IA048 – Aprendizado de Máquina
Exercícios de Fixação de Conceitos (EFC) 3 – 2s2020

Parte 1 – Classificação binária com redes MLP

Problema: detecção de diabetes
Base de dados: dados_diabetes.csv

O conjunto de dados contém informações clínicas de mulheres com idade superior a 21 anos de origem índigena (Pima) na região de Phoenix, Arizona. Mais especificamente, cada amostra é descrita por 8 atributos:

Atributos
Número de gestações
Concentração de glicose
Pressão sanguínea
Espessura da dobra da pele (tríceps)
Nível de insulina
Índice de massa corporal
Função de *pedigree* de diabetes[*]
Idade

A última coluna corresponde ao rótulo associado a cada padrão, sendo igual a ‘1’ para o caso de diagnóstico positivo para diabetes, e ‘0’ caso contrário.

a) Analise o conjunto de dados a partir dos histogramas dos atributos de entrada e discuta, também, o balanceamento das classes.

b) Considere uma rede MLP com uma única camada intermediária. Mostre a evolução da taxa de acerto (acurácia) junto aos dados de validação em função do número de neurônios na camada oculta. Por simplicidade, adote o esquema *holdout* (com estratificação) para a validação cruzada.
Descreva com clareza todas as escolhas feitas para os hiperparâmetros e demais elementos do problema. Pensem na pertinência de realizar algum pré-processamento nos dados (e.g., normalização).

c) Utilizando o número “ótimo” de neurônios na camada intermediária, treine novamente a rede MLP e apresente a matriz de confusão em relação aos dados de validação, bem como as curvas de evolução da função custo ao longo das épocas, tanto para os dados de treinamento quanto para os dados de validação. Discuta os resultados obtidos.

[*] Esse índice fornece uma estimativa da chance de o paciente ter diabetes com base no histórico familiar.

Parte 2 – MNIST, MLP e CNN

Problema: reconhecimento de dígitos numéricos manuscritos
Base de dados: MNIST

Neste exercício, vamos trabalhar com a famosa base de dados MNIST (*Modified National Institute of Standards and Technology*), com imagens de dígitos numéricos manuscritos, com dimensão 28×28, em níveis de cinza. O conjunto de dados possui 60.000 imagens para treinamento e 10.000 imagens para teste.

Ao carregar a base de dados (seja em Python ou no Matlab), lembre-se de transformar as intensidades dos *pixels* das imagens em valores reais entre 0 e 1, através da divisão por 255.

a) Aplique uma rede MLP com uma, duas e três camadas intermediárias e analise (1) a acurácia e (2) a matriz de confusão para os dados de teste obtidas por estas três redes. Descreva a metodologia e as arquiteturas empregadas, bem como todas as escolhas feitas.

b) Monte uma CNN simples contendo: (*i*) uma camada convolucional; (*ii*) uma camada de *pooling*; (*iii*) uma camada de saída do tipo *softmax*. Avalie a progressão da acurácia junto aos dados de teste em função:
   1) Da quantidade de *kernels* utilizados na camada convolucional;
   2) Do tamanho do *kernel* de convolução.

c) Escolhendo, então, a melhor configuração para a CNN simples, refaça o treinamento do modelo e apresente:
   - A matriz de confusão para os dados de teste;
   - A acurácia global;
   - Cinco padrões de teste que foram classificados incorretamente, indicando a classe esperada e a classe estimada pela rede.

   Discuta os resultados obtidos.

d) A partir da CNN básica do item anterior, tente aprimorar o desempenho lançando mão de uma CNN um pouco mais profunda (até três camadas convolucionais, no máximo). Descreva a arquitetura utilizada e apresente os mesmos resultados solicitados no item c) para o conjunto de teste.
   Faça uma breve comparação entre os modelos estudados neste exercício.