

How to avoid machine learning pitfalls: a guide for academic researchers

Este documento ofrece una breve descripción de algunos de los errores comunes que suelen ocurrir al utilizar el aprendizaje automático y cómo evitarlos. Aunque está diseñado para ser comprensible para cualquier persona con un conocimiento básico de técnicas de aprendizaje automático, originalmente fue escrito para estudiantes de investigación y se centra en cuestiones de particular importancia en la investigación académica, como la necesidad de realizar comparaciones rigurosas y llegar a conclusiones válidas. El documento aborda cinco etapas del proceso de aprendizaje automático: qué hacer antes de construir un modelo, cómo construir modelos de manera confiable, cómo evaluar modelos de manera robusta, cómo comparar modelos de manera justa y cómo informar los resultados.

Introducción

Es fácil cometer errores al aplicar el aprendizaje automático (AA), y estos errores pueden resultar en modelos de AA que no funcionan como se esperaba cuando se aplican a datos que no se vieron durante el entrenamiento y la prueba [Liao et al., 2021]. Esto es un problema tanto para los profesionales, ya que conduce al fracaso de proyectos de AA, como para la sociedad, ya que socava la confianza en los hallazgos y productos de AA [Gibney, 2022]. Esta guía tiene como objetivo ayudar a los recién llegados a evitar algunos de estos errores. Está escrita por un académico y se centra en lecciones aprendidas al realizar investigaciones de AA en el ámbito académico. Aunque está dirigida principalmente a estudiantes e investigadores científicos, debería ser accesible para cualquier persona que comienza en AA y solo asume un conocimiento básico de técnicas de AA. Sin embargo, a diferencia de guías similares dirigidas a una audiencia más general, incluye temas que son de particular preocupación para la academia, como la necesidad de evaluar y comparar modelos de manera rigurosa para publicar trabajos. Para hacerlo más legible, la guía está escrita de manera informal, en un estilo de "Hacer y No Hacer". No pretende ser exhaustiva y proporciona referencias (con URL públicas cuando están disponibles) para lecturas adicionales. Dado que no aborda cuestiones específicas de campos académicos particulares, se recomienda consultar guías específicas del campo cuando estén disponibles (por ejemplo, Stevens et al. [2020] para medicina). Se agradece el feedback, y se espera que este documento evolucione con el tiempo. Por esta razón, si lo cita, incluya el número de versión de arXiv (actualmente v3).

Este documento ofrece consejos sobre cómo evitar errores comunes en el uso del aprendizaje automático (AA) en cinco etapas del proceso de AA. Aquí se resumen las recomendaciones de la sección "Antes de comenzar a construir modelos":

2 Antes de comenzar a construir modelos

Antes de apresurarse a entrenar y evaluar modelos, es esencial tomarse el tiempo para comprender los objetivos del proyecto, entender completamente los datos que respaldarán esos objetivos, considerar las limitaciones de los datos y conocer lo que se ha hecho

previamente en el campo de estudio. Si no se hacen estas consideraciones, es posible obtener resultados difíciles de publicar o modelos inapropiados para su propósito previsto.

2.1 Haga un esfuerzo por comprender sus datos

Asegúrese de que sus datos provengan de una fuente confiable, se hayan recopilado utilizando una metodología confiable y sean de buena calidad. Investigue la procedencia de los datos, verifique si están descritos en un artículo publicado en una fuente respetable y compruebe si los autores mencionan limitaciones de los datos. No asuma que un conjunto de datos es de buena calidad solo porque se haya utilizado en varios artículos. Realice análisis exploratorios de datos para buscar registros faltantes o inconsistentes antes de entrenar un modelo.

2.2 No examine todos sus datos detenidamente

Evite mirar detenidamente los datos de prueba en la etapa inicial de análisis exploratorio, ya que podría hacer suposiciones que limiten la generalización del modelo de una manera que no se puede verificar. Evite la filtración de información de los datos de prueba al proceso de entrenamiento.

2.3 Asegúrese de tener suficientes datos

Es fundamental tener suficientes datos para entrenar un modelo que generalice correctamente. La cantidad de datos necesaria depende de la relación señal-ruido en el conjunto de datos. Si la señal es fuerte, puede necesitar menos datos; si es débil, necesitará más. Si no puede obtener más datos, utilice la validación cruzada y técnicas de aumento de datos para mejorar la calidad de sus datos.

