15-06-2021

Ensemble, tipos:

Secuencial: Votación.- utiliza la moda el hard voting. El soft voting trabaja con la media. Si es regresión nos quedamos con la media de los diferentes modelos de regresión

Distintos modelo,

Modo paralelo: Mismo modelo: bagging (distintas muestras con reemplazamiento) y pasting (distintas muestras sin reemplazamiento)

Bootstrap.- se busca que los estimadores sean lo más independientes posible

Bagging, cómo funciona

Boosting.- se ejecuta un algoritmo y con este entrenado vamos a entrenar el siguiente modelo y así sucesivamente

Adabost.- recoge el entrenamiento del anterior modelo

Grid search.-

Pipeline:

le damos los hiperparámetros que queremos para hacer el entrenamiento (quitar columnas, estandarizar, seleccionar las 20 columnas que mejor predicen, …)

Permite probar diferentes modelos con diferentes hiperparámetros, todos a la vez

cv.- cross validation

Simpleimputer.- asigna a los valores missing lo que le diga (una media, etc)

chi2 se le dice que se quede con las x columnas que mejor correlacionen. Nosotros le damos el x

njob=-1; que pruebe todos los núcleos del procesador

simpleImputer por defecto es la media

standarized

16-06-2021

Taller Pensamiento analítico

Tipos de sesgos colectivos:

* Selección.- Sistemas de recolección de datos .
  + De confirmación.- Ignorar deliberadamente evidencia que no se ajusta a una conclusión deseada o preconcebida
  + De supervivencia.- Concentrarse en las personas o cosas que superaron un proceso de selección pasando por alto a aquellas que no lo hicieron. Ignora los fracasos
  + De anclaje.- Confiar demasiado en una parte de la información e interpretar las conclusiones del análisis en torno a esos hallazgos
  + Ausencia de evidencia.- No abordar el impacto de la ausencia de información en las conclusiones analíticas
  + Sesgo de novedad.- darle más peso a la información más reciente que la que tiene más tiempo
  + Sesgo de volumen de información.- Más información no siempre significa una mejor decisión. Sumamos datos que no influyen en el objeto de estudio añadiendo ruido al análisis
  + Sesgo de observación.- Tendencia a buscar en lugares donde se espera obtener buenos resultados o donde es muy conveniente observar
  + Sesgo de financiación.- Tendencia inconsciente a sesgar los modelos, los datos o la interpretación de los datos de forma que favorezcan los objetivos de un patrocinador financiero o empleador
  + Sesgo de muestreo.- La selección de los grupos o de los datos no es representativa(no intencionadamente) del objeto de análisis. Ley de los números pequeños
* Proceso.- Análisis y manejo de datos
  + Sesgo de inducción.- Estimular, indirecta o deliberadamente la respuesta ya en la formulación de la pregunta
  + Sesgo de confianza.- Tener demasiada confianza en nuestras capacidades nos lleva a asumir mayores riesgos y cometer más errores
  + Apofenia.- Sobreestimación del valor de los patrones percibidos en los datos aleatorios. Correlaciones irrelevantes y Falsa Causalidad
  + Sesgo de encuadre.- La forma en que se presentan los datos puede influir en su percepción
  + Sesgo de disponibilidad.- Si algo se puede recordar, debe ser importante, o al menos más importante que los sucesos alternativos que no se recuerdan tan fácilmente
  + Sesgo de escalada irracional.- Justificar el aumento de la inversión de tiempo basándose en la inversión de tiempo anterior y existente
* Sociales.- Toma de decisión, storytelling
  + Pensamiento grupal. Elegir la opción con la que está de acuerdo la mayoría del grupo o ignorar los conflictos dentro del grupo debido al deseo de lograr un consenso
  + Efecto halo.- Aceptar decisiones de personas a las que admiras sin cuestionarlas. O el efecto halo inverso, alguien no te agrada, por lo tanto, no puede tener razón
  + Sesgo de espejo.- Suponer que otros actuarán igual que nosotros al encontrarse en circunstancias similares
  + Estereotipo.- Esperar que un grupo o persona tenga ciertas características sin tener información real sobre la persona o grupo
  + Maldición del conocimiento.- Al comunicarnos con otros individuos, suponer, sin saberlo, que los demás tienen los conocimientos necesarios para entender lo que se dice
  + Punto ciego.- Notar los sesgos cognitivos y motivacionales mucho más en otros que en uno/a mismo/a

Soluciones:

* Habilidades
* Tener tiempo
* Enfoque
* Técnicas
  + Análisis DAFO

Estrategias de mitigación

* Sesgos de selección
  + Utilidad de los datos
  + Utilidad de las fuentes
  + Credibilidad de los datos
  + Lagunas de información
* Sesgos de proceso
  + Técnica de los seis sombreros
* Sesgos sociales
  + Hipótesis alternativa
  + Hipótesis contrapuestas
  + Abogado del diablo
  + Mapeo lógico

Hábitos de los analistas expertos:

