AI responsable

Felipe González, Teresa Ortiz, Roberto Sánchez

2020-04-26

Contenido

Sección 1

Acerca de este material

Este material es reproducible según instrucciones en este repositorio[1].

[1] https://github.com/felipegonzalez/ai-responsable

Sección 2

Introducción

Los métodos de aprendizaje de máquina e inteligencia artificial (ML/IA), que para resumir en este documento llamamos **aprendizaje automático** son cada vez más requeridos y utilizados por tomadores de decisiones para informar acciones o intervenciones en varios contextos, desde negocios hasta política pública. En la práctica, estos métodos se han utilizado con diversos grados de éxito, y con esto ha aparecido la preocupación creciente de cómo entender el desempeño o influencia positiva o negativa de estos métodos en el contexto de la toma de decisiones ((**?**), (**?**)).

> ¿Cuándo estos métodos de aprendizaje automático nos llevan a tomar decisiones que son costosas, riesgosas o pueden inducir estados no deseados (por ejemplo, discriminatorios o injustos) en los sistemas y decisiones en donde se utilizan?

2.1 IA y toma de decisiones

En este documento pretendemos discutir los problemas más comunes en el uso de ML/IA para la toma decisiones, cómo detectar problemas potenciales y evaluar la posiblidad de resultados indeseables para los tomadores de decisiones, para la sociedad o una compañía o institución particular. Consideramos dos arquetipos de inclusión de aprendizaje automático en el proceso de toma de decisiones:

1. **Las decisiones se toman automáticamente, sin intervención humana directa.** Por ejemplo, inclusión de individuos en programas sociales, o otorgamiento de créditos al consumo.

2. **Resultados del aprendizaje automático son insumos para decisiones que se toman de forma más tradicional.** Por ejemplo, decisiones acerca del funcionamiento o refinación de un programa social, pronósticos de variables clave para planeación.

El primer punto que tenemos que tener presente y es guía de este documento es en la construcción y métodos particulares de los sistemas de aprendizaje automático existe una gran variedad de técnicas, conocimiento experto del tema y de modelación en general, y **en este documento** no pretendemos definir cómo deben ser esos procedimientos. En lugar de eso

Principio guía 1

- Nos concentramos en la **evaluación** de estos sistemas, antes y después de poner en producción o utilizar para tomar decisiones.
- La evaluación sólo tiene sentido en términos del contexto de la decisión, y de los **resultados que son deseables para los tomadores de decisiones**, instituciones o compañías involucradas.

La evaluación de un sistema de aprendizaje no tiene sentido fuera de su contexto: qué es una tasa de error apropiado, cuáles son sesgos poco aceptables, y así sucesivamente, sólo pueden considerarse dentro de los propósitos de los tomadores decisiones. Esto quiere decir que aunque mucha de la discusión se concentra en discusiones técnicas, métricas particulares de desempeño o métodos de ML, al final ninguna de estas discusiones tiene sentido fuera del contexto de la **decisión** que se quiere tomar, ya sea una decisión automatizada totalmente o como insumo de una decisión tomada por páneles de expertos o tomadores de decisiones.

2.2 Las tres componentes de la aplicación de ML/AI

Las decisiones que consideramos tienen tres partes importantes: se considera alguna **intervención** que se planea aplicar personas, unidades o procesos. Esta intervención se pretende utilizar en ciertos casos que **dependen de variables desconocidas**. Usamos aprendizaje automático para **predecir o estimar** estas variables desconocidas.

Nuestra pregunta es entonces cómo la calidad o tipo de predicciones o estimaciones pueden influir en la focalización, y más adelante, en el resultado global de la intervención.

1. (**Intervención**) Tenemos una intervención o política o intervención que estamos considerando aplicar a ciertas personas, unidades o procesos. Supondremos generalmente que tenemos alguna idea del beneficio de esa política cuando se aplica a distintos sectores de la población objetivo.

2. (**Focalización**) Esta intervención o política pretende focalizarse en ciertas unidades o circunstancias para mejorar resultados o uso de recursos.

3. (**Predicción o Estimación**) La focalización depende de algunas características desconocidas de las personas, unidades o procesos. Utilizamos entonces predicciones o estimaciones basadas en métodos de ML/AI para informar variables desconocidas y mejorar la focalización.

