Introducción al Aprendizaje de Maquina
Guion del video final

Daniel Felipe Quiñones Ordoñez

August 2021

1 Guion del vídeo

Actualmente hemos visto cómo el aprendizaje profundo o deep learning ha tenido gran exito y acogida en multiples disciplinas y aplicaciones. Desde ayudar a médicos expertos a hacer diagnósticos témpranos de cáncer u otras enfermedades, ayudar a mitigar amenzas de ciberseguridad y seguridad en general, a aplicaciones mas cotidianas como mejores recomendaciones en nuestras plataformas de streaming favoritas y mejoramiento de los procesadores de texto.

En gran medida esto se debe al mejoramiento del hardware que permite procesar grandes volumenes de datos. Con la mejora de las GPUs o servicios en la nube o de servidores que facilitan la realizacion de experimentos hemos alcanzado un "boom" en la investigación y desarrollo de modelos basados en el aprendizaje profundo.

Sin embargo, a pesar de esto, sigue persistiendo un problema que ha sido un reto para grandes mentes de nuestro tiempo. El black box o la caja negra, como comunmente se le llama al proceso que utilizan las redes neuronales profundas para poder aprender y hacer predecciones sobre los datos, es es un misterio para la comunidad científica, en cuanto a que no se comprende la tasa de efectividad al aprender que tienen las redes, teniendo en cuenta la teoria que se ha desarrollado en cuanto a modelos de aprendizaje de máquina.

Me explico mejor: al hablar de generalizacion sobre los datos, nosotros vimos que a partir de un conjunto de entradas conocidas o muestras sobre un dataset, nosotros podriamos saber con cierta probabilidad la clasificacion o la salida sobre alguna entrada nueva que no conociamos previamente. De esta forma entendimos el concepto de aprendizaje.

Esto nos llevo a mirar el error que se podría producir al hacer estas clasificaciones y/o predicciones sobre el conjunto de datos. Asi dimos con unas cotas para el error que se podria producir, y que dependia de algunos parametros como el tamaño del numero de muestras que conociamos, la dimension VC, o el numero de hipotesis para poder aprender.

Ahora bien, ¿qué pasa con las redes neuronales profundas? Que en teoría estas no deberían poder generalizar sobre los datos y predecir sobre ellos porque

rompen estas cotas y la teoria que se ha desarrollado para esto. De ahí la incomodidad que se tiene dentro de la comunidad por entender qué sucede dentro de estas redes y cómo es que pueden aprender con tanta eficacia.

Por eso miremos brevemente qué es una red neuronal, y cómo funcionan. Con esto en mente buscaremos mostrar un poco uno de muchos enfoques que científicos han usado para poder aclarar esta caja negra y es mirar qué ocurre dentro de las capas de la red neuronal.

Una red nueronal profunda es un modelo matemático basado en las redes neuronales biológicas. En otras palabras esta inspirado en nuestros cerebros.

Nosotros vimos el modelo del perceptron que simulaba el funcionamiento de una sola neurona. Basicamente teniamos una funcion que determinaba una salida binaria dependiendo si las entradas superaban una cota establecida. En las redes neuronales tenemos algo similar: tenemos un conjunto de neuronas conectadas entre sí, y distribuidas en capas o layers. Fundamentalmente tenemos una entrada, una salida y en medio de estas los layers, o tambien conocidos como hidden algunas veces; y adicionalmente a estos elementos existen las observaciones sobre cada una de las capas.

Expliquemos un poco mejor esto: La entrada podemos pensarla como el conjunto de parametros de nuestro modelo.Esto puede ser algo sencillo como altura, base, y peso; o tener millones de parametros de input. Las salidas por el contrario suelen ser binarias, aunque dependiendo del problema pueden ser de mas variables. Para los propositos de esta presentacion consideraremos una salida binaria unicamente. En los layers tenemos un conjunto de neuronas. Estas neuronas estan conectadas a todas las neuronas del layer anterior y del siguiente.

