GUION DE PRESENTACIÓN

Aprendizaje de Máquina (2021-I)

Jesús Andrés Rojas Montenegro

(Lo que es de este color, finalmente no se metió al video)

Intoducción.

Las personas que han tratado de expandir el conocimiento lo han hecho por medio de descubrimientos totalmente empíricos teniendo cómo brújula su curiosidad. Justamente, mucho de lo que querían hacer era tratar de ser capaces de abstraer lo suficiente sus problemas para poder enfrentar cosas similares en un futuro. A esto lo llamamos generalizar y es lo que se suele hacer en aprendizaje de máquina. Sin embargo, algunos investigadores dejándose llevar esta misma curiosidad encontraron que el mismo hecho de cómo se ha venido planteando la generalización necesita ser reevaluado.

¿Cuál es el problema general abordado?

Así, el artículo del que voy a hablar tratará de responder qué es lo que distingue a las redes neuronales que generalizan bien de las que no. Para hacer esto los autores hicieron varios experimentos y comprobaron que los enfoques tradicionales no explican por qué las redes neuronales generalizan bien en la práctica.

¿Cómo se relaciona con las temáticas cubiertas en el curso?

Formalmente, como se vio en el curso, lo que se quiere en el aprendizaje de máquina es minimizar la diferencia entre la proporción de equivocaciones en los datos que se usaron para entrenar al modelo y en los que se usaron para probar su efectividad, en otras palabras, es reducir el error de generalización del modelo producido.

Además del concepto fundamental de generalización, hay otro que es necesario conocer y es la regularización. Entonces, el problema de aprendizaje puede ser visto como un problema de optimización, digamos minimizar los errores que comete el modelo, y para este tipo de problemas existen diferentes métodos iterativos para llegar a una solución. Ahora bien, en la práctica queremos el mínimo número de iteraciones posibles y es ahí dónde entra la regularización, son técnicas para agilizar la convergencia. Por ejemplo, restar en cada paso de la iteración un término para forzar la convergencia o modificar el contraste y brillo de las imágenes para resaltar ciertas características importantes.

Estado del arte. Revisión corta de que se sabe respecto al problema y/o como se ha abordado previamente.

Ahora bien, en trabajos previos por **Barlet y Hardt** se analiza el poder de representación de las redes neuronales y los teoremas de aproximación universal para los perceptrones y se ha intentado ver la capacidad de generalización de las redes neuronales dando algunas cotas al error de generalización. Además, también hay un trabajo de **Neyshabur** y algunos colegas suyos que mediante experimentos concluyó que el tamaño de la red no es la principal forma de control de la capacidad para las redes neuronales.

¿Cuál es el problema específico de aprendizaje de maquina abordado?

Así, en el artículo los autores toman el caso particular de las redes neuronales profundas y ejecutan sus experimentos con diferentes arquitecturas y usando las bases de datos de imágenes **CIFAR10** e **ImageNet** para cuestionar la visión tradicional de la generalización.

Estrategias propuestas y/o resultados principales. Se deberá explicar de forma clara y sencilla las estrategias propuestas para el abordaje del problema y/o los resultados principales relacionados con el problema.

Concretamente, el artículo empieza mostrando sus experimentos para analizar cómo afecta la aleatorización en el entrenamiento de un modelo: en la primera prueba usan un conjunto de datos tal y como está, en la siguiente se corrompen parcialmente las etiquetas escogiendo al azar cierto número de datos a los que se les cambia su etiqueta. En el tercer experimento la etiqueta de cada imagen se elige aleatoriamente. En el cuarto se intercambian los píxeles de cada imagen por medio de una permutación específica. En el quinto se intercambian los píxeles aleatoriamente. Finalmente, en el último se añade ruido gaussiano a los pixeles de las imágenes.

Obviamente en la gráfica de los resultados se ve que entre menos cambios más rápido se encuentra una solución. Pero, lo realmente sorprendente es que todos hayan podido llegar a una solución ya que cuando se aleatorizan las etiquetas se está destruyendo cualquier relación entre la imagen y la etiqueta. Las otras dos gráficas describen, según, que tanto se hayan modificado las etiquetas, cuanto tiempo demorará en sobrecargarse el modelo de aprendizaje y la proporción de error de la arquitectura de la red, en ambos casos vemos que converge, lo que es aún más sorprendente.

Así que podemos concluir que las redes neuronales profundas pueden adaptarse a la aleatorización de las etiquetas, lo que teóricamente tiene implicaciones, porque con la medida de complejidad de Rademacher, que analiza la capacidad del conjunto de hipótesis para ajustarse a las asignaciones aleatorias de etiquetas binarias al azar, resulta que es aproximadamente 1 para el caso de etiquetas aleatorias, lo que no se ajusta a lo que en realidad está pasando porque vemos que estamos llegando a una clasificación. También, pasa que muchas otras medidas tradicionales no tienen en cuenta las particularidades de los datos o la distribución de las etiquetas como la dimensión VC.

