38. ¿En qué consiste la ingeniería de características?

Es el proceso de determinar que variables productivas contribuyen mejor al poder predictivo del algoritmo. FE es, quizá, la parte más importante del proceso de minería de datos. Con buenas variables, un modelo simple puede ser mejor que un modelo complicado con malas variables. Es el elemento humano en el modelado: El entendimiento de los datos, más la intuición y la creatividad, hacen toda la diferencia. Es más un arte que una ciencia. Regularmente es un proceso iterativo con el EDA. Un domain expert puede ser de mucha utilidad en esta etapa. Feature Learning es lo que hace Deep Learning y no se verá en este curso.

39. Describe el proceso manual para generar características.

• Brainstorming

o No juzguen en esta etapa

o Permitan y promuevan ideas muy locas.

o Construyan en las ideas de otros

o No divaguen

o No mantengan conversaciones en paralelo

o Sean visuales

o Vayan por cantidad, la calidad se verá luego

o Otros consejos se pueden consultar aquí (https://challenges.openideo.com/blog/seven-tips-on-better-brainstorming, son los mismos)

• Decidir que features crear

o No hay tiempo infinito

• Crear esos features

• Estudiar el impacto de los features en el modelo

• Iterar

40. Describe el proceso automático para generar características.

• Interacción multiplicativa

o (C = A \cdot B)

o Hacer para todas las posibles combinaciones.

o Es importante mencionar que estas nuevas variables benefician mucho a los regresores lineales, pero no afectan mucho a árboles de decisión (e.g. el Random Forest)

• Interacción de razón

o (C = A / B)

o Tener cuidado con dividir por cero (\to) hay que tomar una decisión

o Hacer para todas las posibles combinaciones.

• Transformar una variable numérica en una binaria.

o Se trata de encontrar el cut-off que maximize tu variable dependiente.

o Muy parecido a lo que hacen algoritmos como el J48 (en su versión comercial se conoce como C5).

o Hay un paquete de R que lo implementa: C50.

• Numérica (\to) bin.

• Otras

o (X^2)

o (\log X)

o etc.

41. ¿Cómo puedo crear una variable binaria a partir de una continua? ¿Cómo decido el punto de corte?

Se trata de encontrar el cut-off que maximize tu variable dependiente. Buscar el valor que divida los datos como me interesan con datos continuos y de ahi puedo hacer feature engineer.

Muy parecido a lo que hacen algoritmos como el J48 (en su versión comercial se conoce como C5).

Hay un paquete de R que lo implementa: C50. Numérica (\to) bin

42. ¿Para qué sirve la selección de características?

El proceso de seleccionar variables antes que ejecutar los algoritmos.

• Realiza cross-validation

o Realizar cross-validation sólo en una parte del proceso (i.e. el modelo) es hacer trampa.

• ¡Cuidado! No hagas feature selection en todos tus datos antes de construir el modelo.

o Aumenta el riesgo de over-fitting.

o Aún realizando cross-validation.

• Hay todo un paquete en sklearn : sklearn.feature\_selection

43. ¿Qué es el filtrado basado en las propiedades de la distribución? Da un ejemplo (Luis)

Filtrado basado en las propiedades de la distribución

• Si hay poca variabilidad, no pueden ser usados para distinguir entre clases.

• Podemos utilizar como medidas de variabilidad a la mediana y al inter-quartile range IQR.

• En sklearn puedes utilizar VarianceThreshold

Filtrado basado en las propiedades de la distribución (Algoritmo)

• Obtenga para cada variable su mediana.

• Obtenga para cada variable sus quartiles, en particular, reste el tercer quartil del primero, para obtener el IQR.

• Realice un scatter-plot entre ambas variables, esta gráfica nos da una visión de la distribución de las variables.

• Eliminemos las variables que tengan "baja variabilidad" i.e. que sean menores que un porcentaje del IQR global.

o e.g. (< 1/5) ó (< 1/6).

• ¡Cuidado! Que las variables individuales tengan baja variabilidad, no significa que unidas con otras variables la tengan. Para una posible solución ver "A practical approach to Feature Selection" de Kira and Rendell, 1992.

44. ¿En qué consiste el filtrado por correlación? (Luis)

• Descrito en "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution de Yu & Liu ICML 2003 (https://static.aminer.org/pdf/PDF/000/335/746/feature\_selection\_for\_high\_dimensional\_data\_a\_fast\_correlation\_based.pdf)

• Obtienes un conjunto de variables no muy relacionado entre sí, pero altamente relacionado a la variable de salida.

45. ¿En qué consiste el Fast-correlation filtering? Describe el algoritmo

• Encuentra una medida de relación entre cada par de variables.

o Aquí usaremos la correlación, el artículo usa otra cosa.

• Encuentra la correlación de cada variable con la variable de salida.

• Ordena las variables según su correlación con la variable de salida.