1. Cuestionar los supuestos clave
2. Considerar hipótesis alternativas
3. Buscar datos inconsistentes que proporcionen justificación
4. Concentrarse en los impulsores clave que mejor explique lo que ha ocurrido o lo que está por suceder

17-06-2021

El conjunto de entrenamiento se puede repetir pero no el de validación. Con RepeatedStratifiedKold. Siempre será mejor esa opción que kfold solo

nrepeat.- es el número de épocas

nsplit.- número de iteraciones en cada época

Aprendizaje no supervisado.- No hay target

* PCA
* Cluster
* Detección de anomalías
* Densidad de estimación

PCA (Análisis de componentes principales).-

Reducción de dimensiones. Mínimo una dimensión inferior. Se elige la proyección con mayor información

r2(varianza explicada).- mide la discrepancia entre los datos originales y el PCA

Mejora el tiempo de entrenamiento y puede que el score

Elbow es el punto a partir del cual no interesa seguir reduciendo dimensiones

Pca.explained.variance.ratio.-Nos da el mejor resultado y el peor resultado de las proyecciones

Se le puede pedir que reduzca a x componentes o que reduzca las columnas hasta que de un resultado no inferior a un x%

pca=PCA(0.95)

PCA(n\_components=0. 95)

Hace la reducción de número de componentes hasta el valor inmediatamente inferior al marcado, si no existe exactamente el que le hemos dado

Al darle el porcentaje (.95), nos da automáticamente las columnas a las que ha reducido para que el porcentaje de explicación sea el que le hemos dado (.95).

Al darle un número de componentes a los que queremos que baje, hay que mirar el % de explicación del componente que más explica

Si la suma de todos los ejes es superior al % que se ha puesto, se puede seguir bajando alguna dimensión más

**Cluster**

Kmeans= Kmeans(n\_cluster=K)

Con transform nos da la distancia a cada uno de los centroides

Se pueden inicializar los centroides de los clúster

El error se calcula con la inercia. Distancia media al cuadrado entre cada punto del clúster respecto al centroide

En los errores de kmeans la inercia ha de ser lo menor posible y el score el más alto posible

minibatchKmeans.- se tarda menos tiempo pero la inercia será mayor.

Lo mejor es utilizar directamente kmeans

Para elegir el mejor número de clúster calculamos la inercia para cada número que probemos y cuando la inercia deja de mejorar, mantenemos el número de clúster(hemos encontrado el elbow).

Con silohuete nos da el número de clúster más adecuado

Me crea una nueva columna en la que clasifica las filas por las características de las variables que se han utilizado en el cluster. Esta nueva columna sería mi target y a partir de ahí entreno el modelo y veo si es capaz de acertar con esos grupos con algún modelo supervisado.

df[‘nueva\_columna’] = kmeans.predict(X\_news)

Puede ser buena opción si tenemos muchas dimensiones hacer un PCA y trabajar con las principales.

Se puede establecer donde quieres que inicie los centroides

El error del modelo se calcula con la inercia. Cuanta menor inercia, mejor. La inercia hace referencia a la distancia media cuadrática entre cada valor y su cercanía al centroide. Aunque disminuya la inercia no quiere decir que tenga la mejor ubicación de los centroides

El score a veces da negativo. El mejor es el más cercano a cero

Con que K te quedas? Con el que a partir de cual, aumentar no mejora de forma importante la inercia.

El coeficiente silohuette nos da el porcentaje de mejora entre un K y el siguiente. Se aplica el que da mejor porcentaje pero probablemente el siguiente k de también mejor resultado

###########

h2o es una librería a la que se le pide el numero de modelos que queremos que entrene, la métrica que queremos que utilice, el número de segundos que queremos que trate con cada modelo. Con el código de leader nos selecciona el mejor modelo.

Pdk.- categorizar columnas no categóricas

df[“Rango\_edad”]=pd.cut(df[0], 5] # crea para la columna “0” 5 rangos de edad. Crea una nueva columna que se llama “Rango\_edad”

Se le puede dar una etiqueta a cada rango de edad

Parámetro warm\_start.- entrenamiento incremental. Solo funciona bien con RandomForest

Import gc

gc.collect()

para liberar espacio cuando se están haciendo iteraciones

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

Detección de anomalías.- Hay que trabajar con una cota que es siempre subjetiva (establecer a partir de cuantas anomalías similares ya no se considerará anomalía). La cota se establece como un porcentaje por encima de la desviación típica respecto a la media

Para ser considerado anomalía podemos considerar que es en aquellos casos en los que la cota establecida se supera

######

Segmentación de imágenes (los pixeles van del 0 al 255).-

* Reducir la dimensionalidad
* Encontrar objetos diferentes dentro de la imagen

Con kmeans para un número de pixeles podemos obtener mejor % de aciertos

Aprendizaje semisupervisado

De 100 tenemos la clase de 70 y de 30 no

De los que tenemos la clase entrenamos 50 por ej.