Ejemplos de esto puede ser:

- Predicción del desempeño de alumnos en un sistema educativo (unidad individuo)
- Predicción de ingresos trimestrales para hogares (unidad hogar)
- Predicción de eventos riesgosos, por ejemplo actividad criminal (unidad zona geográfica)
- Predicción de qué documentos o artículos serían útiles sugerir dada una búsqueda (unidad usuario-búsqueda)

Principio guía 2

En este documento, nos concentramos en la parte de **predicción o estimación**, y cómo deficiencias o lagunas en el proceso de validación puede producir resultados indeseables de **focalización**

2.3 Los retos de la aplicación

En estos tres casos, examinaremos los problemas más comunes, diagnósticos para detectarlos, y sugerencias para mitigarlos, divididos en cuatro secciones:

1. **Intrínsecos a los datos**: que se refieren principalmente a deficiencias, sesgos y proceso que genera los datos utilizados (**?**). Este es un punto de bloqueo para muchas aplicaciones.
2. **Relativos a la construcción y desarrollo de los modelos**: que se refiere a métodos y principios importantes para construir modelos robustos y validados correctamente.
3. **Relativos a la evaluación y monitoreo del modelo**: que se refiere a la interpretación del modelo, evaluación una vez puesta en producción y principios de monitoreo para evitar consecuencias inesperadas.
4. **Rendición de cuentas**, que se refiere al entendimiento o la capacidad de explicar cómo es que métodos automatizados toman ciertas decisiones.

Sección 3

Retos intrínsecos a los datos

El primer conjunto de retos que tratamos son los que tienen qué ver con los datos disponibles para aplicar métodos de aprendizaje automático. Estos se pueden separar en dos grandes partes:

1. Calidad y relevancia de las métricas recolectadas
2. Información incompleta acerca de la población objetivo

3.1 Métricas: introducción

En primer lugar, debemos considerar qué tipo de mediciones tenemos a nuestra disposición para aplicar métodos cuantitativos, incluyendo aprendizaje automático, y cuáles son las mediciones ideales para la decisión que se considera tomar.

Términos

- **Métrica objetivo**: medición que se considera ideal para tomar la decisión relevante. En la medición, ese ideal es tipicamente solo alcanzando en parte.

Reto: mala correspondencia de métrica y objetivos

- Las variables medidas corresponden pobremente a las métricas objetivo de interés, quizá con sesgos sistemáticos que vayan en contra de objetivos de la política.

  - En particular, las variables medidas contienen sesgos indeseables en términos de los objetivos de la política

Ejemplos

- Supongamos que una política se considera aplicar a personas de ingresos bajos. Nuestro conjunto de datos de entrenamiento/validación usamos una medición obtenida en una encuesta, según la pregunta, *¿cuánto estima usted que es el ingreso mensual de su familia?*, o algo similar. Esta medición está sujeta a sesgos desconocidos, y existen incentivos para ocultar fuentes de ingresos por parte de los participantes. Predictores de ingreso construidos con estos datos tienen el riesgo de replicar el sesgo de las mediciones, afectando negativamente los resultados de la asignación de la intervención.
- Consideramos pronosticar los niveles de demanda de medicamentos, de manera que pueda satisfacerse la demanda adecuadamente sin incurrir en inventario que caduca. Utilizamos el *número de unidades que fueron requeridas en el sistema* para el medicamento X. Esta no es una medida exacta de la demanda, porque puede ser que cuando los inventarios se agotan, los involucrados dejan de hacer requerimientos a los abastecedores. Pronosticar la demanda con estos datos puede incurrir en subeestimación, con el resultado de que reforzamos o empeoramos la escasez de medicamentos.

Medidas: mala correspondencia de métricas y objetivos

  - (Cualitativa) Las métricas objetivo deben plantearse claramente, aunque sean ideales. Las métricas recogidas deben ser analizadas para entender qué tan adecuadas son para sustituir la métrica objetivo. Se deben identificar sesgos sistemáticos o validez pobre de la métrica sustituto.
  - (Cuantitativa) Estudios adicionales diseñados para capturar métrica objetivo y métrica seleccionada permiten comparar las dos, y entender si hay sesgos que corregir y con qué variables puede lograrse esto.

Ejemplos

- En la estimación de ingreso, generalmente se utilizan fuentes de datos oficiales con metodología bien establecida para estimar el ingreso de un

hogar. Tal metodología debe ser sostenida por validaciones de distinto tipo que muestra posibles sesgos en la medición y de ser posible cuantificaciones del error de medición.
- Para los pronósticos de demanda, es posible que sea necesario identificar fuentes adicionales de datos que indiquen mejor la demanda, ya sea con estudios indirectos (fuentes de datos de los compradores) o construyendo experimentos que nos permitan observar la demanda en ciertas unidades.