• Elige la mejor variable (la de hasta arriba).

• Tira las variables muy correlacionadas con esta.

• Repite el proceso.

46. ¿En qué consiste el Forward Selection?

El cual inicia sin variables y va agregando una a una las variables, hasta que no mejora la metrica de evaluación.

• Ejecuta el algoritmo con cada variable (i.e. de manera individual)

• Si tienes (x) número de variables, ejecutas el algoritmo (x) veces.

• Como siempre, usando cross-validation.

• Elige el mejor modelo y quédate con esa variable.

• Ahora, ejecuta el modelo de nuevo, pero ahora con la variable recién seleccionada y con cada variable restante.

• Elige el mejor modelo y quédate con esas dos variables.

• Repite hasta que no mejore el modelo agregando más variables.

47. ¿En qué consiste el Backward Selection?

Empieza con todas las variables en el modelo, y se van removiendo.

• Backward selection es el mismo algoritmo, pero invertido

• kind-of …

• En sklearn, este algoritmo está implementado con el nombre RFE (Recursive Feature Elimination)

48. ¿Cómo podemos usar los filtros ANOVA para seleccionar variables?

• Si la variable tiene una distribución similar para los posibles valores de la variable a predecir, seguramente no sirve para discriminar.

• Compararemos la media condicionada a los valores de la variable de salida.

• Para las variables que tengamos una confianza estadística elevada de que son iguales a lo largo de los valores de la variable dependiente, serán descartados.

• Para eso usaremos métodos ANOVA.

• En sklearn este método está implementado en la sección de "Univariate feature selection" el cuál contiene los métodos SelectKBest , SelectPercentile, SelectFpr (False positive rate test), SelectFdr (False discovery rate test), entre otros. Estos objetos reciben como parámetro la prueba estadística a utilizar, en particular ANOVA está implementado mediante f\_classif.

############

60. ¿Se crean los modelos finales con cross-validation? Si/no y por qué.

############

61. ¿En qué consiste el leakage? Da un ejemplo

No utilizar validation set en hyperparameter tunning.

(Data Leakage is the creation of unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions.)

############

62. Da un ejemplo de cómo Caret puede ayudar en la selección de modelos

############

63. ¿Cuál es la filosofía UNIX para procesamiento de información?

############

64. Principales comando en UNIX

pwd - print working directory

ls - list

man - manual

> - redirect

cd - change directory

mkdir - make directory

rmdir - remove directory

############

65. ¿En que se parecen los “|” en UNIX con “%>%” en R

Llamadas a funciones de tal forma que se vea con un pipeline de ejecuciones.

############

66. ¿Qué diferencia hay entre una contenedor y una imagen en docker?

La imagen contiene las características de la máquina virtual.

Un ejecutable aislado y autónomo que contiene todo lo necesario para correr algún \_software\_, como código, librerías, binarios, variables de ambiente, etc. El \_template\_ de un contenedor.

Un contenedor es una instancia de una imagen. Una imagen una vez ejecutada resulta en un contenedor que corre la pieza de \_software\_ especificada en ella. Un contenedor está completamente aislado del anfitrión y de otros contenedores, accediendo sólo lo que se le comparte explícitamente.

############

67. ¿Para qué sirve docker?

- En desarrollo e investigación

\* Instalar y configurar es costoso

\* Manejar dependencias es doloroso

\* Un equipo puede tener diferencias vastas en sus ambientes de trabajo

\* "En mi máquina corre"

\* A pesar de herramientas como `virtualenv` y `packrat`, diferencias a nivel de sistema pueden causar dolores de cabeza

\* Diferencias en versiones de herramientas y tecnologías pueden causar problemas de compatibilidad

- En producción

\* Compartir y distribuir configuraciones y dependencias es difícil

\* "En mi máquina corre"

\* El ambiente de desarrollo o investigación nunca es exactamente igual al de producción

\* Distribuir cambios puede ser extremadamente doloroso

Docker nos permite resolver estos problemas, pues provee una manera sencilla y portable de \*\*construir y compartir ambientes completos\*\*, además de que corre nativamente en cualquier servidor con Linux (y algunos con Windows), permitiendo su \*\*completo acceso a los recursos del sistema\*\*. Es posible también crear una red virtual para instancias de Docker, permitiendo aislar y escalar fácilmente cada parte de la infraestructura.

Además, a través de DockerHub, Docker provee la capacidad de utilizar ambientes creados por la comunidad, si la herramienta o tecnología es relativamente popular, seguramente existe en DockerHub.

Por último, orquestadores de contenedores como Kubernetes y Swarm nos permiten \*\*manejar y escalar una infraestructura\*\* basada en Docker.

###########

68. ¿Cómo deep learning puede apoyar en el proceso de ingeniería de características?

###########

69. ¿Qué es lo que realizan las capas intermedias de las redes neuronales?