Nosotros queremos que nuestra red aprenda por lo que es necesario una matriz de pesos que vaya ajustando nuestros datos de entrada. Estos ajustes pueden depender tambien de una funcion de complejidad que dicta qué tan conectados pueden estar unos nodos con otros dentro de la red.

De esta manera podemos hablar tambien de un proceso de entrenamiento y otro de prueba. Donde lo que queremos es que en la prueba el vector resultante se aproxime lo máximo que se pueda al vector de entrenamiento, sin que los datos de prueba esten dentro de los datos de entrenamiento.

Similar a lo que mencionamos previamente para poder aprender, y en otras palabras, queremos minimizar el error que hay entre el vector de salida de la prueba con el de entrenamiento. Para esto teniendo en cuenta que tenemos en las entradas de nuestras matrices una funcion de varias variables, se utiliza el vector gradiente y su caracteristica de tener la dirección máxima sobre el espacio, para elegir su opuesto y asi minimizar nuestro error.

Ahora consideremos un modelo en el cual tenemos una red neuronal maestra y una estudiante, o lo que es lo mismo, una red de entrenamiento y otra de prueba. La red maestra introduce ruido en los datos por lo que tenemos una matriz de pesos que aproxima lo datos de salida y va a condicionar asi la forma de aprendizaje de la red estudiante. La formula para los errores de cada red es como se muestra en la figura.

(Introducir fórmula para el error de entrenamiento y de prueba).

Adicional a esto consideremos que cada una de las variables dependen de una variable t, o variable del tiempo. Asi podemos pensar en la dinamica del sistema a traves del tiempo, y ver las curvas de aprendizaje teniendo en cuenta esta variable.

Cuando pensamos en la dinamica de un modelo o de un sistema, es importante ver las condiciones iniciales o la forma en la que inicializamos las iteraciones. En las redes neuronales es común tomar valores aleatorios en la matriz de pesos como valores iniciales. Sin embargo invetigadores como Lampinen y Ganguli consideran una condicion prticular llamada condición de entrenamiento alineado que puede ser beneficioza al momento de analizar la dinámica de aprendizaje de las redes.

La condicion de entrenamiento alineado hace que podamos descomponer la matriz de pesos de una forma especial. Su descomposición de valor singular se elige de manera que tenga la siguiente forma

$$W = \epsilon U V^t \tag{1}$$

Aqui los vectores singulares de W no cambian, mientras que sus valores singulares evolucionan de acuerdo a la siguiente fórmula.

(Fórmula 9 del artículo)

Este tipo de condicionamiento inicial ha mostrado una solucion analítica cercana a la dinamica de aprendizaje que tienen las redes estudiantes o de prueba, y no solo para el error de entrenamiento, sino tambien para el error de generalización como muestran los resultados obtenidos por Lampinen y Ganguli.

Los mismos autores tambien han llegado a un algoritmo que mejora el método del gradiente descendiente y alcanza una cota mejor para el error de generalización, cuyos detalles pueden encontrarse en la referencia y el link adjunto en la descripción.

Las conclusiones de estos resultados llevan a que el desarrollo teorico de el entrenamiento alineado, da una solucion analitica a la dinamica de aprendizaje, lo que reduce la incertidumbre de el aprendizaje de las redes neuronales profundas. Adicional a esto, pone la mira en la estructura de las tareas, es decir como los layers transfieren información el uno al otro, y no en el tamaño de la red como se podría pensar. Es de resaltar también que con los resultados obtenidos se concluye que las redes aprenden las estructuras principales dentro de los datos.

También como conclusión a estas observaciones vemos que es importante el desarrollo de más herramientas teoricas para poder comprender de forma general el aprendizaje en las redes neuronales profundas. Las ventajas mencionadas anteriormente son bajo el desarrollo de una teoria particular que considera ciertas condiciones para los modelos de las redes y que permitieron dar con una solución analítica concreta, y adicional, a la reducción de un problema tan complejo como es la no linealidad dentro del modelo.

Esto con el fin de llegar a mejores soluciones y generar confianza entre los muchos investigadores y consumidores de los modelos de redes neuronales, que tantos beneficios nos han traido hasta el momento.