3.2 Información incompleta

Cuando buscamos aplicara un método del aprendizaje automático en nuestro contexto, la información puede ser incompleta de distintas maneras:

1. Solo tenemos datos para una muestra de una población.
2. No tenemos todas las mediciones importantes para la toma de decisiones.
3. Tenemos datos del pasado y presente, pero no del futuro, que es cuando tenemos qué tomar decisiones.

3.3 Muestras naturales y diseñadas

Nuestros modelos pretenden ser aplicados a la población objetivo, pero la situación usual es que sólo tenemos disponible una **muestra** de esta población, a partir de la cual queremos desarrollar predicciones o estimaciones que ayuden en la toma de decisiones posterior.

Términos

- **Población objetivo**: son las unidades que se pretende intervenir (personas, hogares, zonas geográficas, etc). Los modelos se construyen entonces con el fin de aplicarse a la población objetivo.
- **Estructura predictiva**: se utiliza para hablar en general del tipo de modelos que se utilizan para hacer predicciones (lineales, bosques aleatorios, redes neuronales), las características que utiliza y cómo las utiliza el modelo (interacciones, transformaciones no lineales).
- **Subpoblaciones de interés** o **Subpoblaciones protegidas**: subpoblaciones de la población objetivo para las cuales queremos tener evaluaciones concretas del desempeño de estimaciones o los modelos.
- **Representatividad**: término alternativo a muestra probabilística bajo el caso más simple de muestreo autoponderado. Este es un requisito que se considera relativamente poco importante o relevante. Lo importante es que los datos provengan de un diseño muestral probabilístico.

Cuando tenemos una muestra, consideramos dos extremos posibles, siendo el primer extremo el más deseable aunque no siempre alcanzable:

1. **Muestreo probabilístico**: Los casos son seleccionados a partir de un diseño muestral probabilístico.
2. **Muestras naturales**: Los casos son seleccionados por un proceso natural mal o parcialmente conocido.

Bajo el caso 1, todos las predicciones y estimaciones que se pretenden aplicar a la población objetivo pueden ser evaluadas en cuanto a su precisión, con **garantías probabilísticas**. Por ejemplo, podemos dar rangos de error para estimaciones de cantidades asociadas a toda la población objetivo, tasas de error calibradas correctamente a la población objetivo, etc.

Bajo el caso 2, estrictamente hablando no es posible saber qué va a pasar cuando apliquemos nuestros modelos a la población general, y no es posible construir rangos de error de predicciones y estimaciones mediante métodos estadísticos que tengan garantías probabilísticas.

El primer reto es entonces:

Reto: muestras naturales

Las muestras naturales de datos pueden resultar en:

- Errores de estimación y de predicción incorrectamente estimados
- Estructuras predictivas distintas a las que observaríamos en la población objetivo (modelos no válidos).
- Extrapolaciones que no son soportadas por los datos.
- Subreprensentación de subgrupos protegidos

Es decir: las cantidades y predicciones estimadas tienen error desconocido, los modelos y características útiles en la muestra pueden no aplicar en la población objetivo, y la situación puede agravarse para grupos protegidos subrepresentados.

Ejemplo

Supongamos que nos interesa predecir la prevalencia de anemia en niños de una población objetivo dada. Decidimos tomar una muestra acudiendo a hospitales que "corresponden" a la población objetivo para aplicar las pruebas correspondientes. La prevalencia de anemia que encontremos en esta muestra tendrá error desconocido, posiblemente grande, como estimación de la prevalencia en la población objetivo. Más aún, modelos de predicción construidos para esta muestra natural no necesariamente generalizarán correctamente a la población

total, y no es posible tener una estimación confiable del error. Puede ser que la estructura predictiva para la población objetivo sea muy distinto de la que obtenemos con nuestra muestra natural.

Medidas: muestras naturales

- (Cualitativa) Entender y describir las dimensiones importantes en las cuales nuestra muestra puede ser diferente a la población, en particular sesgos de selección no medidos. Utilizar literatura relacionada con el tema, información de expertos.
- (Cuantitativa) Aunque los modelos pueden construirse con varias fuentes de datos, diseñadas o naturales,
  la validación debe llevarse a cabo idealmente con una muestra diseñada que permita inferencia estadística a la población objetivo. La muestra de validación debe cubrir apropiadamente la población objetivo y subpoblaciones protegidas.

