Ética

Definición: tiene como objetivo definir y razonar de manera sistemática sobre los conceptos del bien y el mal y sobre cómo aplicarlos a la conducta humana (o de cualquier agente moral).

El aprendizaje automático se aplica cada vez más a actividades de nuestra vida diaria. Los modelos pueden tener sesgos por género por ejemplo, porque los datos con los que entrenamos lo tenían.

Problemáticas:
- Toma de decisiones de forma automática y masiva basadas en prejuicios o sesgos
- Reproducir, perpetuar e incluso agudizar sesgos o injusticias. Se agudizan los problemas que ya tenemos.

Soluciones:
- sociales
    - ser conscientes del impacto que nuestro trabajo tiene en la realidad.
    - regular la producción de los modelos de IA.
- técnicas
    - interpretabilidad
    - fairness: que nuestros modelos sean justos
    - Ética por diseño

Ética por diseño
Idea: **incluir los límites/restricciones morales** dentro del proceso de entrenamiento del modelo (en los datos o en la función de costo). Ej si el modelo se vuelve racista, que empeore la función de costo.

idea en desarrollo ya que no es fácil determinar qué moral “usar” y que requiere una complejidad computacional muchas veces imposible en la práctica.

Interpretabilidad

Contexto
Muchas veces, los modelos que mejor funcionan y terminan saliendo a producción son sumamente complejos y ni los entendemos ni sabemos qué hacen adentro

Si no sabemos porqué un modelo predice como predice, no deberíamos sacarlo a producción

La interpretabilidad viene para ayudarnos a entender que hay dentro de la caja negra (modelo).

Problema:
- si no confiamos/entendemos un modelo o sus predicciones no podemos usarlo
- el desempeño en una base de validación puede ser bueno, pero su comportamiento en el mundo real puede ser impredecible
- no se puede confiar ciegamente en un sistema para la toma de decisiones.

Motivación de la interpretabilidad
- Fairness: No queremos que el algoritmo tenga comportamiento discriminatorio.
- Robustez: confiamos en que funcionará en la vida real?
- Confiabilidad: se comporta correctamente frente a las razones correctas.
- Privacidad

Técnica: LIME
LIME: local interpretable model- agnostic explanations

El **objetivo** es **identificar un modelo interpretable** a partir de una **representación interpretable** que es localmente **fiel al clasificador.**
Una explicación interpretable tiene que ser entendida por humanos, al margen de la complejidad de los descriptores utilizados.

La idea es que para una instancia (que es una perturbación de mi instancia interpretable) lo que voy a hacer es usar mi modelo que se que funciona y para esa instancia voy a crear un modelo interpretable g (árbol, etc) que se parezca a como razona f localmente.

Fairness: Nos da las **herramientas matematicas** para que nuestros sistemas **reflejen nuestros valores**. Fairness no nos dice cuáles son esos valores.

Atributo protegido: Son atributos, datos que no deben ser almacenados/considerados/utilizados.

Proxy: variable que no contiene al atributo protegido pero hace que se pueda inferirlo.
problema: generalmente son variables que son inherentemente útiles para predecir lo que queremos predecir. Y no sabemos cuánta de la atención que le está prestando el modelo es por el feature en sí mismo y cuanto porque sirve para predecir el atributo protegido. No podemos hacer al modelo ciego de los atributos protegidos porque los va a poder inferir por otros lados.

El modelo es insesgado si el bias = 0. Si no es sesgado.

Calibración
Los modelos tienden a exagerar las probabilidades cuando uno se aleja del 0.5. Lo podemos usar para ver que tan sesgado o no sesgado es el modelo.

“La calibración de un modelo de clasificación consiste en **reajustar las probabilidades predichas para que correspondan con la proporción de casos reales observados**. En otras palabras, corregir las probabilidades predichas por un modelo cuando este las subestima o sobreestima. “ (internet)

Group fairness
Para dos grupos comparamos las métricas. Queremos que las métricas nos den igual para todos los grupos.

Teoremas de imposibilidad:
Si la data presenta una distribución distinta para cada grupo:

- **Si tomamos dos métricas cualesquiera es imposible hacer un sistema los iguale para ambos grupos.**

En un mundo injusto hacer un sistema justo es imposible. Esto aplica para cualquier sistema, informático o humano.

Para aplicar soluciones de fairness se necesita conocer los grupos para los que se está aplicando. El objetivo de un modelo es minimizar un error.

Costo de fairness

El objetivo de un modelo es minimizar el error. **Si le agregamos a ese modelo una restricción lógicamente el error aumentará**, aplicamos restricciones para que tales grupos sean iguales en cierta cosa. El modelo ya no hace lo mejor posible respecto de los datos.

El problema de la automatización

Los sistemas humanos nos permiten cambiar cosas sobre la marcha. Casos outliers no contemplados. Todos nosotros estamos expuestos a ser outliers. Con los sistemas automatizados tenemos el problema de que no son capaces de reconocer la intención del sistema, entonces no pueden tomar una determinación propia.

La automatización nos da la oportunidad de medir fairness más que cualquier sistema humano ya que al tener un sistema determinista podemos calcular las métricas que queramos.

Un **Stakeholder es un individuo**, grupo u organización **podría afectar el, ser afectada o percibirse como afectada por cualquier actividad, decisión o resultado de un proyecto.**

Fairness tree: dado un stakeholder, respondo una serie de preguntas para ver cual es la métrica que me conviene.

Adversarial learning: es entrenar por un lado descenso por el gradiente para algo y ascender en el gradiente en otra cosa para evitar que la red neuronal aprenda algo que no queremos.

Individual fairness: dos personas son muy parecidas pero reciben un trato distinto.

ENTONCES LO QUE DEBERIAMOS HACER ES

IDENTIFICAR LOS STAKEHOLDERS, CONOCERLOS Y LEUGO HACER FAIRNESS DE GRUPOS Y VER QUE ONDA.