Luego, en otra tanda de experimentos, se quiso ver el rol que juega la regularización en la generalización comparando la diferencia cuando se activaban o no esos mecanismos de arquitecturas como **Inception, AlexNet o MLP** para un etiquetado aleatorio. Entonces, se compararon las técnicas de **Data Augmetation**, que básicamente es aumentar el conjunto de entrenamiento, por ejemplo, para las imágenes, sería perturbar el brillo o el contraste en algunas zonas para diferenciarlas. Otra técnica fue el **Weight Decay**, que es lo que había dicho antes de forzar la convergencia añadiendo términos en cada paso iterativo o con la técnica de **Dropout**, que es eliminar aleatoriamente neuronas de la red para quitar la dependencia a neuronas individuales y evitar redundancia.

Entonces, con eso en mente, los resultados se resumen en la gráfica de la exactitud del modelo generado según el número de pasos de entrenamiento, las líneas claras son los resultados en los datos de entrenamiento y las oscuras en los de prueba. La otra grafica muestra los resultados con y sin **Batch Normalization**, un operador que normaliza las respuestas de la red neuronal muy usado en la arquitectura **Inception**, y se puede ver que esta técnica ayuda a reducir el ruido y aumenta la estabilidad del aprendizaje, pero tampoco ayuda demasiado para generalizar. Sin embargo, en la gráfica se ve que, si se para el entrenamiento antes, si puede haber una mejora en la generalización

Así, para este caso, la regularización puede mejorar el rendimiento de la generalización, pero no es necesaria ni suficiente para controlar el error y vemos que es poco probable que juegue un papel fundamental y hasta ahora, lo que más ha marcado la diferencia es la arquitectura de red que se escoja. (poner foto de gatos para DA)

Luego, los autores terminan con un aporte teórico que muestra que las redes neuronales grandes pueden expresar cualquier etiquetado de los datos de entrenamiento probando el siguiente teorema. (poner foto de teorema) De hecho, exhibieron una red neuronal profunda simple de dos capas con 2n + d parámetros que puede expresar el etiquetado de cualquier muestra de tamaño n en d dimensiones. Concentrándose así en una muestra finita y no infinita como lo han intentado varios autores.

Finalmente, el articulo acaba con una sección dónde abarcan el hecho de que, aunque un modelo pueda lograr clasificar todos los datos, no significa que lo haga bien. Así que, apelan a los modelos lineales y ven cómo el descenso estocástico del gradiente actúa como un regularizador implícito ya que para estos modelos usar el método del descenso estocástico del gradiente resulta siempre en una convergencia. Lo que puede estar abriendo el camino para saber porqué hay arquitecturas de red neuronal mejores que otras.

¿Por qué este problema es importante en aprendizaje de maquina?

Así las cosas, es crucial lo que se ha hecho en este artículo porque, además de poner un precedente en nuevas formas de experimentar con redes neuronales, funciona como una advertencia de que hay que dar un paso atrás y revisar todo lo que se ha hecho, ya que por mucho tiempo se pensó que una al tener una gran cantidad de parámetros para un modelo había limitar el aprendizaje por medio de regularizadores para garantizar un resultado y una buena generalización. Todo lo contrario, lo que más resalta aquí es que la regularización no es estrictamente necesaria y que los modelos son capaces de adaptar una clasificación para un etiquetado completamente aleatorio.

¿Qué aplicaciones puede tener la solución propuesta?

Ahora bien, para hablar de aplicaciones, basta con ver que el aprendizaje profundo es algo que está inundando a la sociedad, por ejemplo, con reconocimiento el de imagen y de voz o la conducción autónoma son aplicaciones con muchísimo potencial, y cabe recordar que uno de los pilares del aprendizaje de máquina es la buena generalización para poder garantizar que, ante casos de prueba desconocidos, el modelo va a funcionar correctamente y si no, necesitamos poder de alguna manera controlar el potencial error.

¿Qué problemas abiertos quedan?

Sin embargo, en el artículo queda el interrogante de qué es lo que en realidad garantiza una buena generalización y de cómo encontrar una buena medida formal y precisa de la complejidad de un conjunto de datos de prueba, que pueda abarcar todos los posibles cambios que se mostraron. Además, también está el problema de fondo de cómo en entender completamente el funcionamiento del aprendizaje de máquina para un caso general y no con muestras finitas. El trabajo hecho aquí genera más interrogantes de las que soluciona, pero el hecho de que ya se pueda ver en qué posiblemente estemos fallando es algo realmente emocionante.