- La construcción de la muestra de validación debe ser producida bajo un diseño muestral que permita inferencia a la población objetivo (**?**).
- La muestra de validación debe cubrir a subgrupos de interés y protegidos, de manera que sea posible hacer inferencia a sus subpoblaciones. Eso incluye tamaños de muestras adecuados según metodología de muestreo (ver **?**).
- Si no está disponible tal muestra, es **indispensable** un análisis de riesgos y limitaciones de la muestra naturales, conducida por expertos y personas que conozcan el proceso que generó esos datos muestrales.
- Muestras *representativas* no son necesarias ni en construcción de modelos ni en validación. Por ejemplo, en muestreo de personas muchas veces se usan esquemas de selección desconocidos, pero acotado por cuotas. Esto produce balance en las tasas a la que aparecen ciertos grupos, sin embargo, sigue siendo una muestra natural (en este caso a la muestra natural se le llama también *de conveniencia*).

La situación ideal es la de muestreo probabilístico. En este caso, podemos entender exactamente qué subpoblaciones se muestrearon, a qué tasas, y cómo se relacionan estas tasas con las tasas poblacionales. El diseño de la muestra determina nuestro alcance inferencial.

3.4 Muestras y comparaciones predictivas

En muchos casos, parte de la aplicación de un modelo de predicción se concentra en comparaciones predictivas muy particulares. La política se aplica en función de hallazgos del modelo en términos de las variables incluidas en el modelo.

Es necesario ser muy cuidadosos con este tipo de uso (**?**). Si las variables a considerar están asignadas naturalmente por un proceso desconocido, la derivación de políticas a partir de esos modelos puede llevar a decisiones erróneas.

En este caso, aún cuando la muestra puede permitir inferencia a la población, estamos a fin de cuentas haciendo inferencia causal.

Reto: comparación casual

Comparaciones predictivas o contrafactuales de algun tratamiento o variable con datos de muestras que no tienen algún tipo de asignación aleatoria de tal tratamiento o variable pueden ser muy lejanas de comparaciones causales al aplicar el tratamiento en la realidad.

Medidas: comparación causal

- (Cualitativa) Entender y describir las razones por las que la variable tratamiento está correlacionada con variables conocidas y no conocidas. Describir sesgos posibles basados en análisis y conocimiento experto. Considerar qué variables control serían importantes para que la comparación predictiva tenga interpretación causal
- (Cuantitativa) Producir datos diseñados que incluyan la consideración causal. Esto incluye experimentos aleatorios y otras técnicas. Esto puede hacerse antes de construir los modelos, o incluir gradualmente estos datos en el proceso de monitoreo y reajuste.

En muchos casos, suplir con modelación datos creados experimentalmente puede ser muy difícil, y los resultados pueden depender fuertemente del decisiones del modelador, con consecuente incertidumbre rara vez medida.

Observación: este problema es ortogonal al de la representatividad o diseño muestral. Muestras bien diseñadas desde el punto de vista de la población objetivo pueden ser poco apropiadas para hacer inferencia causal, y a la inversa, datos experimentales pueden proveer indicaciones causales correctas en la muestra seleccionada, pero tener dificultades para generalizar a una población objetivo.

3.5 Atributos incompletos

Reto: atributos faltantes o incompletos

- Información crucial acerca de las unidades es totalmente desconocida puede resultar en modelos de desempeño pobre, con poca utilidad para la toma de decisiones. Comparaciones predictivas pueden ser poco útiles y a veces engañosas cuando existen variables omitidas importantes.
- Información parcial completada con procesos de imputación puede producir sesgos, dependiendo de la razón por la que las observaciones son incompletas.

Muchos proyectos de aprendizaje automático están destinados a fallar por ignorar variables o atributos que son importantes para predecir la variable objetivo. Adicionalmente, cuando existe información parcial acerca de los atributos, generalmente la ausencia de esa información muchas veces está asociado a características relevantes de la unidades para las que se quiere predecir.

Medidas: atributos incompletos

- (Cualitativa) Identificar si existen variables omitidas importantes para las cuáles no se tiene mediciones asociadas. Indentificar razones por las que existen datos faltantes: si la falta de datos está fuertemente asociada con la variable a predecir será difícil obtener buenos resultados.
- (Cuantitativa) Los procesos de imputación tienen que ser evaluados en cuanto a su sensibilidad. De preferencia, se deben utilizar métodos de imputación múltiple que permitan evaluar incertidumbre en la imputación (**?**), (**?**).

