Machine Learning

Tipos de problema

1. **Clasificación**
   - Predice una clase
   - Las métricas: que hacen sentido para este tipo de problemas, accuracy.
   - Accuracy: Los que fueron buenos / todos
   - Precision: Indica falsos positivos (f/(t+f))
   - Recall: Indica falsos negativos (t/(t+f))
   - F1-Score: Cuando Precision y Recall tienen la misma importancia
2. **Regresión**
   - Predice un valor.
   - Las métricas: Error MSR2, etc.

Tipos de datos

- Estructurado
- Semiestructurado
- No estructurado: Texto cae acá. Tenemos que pasarlo de un esquema no estructurado a un estructurado, en numeritos.

Aprendizaje

1. Supervisado: Tiene etiquetas
2. No supervisado: No tiene etiquetas, no tiene métricas de desempeño. No sabes qué es la verdad, asi que no podemos decir qué está bien y mal
   - TSNE
   - Análisis de componentes principales
   - Clusterización
     - K-Means
     - OBSCAN
     - Affinity Propagation
3. Por refuerzo

Analisis de los datos

Tipos de datos

- Estructurado
- Semiestructurado
- No estructurado: Texto cae acá. Tenemos que pasarlo de un esquema no estructurado a un estructurado, en numeritos.

Variables

Numeros, cantidades o medidas con los que describimos.

- Discreta: Numeros enteros. (ej. cantidades)
- Continuas: Rango (ej. precio)
- Categóricas: Etiquetas o categorías. El tipo de variable categórica implica distintas forma de codificarlos para el modelo.

- Nominales: El orden no importa
- Ordinales: El orden importa (ej. nivel académico)

Características
- Datos faltantes (ej. null): Los modelos de machine learning no pueden trabajar si faltan datos
- Cardinalidad: Representa la cantidad de vakires distintos por categoría (ej. Genero puede ser Hombre y Mujer, o sea que tiene una casdinalidad 2).
  - Recordemos que las muestras en los sets de prueba y entrenamiento deben tener la misma carnalidad y en un *ratio* equivalente (?.
  - Si la cardanilidad es baja, casi constante, no vale la pena para entrenamiento?
- Distribución

Distribución
Describe la probabilidad de tener uno de los posibles valores que una variable puede tener (ej. prob de que sea hombre y mujer donde hay 4 hombres y 3 mujeres)

El tipo de distribución nos permite adaptar un método de imputación para rellenar los variables faltantes.

Discreta:
- Binomial
- Poisson

Continua:
- Normal: Media, moda y mediana son iguales (simétrica). La inferencia usando una distribución normal sale bien
- Uniforma
- Exponencial
- Entre otras

Medidas de tendencia central

Media: Promedio aritmético de un conjunto de datos Mediana: Valor central de un conjunto de datos ordenados.
- Para los valores atípicos, extremos, que afectan la media. (ej. casi todos ganan $5 y una persona $1000)

Moda: Valor que aparece con mayor frecuencia en un conjunto de datos

Medidas de dispersión

Rango: val_max — val_min Varianza: Dispersión de los valores respecto a la media.
- Poblacional (+30 datos)
- Muestral

Desviación estándar: Raíz cuadrada de la varianza, para tener el valor sobre la unidad original.
Rango inter cuartil: Para obtener los valores atípicos, siendo los que no están entre 1.5 de Q1 y Q3.

Regresión Lineal simple
Un modelo que intente pasar lo más cerca de todos los puntos como sea posible.

> Origen: Un estudio de estaturas de padre-hijo, la descendencia tiende a ser lo que la mayoría de la población. Los descendientes “regresan” a un parámetro que representa toda la población

$y = mx + b$ nos dice que si queremos predecir la variable independiente necesitamos una pendiente y un *interceptor* (u ordenada a el origen).

Es decir que $\hat{y} = w_1 x + w_0$ quiere decir que el valor de la variable dependiente predicho es igual a un primer peso w_1 multiplicado por el valor independiente y sumado con otro peso w_0

Evaluando la regresión lineal

A la distancia entre los puntos y la línea tenemos un **residuo** u **error**.

