Lineamientos para la primera evaluación del curso aprendizaje automático

O. L. Quintero

14 de septiembre de 2019

Resumen

Este documento contiene los lineamientos para la presentación de la actividad evaluativa del 30 % a entregar el dia 18 de Octubre a las 12 del día GMT-5.

1. Introducción

El objetivo del curso es proveer elementos teóricos y conceptuales que le permitan a los estudiantes de Aprendizaje Automático de la Maestría en Ciencia de Datos, enfrentar el problema de construir un modelo compacto (learning machine) que permita representar fenómenos del mundo real.

Consecuentemente, los principios de teoría de aprendizaje fueron adelantados en la primera sesión. Se debe cuestionar y NO desviar la tarea de aprendizaje automático, es decir no "presuponer"la naturaleza del mismo. Debe explorarse la construcción de diversos modelos mediante la aplicación de los conceptos.

Teniendo en cuenta que los estudiantes de la Maestría en Ciencia de los Datos y Analítica tienen diferentes perfiles y fundamentación, el curso pretende adelantar la correcta aplicación de conceptos que permiten construir el modelo, evaluarlo y juzgar con perspectiva científica su desempeño.

Esta actividad evaluativa consiste en entrenar redes neuronales artificiales y redes convolucionales usando las capacidades computacionales de la Universidad EAFIT

1. Solicitar cita al centro de computacion cientifica APOLO y notificar que es estudiante de la MCD en el curso de Aprendizaje Automatico
2. Realizar el ejercicio de redes neuronales
3. Realizar el ejercicio de redes convolucionales
4. Realizar el informe

Los conceptos generales se pueden revisar directamente del libro "Machine Intelligence for decision making"(en borrador para uso de los estudiantes de este

curso y bajo edicion por Springer), y de las diversas fuentes citadas en el libro con artículos científicos y otros libros mas especializados en cada tema.

Para comenzar a realizar su proceso de aprendizaje (me refiero a practicar en datos juguete antes de abordar el problema real), el estudiante puede usar conjuntos de datos sintéticos que haya como ejemplo en cualquier programa o suite. Se sugiere Orange3 como elemento de práctica para aprehender los conceptos y realizar ejercicios del proceso que se indicara a continuación primero con los datos de juguete.

2. Contenidos

El estudiante debe seleccionar un problema relevante, ojalá el problema que está supuesto a trabajar en proyecto integrador y realizar este ejercicio de manera INDIVIDUAL.

El estudiante puede usar cualquier tipo de programa o suite compatible con EL CENTRO DE COMPUTACIÓN CIENTIFICA APOLO para realizar el ejercicio que se detallará en la siguiente sección.

En el ejercicio anterior, los estudiantes abordaron el entrenamiento de modelos de aprendizaje siguiendo los siguientes elementos:

1. Error real deseado
2. Error de entrenamiento deseado
3. Garantia probable de aprendizaje
4. Tamaño óptimo de la muestra
5. Dividir el conjunto de muestra en: entrenamiento, validación y prueba.
6. Visualizar los datos con los algoritmos de visualización por defecto de la suite, extraer características si hace falta para aumentar la dimensionalidad
7. Visualizar los datos con el embebimiento BH tsne https://lvdmaaten.github.io/tsne/
8. Entrenar el modelo con los datos en altas dimensiones (espacio original)
9. Entrenar el modelo con los datos en dimensiones reducidas (luego del embebimiento del paso 7)
10. Realizar la comparación del modelo del paso 8 contra el modelo del paso 9

En esta ocasión, dado que no existe la posibilidad de definir garantias de aprendizaje, el ejercicio evaluativo consiste en:

- Seleccionar el mismo problema que abordó en la práctica pasada

- Dividir el conjunto de muestra en: entrenamiento, validación y prueba.
- Entrenar (ver 2.1) el modelo con los datos en altas dimensiones (espacio original)
- Entrenar (ver 2.1) el modelo con los datos en dimensiones reducidas (luego del embebimiento)
- Realizar la comparación del modelo de dimensiones completas contra el modelo de dimensiones reducidas
- Entrenar la red profunda InceptionV3 sobre Imagenet (ver 2.2)

2.1. Entrenamiento de la red neuronal

Para su problema con m entradas, n salidas y N cantidad de instancias (patrones o tamaño de la muestra).

Seleccionar una red neuronal en configuracion perceptron multicapa con m entradas, n salidas.

Realizar el entrenamiento del perceptron multicapa cn la combinatoria de arquitecturas de la siguiente manera:

- Con L capas ocultas, donde L = 1,...,10.
- Variando el numero de neuronas de las capas ocultas l_i = 1,...,10

De esta manera, usted verificará el proceso de aprendizaje de la red neuronal a medida que va aumentando el número de pesos que debe estimar, usando el mismo tamaño de la muestra N.

Como podra observar, lo que se solicita es que entrene perceptrones multicapa de menor a mayor tamaño y profundidad desde una red mínima (m entradas, 1 neurona en la capa oculta y n salidas), hasta una red de gran tamaño (m entradas, 10 capas ocultas, 10 neuronas en la capa oculta y n salidas).

Este entrenamiento se realizará variando la tasa de aprendizaje en tres valores: 0.2, 0.5 y 0.9.

Dado que no todos los estudiantes van a trabajar en la misma libería, se sugiere no modificar los parametros de los regularizadores sino dejar los valores por defecto.

El entrenamiento debe realizarse usando un criterio de parada de error tolerancia 1e-2 y numero maximo de epocas de entrenamiento de 50.

Debe observar las curvas de entrenamiento de cada una de las redes y además evaluar el desempeño de las mismas usando la sensitividad, especificidad y capacidad de predicción en una curva ROC.

Note que algunos de los problemas de los estudiantes son de clasificación, pero otros son de ajustes de curva.

Debe comparar los resultados obtenidos con los modelos del trabajo numero 1.

2.2. Entrenamiento de la red profunda Inception v3

Los estudiantes deben hacer el ejercicio de entrenamiento de la arquitectura Inception v3 de google EN EL SUPER COMPUTADOR APOLO.

"*Se debe entrenar al modelo antes de utilizarlo para reconocer imágenes. Generalmente, el entrenamiento se realiza a través del aprendizaje supervisado con un conjunto extenso de imágenes etiquetadas. Aunque Inception v3 se puede entrenar a partir de conjuntos de imágenes etiquetadas diferentes, ImageNet es un conjunto de datos común a elección*"

(Pagina de google: https://cloud.google.com/tpu/docs/inception-v3-advanced?hl=es-419)

Conjunto de datos de Imagenet en el siguiente articulo: http://www.image-net.org/papers/imagenet$_c vpr$09.*pdf*

Antes de enviar sus codigos a APOLO por favor validar que estan funcionando en el ambiente google collab con el optimizador de gradiente descendiente estocástico.

2.3. Entregables

Se debe entregar:

- Documento informe
- Codigos elaborados

Por favor asistir a asesoría con el profesor del curso en la oficina Bloque 38-434, solicitar su cita con anticipación al correo oquinte1@eafit.edu.co.

El dia 23 de septiembre y para la semana del 30 de septiembre al 4 de octubre.