Sección 4

Retos en la construcción y desarrollo de los modelos

En primer lugar, consideramos el proceso usual de construcción de métodos predictivos con aprendizaje automático (**?**), (**?**):

1. Preprocesamiento y limpieza de datos
2. Entrenamiento o ajuste de métodos de predicción
3. Estimación de métricas de error y selección de predictores
4. Validación final de predictor seleccionado

Esto generalmente involucra al menos dos muestras (1 y 2), y de preferencia tres:

1. Datos de entrenamiento
2. Datos de validación
3. Datos de prueba

4.1 Ausencia de validación

Uno de los primeros errores graves en este proceso es la no consideración de etapas robustas de validación y prueba de los modelos

Reto: ausencia de muestras de validación

Los resultados de la construcción de modelos se presentan según su desempeño con el conjunto de datos que se usó para entrenarlos. Las métricas

de desempeño, en este caso, en general no pueden utilizarse para evaluar el verdadero comportamiento del modelo para las nuevas muestras con las que se pretende usar.

Este problema de validación inexistente o pobre ocurre muchas veces con **pronósticos de series de tiempo**, donde generalmente tenemos poca información a futuro para garantizar buen desempeño, o se trata de procesos altamente dinámicos que son en cualquier escenario difíciles de predecir.

Medida: ausencia de muestras de validación

- (Cuantitativa 1) Construir muestras de validación y prueba preparadas adecuadamente, como discutiremos más adelante. Esto incluye tamaño apropiado para estimar el error con precisión razonable.
- (Cuantitativa 2) Existen estrategias de remuestreo o consideraciones estadísticas teóricas fundamentadas para justificar la generalización del desempeño en entrenamiento.

Argumentos teóricos requieren cuidado adicional, y sus supuestos deben ser evaluados.

4.2 Fugas de información

Las fugas de información (**?**) ponen en duda la validación de modelos como manera de estimar el desempeño en producción de los métodos de aprendizaje automático. Esto ocurre de dos maneras:

- La muestra de entrenamiento recibe *fugas* de los datos de validación, lo que implica el uso de datos de validación en entrenamiento e invalida la estimación del error de predicción.
- Muestras de validación y entrenamiento tienen agrupaciones temporales o de otro tipo que no se conservan en el proceso de entrenamiento y validación. En este caso, entrenamiento y validación recibe *fugas* de información que no estará disponible el momento de hacer predicciones.

Reto: fugas entrenamiento validación

Si alguna parte de los datos de validación/prueba se utiliza en la construcción de los modelos durante entrenamiento, la muestra de validación prueba no cumple su función de dar una estimación realista del error en producción.

Ejemplo

Validación cruzada con selección de variables usando todos los datos (**?**)

Medidas: fugas entrenamiento validación

Cualquier procesamiento y perparación de datos de entrenamiento debe evitar usar los datos de validación o prueba de ninguna manera. Se debe mantener una barrera sólida entre entrenamiento vs validación y prueba.

Esto incluye recodificación de datos, normalizaciones, selección de variables, identificación de datos atípicos y cualquier otro tipo de preparación de cualquier variable a ser incluida en los modelos.

El segundo tipo de filtración

Reto: fugas de datos no disponbiles en la predicción

Algunos modelos son riesgosos de poner en producción pues utilizan variables en entrenamiento y validación que no estarán disponibles en la misma forma al momento de poner en producción. Esto generalmente tiene ver con temporalidad de los datos o agrupaciones particulares.

Ejemplo

Un modelo hace predicción de actividad criminal en distintas zonas geográficas para el tiempo t. En la extracción de datos se usa como variable de entrada el número de unidades de policía que antendieron la zona de interés al tiempo t. Esto representa una fuga en la predicción, pues al momento de predecir actividad criminal al tiempo t no estará disponible las unidades de policía al tiempo t. El modelo puede parecer preciso, pero en producción su exactitud se verá considerablemente degradada.

En el caso más extremo, aunque quizá más fácil de detectar, existen variables presentes en datos de entrenamiento que no estarán disponibles en producción

(por ejemplo, cantidad impagada si estamos haciendo predicción de impago). **En casos más sutiles este error puede ser difícil de detectar**.

- En entrenamiento: pueden existir variables acumuladas hasta el momento donde se registra la variable a predecir.
- En producción: las variables están acumuladas hasta el momento donde se hacen las predicciones. La variable a predecir ocurre en el futuro.

Este tipo de error generalmente produce modelos que parecen muy optimistas, y ocurre de muchas maneras.

Reto: fugas de datos no disponbiles en la predicción

El esquema de validación debe **replicar tan cerca como sea posible** el esquema bajo el cual se aplicarán las predicciones. Esto incluye que hay que replicar

- Ventanas temporales de observación y registro de variables y ventanas de predicción
- Si existen grupos en los datos, considerar si tendremos información disponible de cada grupo cuando hacemos la predicción, o es necesario predecir para nuevos grupos.

