TP 1

Redes feedforward multicapa

Redes Neuronales

Aprendizaje supervisado

DEPARTAMENTO DE COMPUTACION

1

Introducción

En este documento se detalla el Trabajo Práctico 1 de la materia Redes Neuronales.

El objetivo de este trabajo es que, dados problemas del mundo real, los alumnos puedan implementar y experimentar soluciones a los mismos, utilizando redes neuronales artificiales bajo el paradigma del aprendizaje supervisado. Los datos de los problemas pertenecen a problemas reales y es posible que para que puedan ser utilizados en forma correcta deban ser preprocesados. Se espera que basándose en los conjuntos de datos se construyan modelos de redes neuronales feedforward multicapa que tengan una capacidad aceptable de generalización.

1.1. Consignas

1. Se recomienda implementar el TP en Python. En el caso de desear implementarlo en otro lenguaje de programación, deberán solicitar autorización previa a los docentes.
2. El avance científico es fruto de la investigación personal, y de la discusión entre los pares. No está permitido compartir código fuente entre los grupos, pero sí se fomenta la discusión e interacción dentro del grupo, y entre los distintos grupos (i.e., está permitido probar experimentalmente ideas que les funcionaron bien a otros grupos, citando la fuente "Sugerencia de los colegas del Grupo X").
3. Se deberán partir los datasets de los problemas en conjuntos de entrenamiento, validación y testing, en proporciones razonables.
4. Para mostrar la evolución del error, se deberán presentar en el informe los gráficos de error vs. cantidad de épocas, de los conjuntos anteriores.
5. Se deberán realizar al menos los siguientes experimentos, sacar conclusiones y reportarlas en el informe:
   (a) Variación del número de capas ocultas.
   (b) Variación del número de neuronas ocultas.
   (c) Performance de la red, con entrenamiento sin momentum y con momentum.
   (d) Performance de la red, con entrenamiento sin y con parámetros adaptativos.
   (e) Performance de la red, con entrenamiento estocástico, batch y mini-batch.
   (f) Performance de la red, variando simultáneamente el factor de aprendizaje μ, y el parámetro α del momentum.
   (g) Performance de la red, con distintas técnicas de inicialización de los pesos de la red.
   (h) Performance de la red, sin y con preprocesamiento de los patrones.
   (i) Performance de la red, sin y con early-stopping.
   (j) Performance de la red, variando las funciones de activación y/o sus parámetros.

Los items anteriores pueden ser combinados (por ejemplo, entrenamiento con early stopping y con momentum), o agregar otros experimentos que considere interesantes y reportar los resultados. En base a los experimentos anteriores, se justificará la arquitectura y método elegido para la solución óptima propuesta. De la solución óptima, se debe reportar indefectiblemente, el error final de los conjuntos de entrenamiento, validación y testing.

1.2. Detalles de la entrega

El programa deberá permitir seleccionar un conjunto de datos, y tener la opción de trabajar con una red neuronal nueva o ya entrenada. Si se utiliza una red nueva deberá ser posible entrenarla con los datos seleccionados más los parámetros necesarios. Si se utiliza una red ya entrenada se deberán poder utilizar los datos seleccionados para testearla. En ambos casos el programa deberá mostrar de forma clara el desempeño del modelo.

En caso de que existan dificultades en la ejecución del programa para su evaluación, el trabajo puede llegar a ser rechazado. Por esto se recomienda enfáticamente utilizar Python.

El informe deberá ser conciso y podrá contener los siguientes tópicos, además de los que consideren necesarios para que sea claro:

- Breve introducción al problema.
- Detalle de las opciones que acepta el programa y su modo de uso.
- Detalle de los requerimientos computacionales para ejecutar el programa.
- Resultados experimentales de la sección 1.1 *"Consignas"*.
- Descripción, justificación y performance, de la solución óptima propuesta.
- Decisiones tomadas y su justificación.
- Conclusiones.

El informe debe entregarse impreso y además en un archivo comprimido junto al código fuente a la dirección: entregas.redneu@gmail.com

2

Problemas

Para los siguientes problemas de aprendizaje supervisado, se deben realizar los experimentos solicitados en 1.1, y proponer una arquitectura de redes neuronales multicapa óptima.

2.1. Diagnóstico de cáncer de mamas

Este conjunto de datos contiene los resultados de un examen específico que es utilizado en el diagnóstico del cáncer de mamas. Cada entrada corresponde a los datos obtenidos para distintos pacientes y contiene 10 características provenientes de imágenes digitalizadas de muestras de células. Junto con estas características se encuentra también el diagnóstico final, determinado junto con otras pruebas, en donde se indica si la muestra analizada pertenecía a un tumor maligno o benigno. La información de cada una de estas características es obtenida en valores reales a partir de algunos atributos como los que se detallan a continuación:

1. Radio (media de la distancia desde el centro a los puntos de perímetro)
2. Textura (desviación estándar de los valores en escala de gris)
3. Perímetro
4. Área
5. Suavidad (variaciones locales en la longitud del radio)
6. Compacidad (perímetro^2/ área - 1)
7. Concavidad (severidad de las porciones cóncavas del contorno)
8. Puntos cóncavos (proporción de porciones cóncavas del contorno)
9. Simetría
10. Densidad
11. *Respuesta:* Diagnóstico (M = maligno, B = benigno)

2.2. Eficiencia energética

Este problema consiste en determinar los requerimientos de carga energética para calefaccionar y refrigerar edificios en función de ciertas características de los mismos. El análisis energético se realizó utilizando edificios de distintas formas que difieren con respecto a la superficie y distribución de las áreas de reflejo, la orientación y otros parámetros. Cada entrada en el conjunto de datos corresponde a las características de un edificio distinto junto a dos valores reales que representan la cantidad de energía necesaria para realizar una calefacción y refrigeración adecuadas. El conjunto de datos contiene 8 atributos y 2 respuestas que se especifican a continuación:

1. Compacidad Relativa
2. Área de la Superficie Total
3. Área de las Paredes
4. Área del Techo
5. Altura Total
6. Orientación
7. Área de Reflejo Total
8. Distribución del Área de Reflejo
9. *Respuesta:* Carga de Calefacción
10. *Respuesta:* Carga de Refrigeración

2.3. Reconocimiento de dígitos manuscritos *(Opcional: bonus)*

Este ejercicio es de entrega opcional. En el caso de entregarlo, sirve como un *"bonus"*, por si se hiciese algo incorrecto en los ejercicios obligatorios.

La base de datos MNIST ("Modified National Institute of Standards and Technology") es una base de datos de dígitos manuscritos. Consiste de imágenes de números del 0 al 9, con su correspondiente etiqueta. El problema consiste en identificar el número, a partir de la imagen.

La base de datos se encuentra disponible para su descarga en:
https://www.kaggle.com/c/digit-recognizer

Opcional del opcional (for the glory!): Si lo desean, los integrantes del grupo se podrán transformar en *"Kagglers"*, submitir los resultados obtenidos a *www.kaggle.com*, y reportar en el informe el ranking obtenido en la competencia.