2.4 Hable con expertos en el dominio

Consulte a expertos en el campo de estudio para comprender los problemas a resolver, seleccionar características y modelos apropiados, y publicar en revistas especializadas.

2.5 Investigue la literatura

Realice una revisión bibliográfica antes de comenzar a trabajar en su proyecto para comprender lo que se ha hecho previamente en el campo y aprovechar el conocimiento existente.

2.6 Piense en cómo se implementará su modelo

Considere desde el principio cómo se implementará su modelo en el mundo real, teniendo en cuenta limitaciones de recursos, restricciones de tiempo y cómo se integrará en sistemas de software más amplios.

Estas recomendaciones ayudarán a evitar errores comunes y mejorar la calidad de su proyecto de aprendizaje automático.

Esta sección aborda cómo construir modelos de manera confiable en el aprendizaje automático (AA). Aquí se resumen las recomendaciones clave:

3 Construir modelos de manera confiable

Construir modelos es una parte gratificante del AA, pero es importante abordar este proceso de manera organizada para evitar experimentos desordenados y difíciles de justificar o escribir. Asegúrese de utilizar los datos correctamente y tome decisiones cuidadosas sobre la elección de modelos.

3.1 No permita que los datos de prueba se filtren en el proceso de entrenamiento

Evite que la información sobre los datos de prueba afecte la configuración, el entrenamiento o la selección de modelos. Esto garantiza que los datos de prueba sigan siendo una medida confiable de la generalización del modelo.

3.2 Pruebe una variedad de modelos diferentes

No existe un solo modelo de AA que funcione mejor en todos los casos. Pruebe varios modelos para encontrar el que funcione mejor para su problema específico. Utilice una validación en lugar de los datos de prueba para evaluar los modelos y optimice los hiperparámetros de manera adecuada.

3.3 No use modelos inapropiados

Evite aplicar modelos que no sean adecuados para su conjunto de datos. Asegúrese de que el modelo elegido coincida con el tipo de datos y el problema que está abordando. No asuma que las redes neuronales profundas son siempre la mejor solución.

3.4 Manténgase actualizado con los desarrollos recientes en el aprendizaje profundo

El aprendizaje automático es un campo en constante evolución. Si utiliza aprendizaje profundo, manténgase al tanto de los avances recientes en esta área para aprovechar las mejores prácticas y las arquitecturas más efectivas.

3.5 No asuma que el aprendizaje profundo siempre es la mejor opción

No suponga que las redes neuronales profundas son la solución óptima para todos los problemas. Evalúe cuidadosamente si otro enfoque de AA, como modelos de árboles aleatorios o máquinas de vectores de soporte, podría ser más apropiado.

3.6 Optimice los hiperparámetros de su modelo

Ajuste los hiperparámetros de su modelo de manera adecuada para obtener el mejor rendimiento. Utilice estrategias de optimización de hiperparámetros en lugar de ajustarlos manualmente.

3.7 Tenga cuidado con la selección de características y la optimización de hiperparámetros

Realice la selección de características y la optimización de hiperparámetros como parte del proceso de entrenamiento, utilizando solo los datos de entrenamiento, para evitar la filtración de información de los datos de prueba.

3.8 Evite aprender correlaciones espurias

Tenga cuidado con las correlaciones espurias en sus datos, que son relaciones aparentes pero sin significado real. Estas pueden llevar a modelos que no generalizan bien. Utilice técnicas de regularización y evaluación cuidadosa del modelo para abordar este problema.

Estas recomendaciones ayudarán a construir modelos de AA confiables y efectivos.

4 Cómo evaluar modelos de manera robusta

Para contribuir al progreso en su campo, es fundamental tener resultados válidos y poder sacar conclusiones fiables. Sin embargo, es fácil evaluar modelos de AA de manera injusta, lo que puede dificultar el avance en la investigación. Por lo tanto, es importante pensar cuidadosamente en cómo se van a utilizar los datos en sus experimentos, cómo medir el rendimiento real de sus modelos y cómo informar sobre este rendimiento de manera significativa e informativa.