Ejemplo

Supongamos que queremos predecir, en varias regiones o ciudades, el daño de edificios a partir de fotos aéreas después de un temblor, usando como métrica objetivo peritajes de los edificios seleccionados. En la validación podríamos cometer el error de no respetar la agrupación regional, y el modelo podría parecer dar buenas predicciones. En la realidad, aplicaríamos para una región sobre la cual no tenemos información. La validación debe considerar la necesidad de predecir para puntos en regiones enteras sin tener información adicional de tal región (es decir, la validación debe estratificar por región).

4.3 Clasificación: probabilidades y clases

En muchos problemas de clasificación, por su naturaleza, es difícil acercarse a tener certidumbre acerca de la clase de una observación según sus covariables. En estos casos, es más útil usar probabilidades

Reto: puntos de corte arbitrario

En problemas de clasificación, los puntos de corte o decisiones de clasificación se toman con criterios vagamente relacionados con el contexto de la decisión (por ejemplo, escogiendo una sensibilidad o especificidad dadas).

Muchas veces se toma erróneamente un punto de corte de 1/2 para clasificación binaria, por ejemplo. Esta decisión se toma fuera del contexto de la decisión que se quiere tomar.

Medida: puntos de corte arbitrario

- En problemas de clasificación ruidosos (no es posible acercarse a tener certidumbre para muchos casos), las **probabilidades de clasificación** en cada clase son instrumentos más apropiados para la toma de decisiones.
- Costos y utilidades pueden utilizarse, en combinación con las probabilidades, para tomar mejores decisiones caso por caso.

Ejemplo: churn de clientes, evaluación de alumnos

4.4 Clasificación: Datos desbalanceados

En problemas de clasificación muchas veces se presenta el problema de que algunas clases tienen representación relativamente baja (por ejemplo, clases con menos de 1% de los casos totales). Estas clases presentan dificultades considerables en los modelos predictivos, pues puede ser que tengamos poca información acerca de esas clases y sea difícil discriminarlas existosamente de otras clases, aún cuando contemos con la información correcta.

Reto: desbalance de clases

En datos con desbalance grande, **predictores de clase** pueden tener desempeño malo (por ejemplo, nunca hacen predicciones de la clase minoritaria).

La solución es considerar las probabilidades de clase como salida principal:

Medidas: desbalance de clases

- Hacer **predicciones de probabilidad** en lugar de clase. Estas probabilidades pueden ser incorporadas al proceso de decisión posterior como tales. Evitar puntos de corte estándar de probabilidad como 0.5, o predecir según máxima probabilidad.
- Cuando el número absoluto de casos minoritarios es muy chico, puede ser muy dificil encontrar información apropiada para discriminar esa clase. Se requiere **recolectar** más datos de la clase minoritaria.
- Submuestrar la clase dominante (ponderando hacia arriba los casos para no perder calibración) puede ser una estrategia exitosa para reducir el tamaño de los datos y tiempo de entrenamiento.

Ejemplos

- Consideremos que tenemos 1 millón de datos, 999 mil negativos y mil positivos. Puede ser buena idea submuestrar los negativos por una fracción dada (por ejemplo 10%) ponderando cada caso muestreado por 10 en el ajuste y el postproceso.
- Consideremos que tenemos 1 millón de datos, 999,950 mil negativos y 50 positivos. Puede ser imposible discriminar apropiadamente los 50 datos positivos. Construir conjuntos de validación empeora la situación: no es posible validar el desempeño predictivo ni construir un modelo con buen desempeño.

Ejemplo: CTR en ligas (Google)

Observaciones:

- Sub y sobremuestreo alteran la proporción natural de positivos y negativos en los datos. Esto quiere decir que las probabilidades obtenidas están mal calibradas y tienen menos utilidad para la toma de decisiones.

4.5 Subajuste y sobreajuste

Subajuste y sobreajuste ocurren cuando la información predictiva en los datos es usada de manera poco apropiada: en subajuste agrupamos de más y damos demasiado poco peso a características individuales de los casos, y en subajuste les damos demasiado peso.

Términos

- **Sobreajuste**: un modelo demasiado complejo para los datos disponibles tiende a capturar características no informativas como parte de la estructura predictiva. Esto se refleja muchas veces en una brecha de error grande entre entrenamiento y validación. Estos pueden ser modelos ruidosos difíciles de interpretar, y las predicciones pueden ser inestables dependiendo del conjunto de datos particular que se utilice.

- **Subajuste** uno modelo demasiado simple para los datos disponibles tiende a ignorar patrones sólidos en la estructura predictiva. Esto se refleja en errores sistemáticos e identificables, por ejemplo, sub/sobre predicción sistemática para ciertos grupos o valores de las variables de entrada.

