peso de uma rede neural previamente treinada. O algoritmo de poda tentará então remover neurônios sem interromper a saída da rede neural. A figura abaixo mostra o algoritmo utilizado para a poda seletiva:



Como você pode ver, o algoritmo de poda tem uma abordagem de teste e erro. O algoritmo de poda tenta remover neurônios da rede neural até não poder remover neurônios adicionais sem degradar o desempenho da rede.

Para iniciar este processo, o algoritmo de poda percorre cada um dos neurônios ocultos. Para cada neurônio oculto encontrado, o programa avalia o nível de erro da rede neural com e sem o neurônio especificado. Se a taxa de erro salta para além de um nível predefinido, o programa mantém o neurônio e avalia o próximo. Se a taxa de erro não melhorar significativamente, o programa remove o neurônio. Uma vez que o programa avaliou todos os neurônios, ele repete o processo. Este ciclo continua até que o programa tenha feito uma passagem através dos neurônios ocultos sem remover um único neurônio. Uma vez que este processo está completo, uma nova rede neural é criada e que tem performance aceitável próximo da original, mas com menos neurônios ocultos.