4.1 Use un conjunto de prueba apropiado

Siempre use un conjunto de prueba para medir la generalización de un modelo de AA. El rendimiento de un modelo en el conjunto de entrenamiento es casi irrelevante, ya que un modelo suficientemente complejo puede aprender por completo el conjunto de entrenamiento sin capturar ningún conocimiento generalizable. También es importante asegurarse de que los datos en el conjunto de prueba sean apropiados, es decir, no deben superponerse con el conjunto de entrenamiento y deben ser representativos de la población más amplia.

4.2 No realice la ampliación de datos antes de dividir sus datos

La ampliación de datos puede ser una técnica útil para equilibrar conjuntos de datos y mejorar la generalización y robustez de los modelos de AA. Sin embargo, es importante realizar la ampliación de datos solo en el conjunto de entrenamiento y no en los datos que se utilizarán para las pruebas. Incluir datos ampliados en el conjunto de prueba puede llevar a problemas de sobreajuste y filtración de información.

4.3 Use un conjunto de validación

Cuando entrene varios modelos sucesivamente, utilice un conjunto de validación separado para medir el rendimiento en lugar del conjunto de prueba. No utilice el conjunto de prueba para guiar el proceso de entrenamiento, ya que esto puede llevar al sobreajuste y no permitirá una evaluación independiente del modelo.

4.4 Evalúe un modelo varias veces

Muchos modelos de AA son inestables, lo que significa que su rendimiento varía significativamente cuando se entrenan varias veces o cuando se realizan pequeños cambios en los datos de entrenamiento. Para obtener una evaluación confiable, realice múltiples evaluaciones utilizando diferentes subconjuntos de los datos de entrenamiento.

4.5 Guarde algunos datos para evaluar su instancia final del modelo

Es importante distinguir entre evaluar el potencial de un modelo en general y evaluar el rendimiento de una instancia específica del modelo. Si tiene suficientes datos, es mejor reservar algunos para evaluar de manera imparcial la instancia final del modelo seleccionada.

4.6 No use la precisión con conjuntos de datos desequilibrados

Tenga cuidado al elegir métricas de evaluación para modelos de clasificación en conjuntos de datos desequilibrados. La precisión puede ser engañosa en tales casos. En su lugar, considere métricas como la puntuación F1, el coeficiente kappa de Cohen o el coeficiente de correlación de Matthews, que son menos sensibles al desequilibrio de clases.

4.7 No ignore las dependencias temporales en los datos de series temporales

Los datos de series temporales son sensibles al orden de las observaciones. Ignorar esta dependencia temporal puede llevar a problemas de sesgo y sobreajuste. Utilice técnicas de

validación cruzada que respeten las dependencias temporales y evite el sesgo de mirar hacia el futuro al procesar datos de series temporales.

Estas recomendaciones ayudarán a garantizar una evaluación robusta y confiable de los modelos de aprendizaje automático.

5 Cómo comparar modelos de manera justa

Comparar modelos es la base de la investigación académica, pero puede ser sorprendentemente difícil hacerlo correctamente. Si realiza una comparación de manera injusta y la publica, otros investigadores pueden verse desviados posteriormente. Por lo tanto, asegúrese de evaluar diferentes modelos en el mismo contexto, explorar múltiples perspectivas y utilizar pruebas estadísticas adecuadas.

5.1 No suponga que un número más grande significa un mejor modelo

No es raro que un artículo afirme algo como "En investigaciones previas, se informaron precisión de hasta el 94%. Nuestro modelo logró el 95% y, por lo tanto, es mejor". Existen varias razones por las cuales un número más alto no implica necesariamente un mejor modelo. Pequeñas diferencias en el rendimiento pueden deberse a diferentes particiones del mismo conjunto de datos o al uso de conjuntos de datos completamente diferentes. Además, es importante llevar a cabo la misma cantidad de optimización de hiperparámetros al comparar modelos para que la comparación sea justa. Para una comparación justa, implemente todos los modelos desde cero, optimícelos por igual, realice múltiples evaluaciones y utilice pruebas estadísticas para determinar si las diferencias en el rendimiento son significativas.