Reto: sub y sobreajuste

- Modelos que presentan subajuste o sobreajuste son particularmente difíciles de interpretar, y comparaciones predictivas pueden ser malas.
- Modelos subajustados pueden cometer errores sistemáticos que pueden afectar negativamente, por ejemplo, al tratamiento de grupos protegidos.
- Modelos sobreajustados pueden tener predicciones inestables que cambian mucho dependiendo de los datos, por ejemplo, con cada actualización.

Aunque sub y sobre ajuste puede producir resultados predictivos subóptimos, pueden producir rangos de error aceptables. Sin embargo, están expuestos a los problemas señalados arriba.

Medidas: sub y sobreajuste

- Sobreajuste: debe evitarse modelos cuya brecha validación - entrenamiento sea grande (indicios de sobreajuste).
- Subajuste: deben checarse subconjuntos importantes de casos (por ejemplo grupos protegidos) para verificar que no existen errores sistemáticos indeseables.

Ejemplo: reconocimiento de imágenes

4.6 Equidad y desempeño diferencial de predictores

Métodos basados en aprendizaje automático pueden producir predicciones, que cuando no son usadas apropiadamente en la toma de decisiones, pueden producir resultados injustos o discriminatorios ((**?**), (**?**), (**?**)).

En primer lugar establecemos que no siempre el contexto completo del problema de decisión puede enmarcarse dentro de la parte correspondiente al *aprendizaje automático*. Generalmente habrá varios objetivos de los tomadores de decisiones que van más allá de un pronóstico dado, un sistema de clasificación, etc.

Un ejemplo es el de paridad demográfica (por ejemplo, que dos grupos de interés obtengan tasas de clasificación positiva muy similar), que es un tipo de paridad puede ser deseable para los tomadores de decisiones, pero no necesariamente uno que aparece naturalmente en lso métodos predictivos. De esta forma separamos dos preocupaciones importantes:

- Objetivos de política pública o estrategia, orientados a la justica algorítmica, que tienen que incorporarse en el proceso de toma de decisiones.
- Fallas técnicas en los modelos que producen disparidad de resultados para grupos protegidos.

En esta parte discutiremos principalmente el segundo punto.

4.6.1 Términos

- **Atributo protegido**: una característica o variable **protegida** A es una para la que queremos que se cumplan cierto criterio de equidad en las predicciones.

Nuestro objetivo es establecer lineamientos para evitar que deficiencias en los modelos produzcan disparidades indeseables según los distintos subgrupos asociados a una variable protegida A (por ejemplo, A puede ser género, raza, nivel de marginación).

Dos estrategias no muy útiles para prevenir disparidades entre los grupos de A son: *ignorar* la variable A y buscar *paridad demográfica* de predicciones. En el primer caso, se pretende eliminar la posibilidad de disparidad **no** incluyendo la variable G en el proceso de construcción de predictores. Este enfoque en

pocos casos produce resultados deseables, pues típicamente existen otros atributos asociados a A que pueden producir resultados similares aunque A no se considere. El segundo caso, paridad demográfica de predicciones buscamos que las predicciones de los distintos grupos de A sean similares: en el caso de clasificación, por ejemplo, que la tasa de positivos sea similar. Esto poco deseable por sí solo: por ejemplo, si quisiéramos construir un clasificador para cierta enfermedad, consideramos que es posible que mujeres y hombres sean afectados de manera distinta. Sin embargo, *paridad demográfica* puede ser un objetivo de los tomadores de decisiones, y eso debe tomarse en cuenta al momento de tomar la decisión asociada a la predicción.

El concepto de **equidad de posibilidades** ((**?**)) es uno menos dependiente de los objetivos de los tomadores de decisiones, y se refiere al desempeño predictivo a lo largo de distintos grupos definidos de A. Si Y es la variable que queremos predecir, y $\hat{Y}$ es nuestra predicción, decimos que nuestra predicción satisface **equidad de posibilidades** cuando

- $\hat{Y}$ y A son independientes dado el valor verdadero Y

Esto quiere decir que A no debe influir en la predicción cuando conocemos el valor verdadero Y, o dicho de otra manera: A sólo puede influir en la predicción a través de su efecto sobre Y. Consideramos predictores que se alejan mucho de este criterio son suceptibles de incluir disparidades asociadas a la variable protegida A

Una primera implicación de este criterio es:

- Bajo el suspuesto de equidad de posibilidades, las tasas de error predictivo sobre cada subgrupo de A son similares

En problemas de clasificación binaria, el criterio es_

- En cada subgrupo, las tasas de falsos positivos y de falsos negativos son similares

Reto: inequidad algorítmica

Aún conociendo el verdadero valor de la variable que queremos predecir, las predicciones de un método dado dependen fuertemente de una variable protegida. En particular, las tasas de error de distintos grupos de la variable protegida pueden ser muy distintos.