Error cuadrático medio (MSE)

1. Podemos obtener la distancia de todos los puntos y el valor predicho.
2. Tomar su valor absoluto (elevarlo al cuadrado).
3. Obtener el promedio

$\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i$

Root mean square error

Si al MSE le sacamos la raíz cuadrada obtenemos

$\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i}$

Mean absolute error

Usa el valor absoluto en lugar de MSE/RMSD

Ordinary Least Squares @ OLS

Es una forma de evaluar el modelo, queremos encontrar una línea que pase por todos los puntos, que la distancia entre todos los puntos sea la más mínima.

En esta forma de evaluar tenemos:

- 1 función original que suma todas las diferencias (no obtiene el promedio)
- Derivamos con respecto a β, ahora tenemos una función que podemos evaluar en 0 para obtener los mínimos y máximos.
- Podemos resolverlo a mano con álgebra lineal

Validar que una regresión lineal es válida

Linearidad: Entre la variable a predecir y las variables que lo predicen tiene que existir una relación lineal.

> Lo que hicimos sobre el residuo, y observar si hay patrones claros, si se presentan quiere decir que no se cumple

Homoscedasticidad: Cuando los datos *tienden*? a la misma dirección. Misma varianza. (la contra e hetero- distinta varianza)

- Ejemplo: Si hay una distribución de datos que tienden a 0 y a infinito (como un cono/embudo), entonces una regresión lineal es una mala idea. Esto porque el error crece en ambas (?, direcciones

Normalidad de los errores: La distribución de los errores es normal. “Q-Q Plot”

> Normalmente hacerte pasar por la validación del uso de una distribución normal es raro, suele estar automatizado.

Prueba de hipótesis

- H_0 hipótesis nula, es el status quo, no tuvo efecto
- H_1 hipótesis alternativa, ej. de hecho si tuvo un efecto

- P qué tan probable es que los resultados obtenidos sean verdaderos, dependiendo del valor de P sabemos s
- Nivel de significancia, el umbral de riesgo que queremos asumir de error. 0.05 es 5%. No nos sirve de nada

Hay pruebas de hipótesis que nos permiten medir si los datos provienen de una distribución normal:

- Kolmogorov-Smirnof: Compara la distribución de los residuos con una distribución normal.
- Shapiro : Similar a la de Smirnof pero limitada solo a distribución normal, mientras que la de Kolmogorov-Smirnof permite hacer con cualquier tipo de distribución.

Correlación

- Cuando el movimiento de una variable es proporcional otra, se dice que se correlacionan.
- Ejemplos: https://www.tylervigen.com/spurious-correlations
- Recordar que correlación no implica causa.
- Para datos lineales podemos saber si se correlacionan con el índice de correlación de Pearson.
- Ver: https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

Primer parcial solo regresión. En el examen habrá dos modelos:
- Regresión lineal donde tenemos que evaluar la validez de usar un modelo lineal con las pruebas que vimos.
- Lo que queramos.

Regresión Lineal Múltiple

Correlación

Cuanto influye el cambio de una variable, respecto a otra

Correlación negativa −1 -> sube una, baja la otra
Correlación positiva 0 -> no hay relación
Correlación Positiva 1 -> Sube una, sube la otra

Coeficiente de correlación de Pearson

Permite cuantificar la correlación entre variables.

https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson

Feature Scaling

Por la magnitud y/o unidades de las cosas podría ser que la diferencia de rangos por magnitud (ej. edad y salario) puede que haga que el efecto de la variable de menor rango de valores tenga un efecto nulo.

Ej 1000000000 (x) + 17 (edad), el 17 no hace nada.

La distribución normal estándar tiene promedio en 0 y desviación estandar de 1. Es decir centrado en 0

Hay distintos tipos de escalamiento:
- Standard Scaling aka Estandarización: La media de todo queda en 0, todo queda expresado en términos de desviación estándar, lo que hace que sea más sencillo modelarlo sin que las magnitudes afecten.

Feature Leak en Scaling

Dividimos el modelo en set de prueba y de entrenamiento. Una manera de filtrar información es el restar la media a todo; sin embargo, la media, si toma en cuenta la información del set de prueba, va a afectar las métricas del modelo porque ya habrá considerado datos que se supone no conoce.

Se usa solo los datos del set de entrenamiento para evitar este *leak*, los datos de prueba no recalculan la media ni nada, se mantienen usando la del set de entrenamiento.

Se estandarizan ambos sets por separado:

- Los de entrenamiento se escalan con media de esos datos
- Los de prueba se escalan usando la media de los de entrenamiento.