5.2 Use pruebas estadísticas al comparar modelos

Si desea convencer a las personas de que su modelo es mejor que el de otra persona, una prueba estadística es una herramienta muy útil. Hay dos categorías principales de pruebas para comparar modelos de AA. La primera se utiliza para comparar instancias individuales del modelo, por ejemplo, dos árboles de decisiones entrenados. La segunda categoría se utiliza para comparar dos modelos de manera más general, por ejemplo, si un árbol de decisión o una red neuronal es una mejor opción para los datos. Utilice pruebas como la prueba de McNemar para comparar dos clasificadores o pruebas como la prueba U de Mann-Whitney para comparar distribuciones de rendimiento. Estas pruebas requieren múltiples evaluaciones de cada modelo, que se pueden obtener mediante validación cruzada u otros métodos de muestreo repetido.

5.3 Corrija para comparaciones múltiples

Cuando desee utilizar pruebas estadísticas para comparar más de dos modelos, tenga en cuenta que hacer múltiples pruebas puede llevar a interpretaciones excesivamente optimistas de la significación estadística. Cada vez que realice una comparación entre dos modelos con una prueba estadística, existe una probabilidad de que descubra diferencias significativas incluso cuando no las haya. Para abordar este problema, aplique una corrección para pruebas múltiples, como la corrección de Bonferroni, para reducir el umbral de significación. Esto evita la acumulación de errores tipo I en las pruebas múltiples.

5.4 No siempre confíe en los resultados de conjuntos de datos de referencia de la comunidad

El uso de conjuntos de datos de referencia para evaluar nuevos modelos de AA es común en algunos dominios. Sin embargo, esta práctica tiene sus desventajas, ya que el acceso sin restricciones a los conjuntos de prueba puede llevar a resultados sesgados. Además, el uso generalizado de un conjunto de pruebas puede aumentar la probabilidad de que el mejor modelo sea el resultado del sobreajuste al conjunto de pruebas en lugar de una generalización sólida. Sea cauteloso al interpretar los resultados de conjuntos de datos de referencia y no asuma que un pequeño aumento en el rendimiento es significativo.

5.5 Considere combinaciones de modelos

A menudo, tiene sentido utilizar combinaciones de modelos en lugar de elegir entre diferentes modelos. Las combinaciones de modelos, conocidas como ensambles, pueden compensar las debilidades de un modelo utilizando las fortalezas de otro. Los ensambles pueden formarse a partir del mismo tipo de modelo base o combinando diferentes tipos de modelos. Las estrategias de combinación varían desde votación simple hasta métodos más complejos que utilizan otro modelo de AA para combinar las salidas de los modelos base. El uso de ensambles puede mejorar el rendimiento general del modelo.

Estas recomendaciones ayudarán a comparar modelos de manera justa y evitar interpretaciones sesgadas o erróneas en la investigación de aprendizaje automático.

La presentación adecuada de resultados en la investigación académica es esencial para contribuir al conocimiento de manera efectiva. Aquí se resumen las principales recomendaciones para informar sobre los resultados de manera adecuada:

6 Cómo informar sobre sus resultados

El objetivo de la investigación académica no es el autoelogio, sino una oportunidad para contribuir al conocimiento. Para contribuir de manera efectiva, es necesario proporcionar una

imagen completa de su trabajo, cubriendo tanto lo que funcionó como lo que no. El aprendizaje automático a menudo implica compensaciones, y rara vez un modelo es mejor en todos los aspectos que importan. Debe esforzarse por reflejar esto con un enfoque matizado y considerado al informar los resultados y conclusiones.

6.1 Sea transparente

En primer lugar, siempre intente ser transparente acerca de lo que ha hecho y lo que ha descubierto, ya que esto facilitará que otras personas construyan sobre su trabajo. Es buena práctica compartir sus modelos de una manera accesible. Por ejemplo, si utilizó un guión para implementar todos sus experimentos, comparta el guión al publicar los resultados. Esto significa que otras personas pueden repetir fácilmente sus experimentos, lo que aumenta la confianza en su trabajo. También facilita la comparación de modelos, ya que otras personas no tienen que volver a implementar todo desde cero para garantizar una comparación justa. Saber que compartirá su trabajo también lo alentará a ser más cuidadoso, documentar bien sus experimentos y escribir un código limpio, lo que beneficia tanto a usted como a los demás. También es importante tener en cuenta que los problemas relacionados con la reproducibilidad están ganando importancia en la comunidad de aprendizaje automático, por lo que en el futuro es posible que no pueda publicar su trabajo a menos que su flujo de trabajo esté adecuadamente documentado y compartido.