Esta situación de inequidad muchas veces implica que la estructura predictiva depende fuertemente o *abuse* (**?**) de la información que contiene la variable protegida A acerca de la variable respuesta, con el riesgo de producir sesgos injustos a lo largo de distintos valores de la variable protegida.

En el caso de clasificacón binaria, cuando una de las alternativas es *deseable* para los individuos (por ejemplo, calificar para un beneficio, crédito, candidatura de un trabajo, etc), Un criterio menos exigente de equidad puede ser la **equidad de oportunidad**:

- Bajo el suspuesto de equidad de oportunidad, las tasas de falsos negativos de cada subgrupo de A son similares.

En la práctica puede considerarse cuál es más apropiado. Equidad de oportunidad muchas veces es un criterio aceptable, que introduce criterios de justicia algorítmica permitiendo también optimizar otros resultados deseables.

Ejemplo

Supongamos que queremos predecir si un potencial empleado va a durar más de 2 años en cierta compañía, y que la variable protegida A toma dos valores (que supondremos en este caso toma dos valores: se autodenomina con religión o sin religión). El predictor satisface **equidad de posibilidades** cuando tanto la tasa de falsos positivos como la de falsos negativos son iguales para personas con religión y sin religión.

Medidas: inequidad algorítmica

- Cuando existen atributos protegidos, debe evaluarse qué tanto se alejan las predicciones de la **equidad de posibilidades o de oportunidad**
- Posprocesar adecuadamente las predicciones, si es necesario, para lograr equidad de posibilidades o oportunidad (**?**).
- En el caso de clasificación, puntos de corte para distintos subgrupos pueden ajustarse para lograr equidad de oportunidad.

Esto en general implica que además de la decisión tomada en función de las predicciones depende de esta métrica adicional de equidad, y no solo del análisis costo-beneficio.

Sección 5

Retos de uso durante ejecución

Una vez que los métodos de aprendizaje automático se comienzan utilizar para tomar decisiones, es necesario:

- Monitorear, en general, desempeño y características usadas en el tiempo.
- Monitorear en particular resultados indeseables que pueden resultar de la interacción de usuarios con algoritmos.
- Tomar decisiones acerca del proceso generador de datos para mejorar desempeño.

5.1 Monitoreo de desempeño

** Reto ** El desempeño de un modelo puede degradarse en el tiempo, debido a cambios en la población sobre la que se hacen predicciones

** Medidas ** - (Cuantitativa) Monitorear varias métricas asociadas a las predicciones, en subgrupos definidos con antelación (incluyendo variables protegidas). - (Cuantitativa) Monitorear deriva en distribuciones de características con respecto al conjunto de entrenamiento. - (Cualitativa) Cuando sea aplicable y factible, una fracción de las predicciones deberán ser examinadas por humanos y calificadas según alguna rúbrica o mediciones de las variable que se busca predecir.

Sección 6

Retos de rendición de cuentas

6.1 Interpretabilidad y explicación de predicciones.

Es difícil dar una definición técnica de **interpretabilidad** o **explicabilidad**, que en general se refieren a hacer inteligible para humanos el funcionamiento de un algoritmo ((**?**), (**?**)). Hay varias razones por las que tener cierto grado de interpretabilidad en los algoritmos que se usan para tomar decisiones es importantes ((**?**)):

1. Aprendizaje acerca del dominio del problema.
2. Aceptación social.
3. Detección de sesgos potenciales de los algoritmos.
4. Depuración y mejora de modelos.

Los puntos 1 y 2 son más difíciles de definir, y advertimos que cualquier técnica aplicada a interpretar los modelos con estos dos fines tienen varias dificultades que superar. El *aprendizaje automático*, como se usa tipicamente hoy en día, dificilmente se acerca a explicaciones causales o mecanísticas.

El punto 3 y 4, que veremos más abajo, son más suceptibles de análisis técnico. Los sesgos potenciales pueden ocurrir cuando en el proceso de aprendizaje se aprenden de características que son irrelevantes, pero caracterizan a los conjuntos de entrenamiento y validación/prueba que fueron utilizados.

Por otra parte, existe en muchos casos la necesidad de dar **explicaciones individuales** de cómo fueron tomadas ciertas decisiones (por ejemplo, por qué a