6.2 Informe el rendimiento de múltiples maneras

Una forma de lograr una mayor rigurosidad al evaluar y comparar modelos es utilizar múltiples conjuntos de datos. Esto ayuda a superar las deficiencias asociadas con conjuntos de datos individuales y le permite presentar una imagen más completa del rendimiento de su modelo. También es una buena práctica informar múltiples métricas para cada conjunto de datos, ya que diferentes métricas pueden proporcionar diferentes perspectivas sobre los resultados y aumentar la transparencia de su trabajo. Por ejemplo, si utiliza la precisión, también es una buena idea incluir métricas que sean menos sensibles a los desequilibrios de clase. En dominios como la medicina y la seguridad, es importante saber dónde se están cometiendo errores; por ejemplo, cuando su modelo comete errores, ¿tiende más a falsos positivos o falsos negativos? Las métricas que resumen todo en un solo número, como la precisión, no ofrecen información sobre esto. Por lo tanto, es importante incluir métricas parciales como precisión y exhaustividad, o sensibilidad y especificidad, ya que estas proporcionan información sobre los tipos de errores que produce su modelo. Asegúrese de que quede claro qué métricas está utilizando.

6.3 No generalice más allá de los datos

Es importante no presentar conclusiones inválidas, ya que esto puede llevar a otros investigadores por el camino equivocado. Un error común es hacer declaraciones generales que no están respaldadas por los datos utilizados para entrenar y evaluar los modelos. Por ejemplo, si su modelo funciona muy bien en un conjunto de datos, esto no significa que funcionará bien en otros conjuntos de datos. Aunque puede obtener información más sólida utilizando múltiples conjuntos de datos, siempre habrá un límite en lo que puede inferir de cualquier estudio experimental. Existen numerosas razones para esto, muchas de las cuales están relacionadas con la forma en que se curan los conjuntos de datos. Un problema común es el sesgo o error de muestreo: que los datos no sean lo suficientemente representativos del mundo real. Otro problema es la superposición: múltiples conjuntos de datos pueden no ser independientes y pueden tener sesgos similares. También está el problema de la calidad, que es especialmente importante en conjuntos de datos de aprendizaje profundo, donde la necesidad de cantidad de datos limita la cantidad de verificación de calidad que se puede realizar. En resumen, no exagere sus hallazgos y sea consciente de sus limitaciones.

6.4 Tenga cuidado al informar la significación estadística

Ya he discutido las pruebas estadísticas (ver Use pruebas estadísticas al comparar modelos) y cómo se pueden utilizar para determinar las diferencias entre modelos de aprendizaje automático. Sin embargo, las pruebas estadísticas no son perfectas. Algunas son conservadoras y tienden a subestimar la significación, mientras que otras son liberales y tienden a sobreestimarla. Esto significa que una prueba positiva no siempre indica que algo sea significativo, y una prueba negativa no significa necesariamente que algo no sea significativo. Luego está el problema de usar un umbral para determinar la significación; por ejemplo, un umbral de confianza del 95% (es decir, cuando el valor $p < 0.05$) significa que 1 de cada 20 veces una diferencia marcada como significativa no lo será. De hecho, los estadísticos argumentan cada vez más que es mejor no utilizar umbrales y, en cambio, informar los valores p y dejar que el lector los interprete. Además de la significación estadística, otro aspecto a considerar es si la diferencia entre dos modelos es realmente importante. Si tiene suficientes muestras, siempre puede encontrar diferencias significativas, incluso cuando la diferencia real en el rendimiento es minúscula. Para dar una mejor indicación de si algo es importante, puede medir el tamaño del efecto. Existen una variedad de enfoques para esto: la estadística de Cohen es probablemente la más común, pero enfoques más sólidos, como el de Kolmogorov-Smirnov, son preferibles.

6.5 Analice sus modelos

Los modelos entrenados contienen mucha información útil. Desafortunadamente, muchos autores solo informan las métricas de rendimiento de un modelo entrenado, sin proporcionar ninguna idea sobre lo que realmente aprendió el modelo. Recuerde que el objetivo de la investigación no es obtener una precisión ligeramente mayor que la de todos los demás. Más bien, es generar conocimiento y comprensión y compartirlo con la comunidad de investigación. Si puede hacer esto, es mucho más probable que obtenga una publicación decente de su

trabajo. Entonces, analice sus modelos y trate de entender cómo toman decisiones. Para modelos relativamente simples como los árboles de decisión, también puede ser beneficioso proporcionar visualizaciones de sus modelos, y la mayoría de las bibliotecas tienen funciones que lo hacen por usted. Para modelos complejos como las redes neuronales profundas, considere utilizar técnicas de inteligencia artificial explicativa (XAI) para extraer conocimiento. Es poco probable que le digan exactamente lo que está haciendo el modelo, pero pueden proporcionar información útil.

Estas recomendaciones ayudarán a informar sobre los resultados de manera efectiva y a evitar conclusiones erróneas en la investigación de aprendizaje automático.

User

7 Final thoughts

This document doesn't tell you everything you need to know, the lessons sometimes have no firm conclusions, and some of the things I've told you might be wrong, or at least debateable. This, I'm afraid, is the nature of research. The theory of how to do ML almost always lags behind the practice, academics will always disagree about the best ways of doing things, and what we think is correct today may not be correct tomorrow. Therefore, you have to approach ML in much the same way you would any other aspect of research: with an open mind, a willingness to keep up with recent developments, and the humility to accept you don't know everything.

20

8 Acknowledgements

Thanks to everyone who gave me feedback on the draft manuscript, and to everyone who has since sent me suggestions for revisions and new content.

RESUMIR EN ESPAÑOL

ChatGPT

En resumen, este documento proporciona consejos y directrices importantes para llevar a cabo investigaciones en el campo del aprendizaje automático de manera efectiva. Se enfoca en diversos aspectos, desde la preparación de datos hasta la evaluación de modelos y la presentación de resultados. Algunas de las recomendaciones clave incluyen:

Preparación de datos: Asegúrese de comprender sus datos y realizar una limpieza adecuada. Tenga en cuenta los desequilibrios en los conjuntos de datos y aplique técnicas apropiadas para abordarlos.

Selección de características: Elija características relevantes y utilice métodos de selección de características si es necesario. Considere la normalización de datos para modelos sensibles a la escala.

División de datos: Separe sus datos en conjuntos de entrenamiento, validación y prueba de manera adecuada. Evite fugas de datos y realice una validación cruzada adecuada para obtener estimaciones confiables del rendimiento del modelo.

Entrenamiento de modelos: Experimente con diferentes algoritmos y ajuste los hiperparámetros de sus modelos. Asegúrese de entender las limitaciones de cada algoritmo y considere enfoques de aprendizaje profundo cuando sea apropiado.

Evaluación de modelos: No se base únicamente en una métrica de rendimiento, utilice varias métricas y considere la interpretación de errores. Realice pruebas estadísticas para comparar modelos de manera justa y aplique correcciones para comparaciones múltiples.

Presentación de resultados: Sea transparente en su informe y comparta sus modelos y códigos. Informe el rendimiento de múltiples maneras y evite generalizar más allá de los datos utilizados. Tenga precaución al informar sobre la significación estadística y considere el tamaño del efecto. Analice sus modelos para comprender cómo toman decisiones.

Pensamientos finales: Reconozca que la investigación en aprendizaje automático es un campo en constante evolución y que las mejores prácticas pueden cambiar con el tiempo. Mantenga una mente abierta, esté al tanto de los desarrollos recientes y sea humilde al reconocer que no lo sabe todo.

Este documento reconoce que la investigación es un proceso en constante evolución y que no todas las recomendaciones son definitivas. Sin embargo, proporciona una base sólida para llevar a cabo investigaciones de calidad en aprendizaje automático.

Agradecimientos a todos los que contribuyeron con comentarios y sugerencias para este documento.

El objetivo de la investigación académica no es el autoelogio, sino una oportunidad para contribuir al conocimiento. Para contribuir de manera efectiva, es necesario proporcionar una imagen completa de su trabajo, cubriendo tanto lo que funcionó como lo que no. El aprendizaje automático a menudo implica compensaciones, y rara vez un modelo es mejor en todos los aspectos que importan. Debe esforzarse por reflejar esto con un enfoque matizado y considerado al informar los resultados y conclusiones.

6.1 Sea transparente

En primer lugar, siempre intente ser transparente acerca de lo que ha hecho y lo que ha descubierto, ya que esto facilitará que otras personas construyan sobre su trabajo. Es buena práctica compartir sus modelos de una manera accesible. Por ejemplo, si utilizó un guión para implementar todos sus experimentos, comparta el guión al publicar los resultados. Esto significa que otras personas pueden repetir fácilmente sus experimentos, lo que aumenta la confianza en su trabajo. También facilita la comparación de modelos, ya que otras personas no tienen que volver a implementar todo desde cero para garantizar una comparación justa. Saber que compartirá su trabajo también lo alentará a ser más cuidadoso, documentar bien sus experimentos y escribir un código limpio, lo que beneficia tanto a usted como a los demás. También es importante tener en cuenta que los problemas relacionados con la reproducibilidad están ganando importancia en la comunidad de aprendizaje automático, por lo que en el futuro es posible que no pueda publicar su trabajo a menos que su flujo de trabajo esté adecuadamente documentado y compartido.

6.2 Informe el rendimiento de múltiples maneras

Una forma de lograr una mayor rigurosidad al evaluar y comparar modelos es utilizar múltiples conjuntos de datos. Esto ayuda a superar las deficiencias asociadas con conjuntos de datos individuales y le permite presentar una imagen más completa del rendimiento de su modelo. También es una buena práctica informar múltiples métricas para cada conjunto de datos, ya que diferentes métricas pueden proporcionar diferentes perspectivas sobre los resultados y aumentar la transparencia de su trabajo. Por ejemplo, si utiliza la precisión, también es una buena idea incluir métricas que sean menos sensibles a los desequilibrios de clase. En dominios como la medicina y la seguridad, es importante saber dónde se están cometiendo errores; por ejemplo, cuando su modelo comete errores, ¿tiende más a falsos positivos o falsos negativos? Las métricas que resumen todo en un solo número, como la precisión, no ofrecen información sobre esto. Por lo tanto, es importante incluir métricas parciales como precisión y exhaustividad, o sensibilidad y especificidad, ya que estas proporcionan información sobre los tipos de errores que produce su modelo. Asegúrese de que quede claro qué métricas está utilizando.

6.3 No generalice más allá de los datos

Es importante no presentar conclusiones inválidas, ya que esto puede llevar a otros investigadores por el camino equivocado. Un error común es hacer declaraciones generales que no están respaldadas por los datos utilizados para entrenar y evaluar los modelos. Por ejemplo, si su modelo funciona muy bien en un conjunto de datos, esto no significa que funcionará bien en otros conjuntos de datos. Aunque puede obtener información más sólida utilizando múltiples conjuntos de datos, siempre habrá un límite en lo que puede inferir de cualquier estudio experimental. Existen numerosas razones para esto, muchas de las cuales están relacionadas con la forma en que se curan los conjuntos de datos. Un problema común es el sesgo o error de muestreo: que los datos no sean lo suficientemente representativos del mundo real. Otro problema es la superposición: múltiples conjuntos de datos pueden no ser independientes y pueden tener sesgos similares. También está el problema de la calidad, que es especialmente importante en conjuntos de datos de aprendizaje profundo, donde la necesidad de cantidad de datos limita la cantidad de verificación de calidad que se puede realizar. En resumen, no exagere sus hallazgos y sea consciente de sus limitaciones.

6.4 Tenga cuidado al informar la significación estadística

Ya he discutido las pruebas estadísticas (ver Use pruebas estadísticas al comparar modelos) y cómo se pueden utilizar para determinar las diferencias entre modelos de aprendizaje automático. Sin embargo, las pruebas estadísticas no son perfectas. Algunas son conservadoras y tienden a subestimar la significación, mientras que otras son liberales y tienden a sobreestimarla. Esto significa que una prueba positiva no siempre indica que algo sea significativo, y una prueba negativa no significa necesariamente que algo no sea significativo. Luego está el problema de usar un umbral para determinar la significación; por ejemplo, un umbral de confianza del 95% (es decir, cuando el valor $p < 0.05$) significa que 1 de cada 20 veces una diferencia marcada como significativa no lo será. De hecho, los estadísticos argumentan cada vez más que es mejor no utilizar umbrales y, en cambio, informar los valores p y dejar que el lector los interprete. Además de la significación estadística, otro aspecto a considerar es si la diferencia entre dos modelos es realmente importante. Si tiene suficientes muestras, siempre puede encontrar diferencias significativas, incluso cuando la diferencia real en el rendimiento es minúscula. Para dar una mejor indicación de si algo es importante, puede medir el tamaño del efecto. Existen una variedad de enfoques para esto: la estadística de Cohen es probablemente la más común, pero enfoques más sólidos, como el de Kolmogorov-Smirnov, son preferibles.

6.5 Analice sus modelos

Los modelos entrenados contienen mucha información útil. Desafortunadamente, muchos autores solo informan las métricas de rendimiento de un modelo entrenado, sin proporcionar ninguna idea sobre lo que realmente aprendió el modelo. Recuerde que el objetivo de la investigación no es obtener una precisión ligeramente mayor que la de todos los demás. Más bien, es generar conocimiento y comprensión y compartirlo con la comunidad de investigación. Si puede hacer esto, es mucho más probable que obtenga una publicación decente de su

trabajo. Entonces, analice sus modelos y trate de entender cómo toman decisiones. Para modelos relativamente simples como los árboles de decisión, también puede ser beneficioso proporcionar visualizaciones de sus modelos, y la mayoría de las bibliotecas tienen funciones que lo hacen por usted. Para modelos complejos como las redes neuronales profundas, considere utilizar técnicas de inteligencia artificial explicativa (XAI) para extraer conocimiento. Es poco probable que le digan exactamente lo que está haciendo el modelo, pero pueden proporcionar información útil.

Estas recomendaciones ayudarán a informar sobre los resultados de manera efectiva y a evitar conclusiones erróneas en la investigación de aprendizaje automático.

En resumen, este documento proporciona consejos y directrices importantes para llevar a cabo investigaciones en el campo del aprendizaje automático de manera efectiva. Se enfoca en diversos aspectos, desde la preparación de datos hasta la evaluación de modelos y la presentación de resultados. Algunas de las recomendaciones clave incluyen:

Preparación de datos: Asegúrese de comprender sus datos y realizar una limpieza adecuada. Tenga en cuenta los desequilibrios en los conjuntos de datos y aplique técnicas apropiadas para abordarlos.

Selección de características: Elija características relevantes y utilice métodos de selección de características si es necesario. Considere la normalización de datos para modelos sensibles a la escala.

División de datos: Separe sus datos en conjuntos de entrenamiento, validación y prueba de manera adecuada. Evite fugas de datos y realice una validación cruzada adecuada para obtener estimaciones confiables del rendimiento del modelo.

Entrenamiento de modelos: Experimente con diferentes algoritmos y ajuste los hiperparámetros de sus modelos. Asegúrese de entender las limitaciones de cada algoritmo y considere enfoques de aprendizaje profundo cuando sea apropiado.

Evaluación de modelos: No se base únicamente en una métrica de rendimiento, utilice varias métricas y considere la interpretación de errores. Realice pruebas estadísticas para comparar modelos de manera justa y aplique correcciones para comparaciones múltiples.

Presentación de resultados: Sea transparente en su informe y comparta sus modelos y códigos. Informe el rendimiento de múltiples maneras y evite generalizar más allá de los datos utilizados. Tenga precaución al informar sobre la significación estadística y considere el tamaño del efecto. Analice sus modelos para comprender cómo toman decisiones.

Pensamientos finales: Reconozca que la investigación en aprendizaje automático es un campo en constante evolución y que las mejores prácticas pueden cambiar con el tiempo. Mantenga una mente abierta, esté al tanto de los desarrollos recientes y sea humilde al reconocer que no lo sabe todo.

Este documento reconoce que la investigación es un proceso en constante evolución y que no todas las recomendaciones son definitivas. Sin embargo, proporciona una base sólida para llevar a cabo investigaciones de calidad en aprendizaje automático.

Agradecimientos a todos los que contribuyeron con comentarios y sugerencias para este documento.