Sesiones de vídeo

‘Índice de contenido de los vídeos’

Autor: Guillermo Villarino

Actualizado Diciembre 2024

INTRODUCCION.

Este documento pretende establecer una guía de las distintas partes abordadas en las sesiones de vídeo con el objetivo de hacer más llevadera la navegación por los mismos y poder identificar las partes concretas de interés para cada cual.

CONCEPTOS INICIALES

0 - 1:50 Introducción. Definiciones.

1:50 - 6:45 Diferencias entre estadística clásica y minería de datos. Problemas con la pontencia de los test y esquema de validación por remuestreo. Training-test y concepto de validación cruzada.

6:45 - 7:40 Metodologías en minería de datos.

7:40 - 12:45 Conceptos importantes sobre modelos predictivos. Aprendizaje supervisado, concepto de modelo nulo o predicción a falta de variables explicativas. Concepto de sesgo varianza.

12:45 - 15:20 Complejidad del modelo y concepto de sobreajuste. Modelos muy buenos en training pero malos en test. Capacidad de generalización.

15:20 - 17:50 Preparación de datos de cara a la modelización. Fases: tipologías de variables, valores mal codificados, outliers, missings. Relaciones a priori entre variables. Estrategias de visualización y test estadísticos adecuados según el cruce de las variables.

17:50 - 21:50 Concepto de outlier. Baja incidencia. Estrategias multicriterio para detección de outliers según la simetría de las variables.

21:50 - 24:45 Estrategias de tratamiento de outliers. "Winsorización" y conversión a valor perdido. Ventajas e inconvenientes.

24:45 - 27:55 Datos missing. Concepto. Estrategias para detección de patrones y tratamiento. Eliminación, recategorización e imputación.

27:55 - 32:30 Tipos de imputación de datos missing. Imputación simple por valor central o valor aleatorio. Imputación por modelos multivariantes teniendo en cuenta otras variables del archivo. KNN, regresión…Ventajas e inconvenientes.

32:30 - 35:30 Transformación de variables continuas. Discretización o binning. Transformaciones tipo box-cox, potencias y raíces. Transformaciones para aumentar capacidad predictiva con la variable objetivo.

35:30 - 38.30 Cuantificación de las relaciones entre variables. Test chi-cuadrado. Concepto básico y relación con la teoría de probabilidades. Frecuencias observadas y

esperadas. Hipótesis nula, estadístico de contraste y p-valor. Inconvenientes del valor chi-cuadrado como métrica para evaluar asociaciones.

38.30 - 39:50 Transformaciones del valor chi-cuadrado. Estadístico V de Cramer. Razón de ser, expresión y ventajas.

DEPURACIÓN DE DATOS

Trabajo práctico en Python para la depuración de los datos de vinos.

Depuración_1

0 - 5:00 Diferentes archivos. HTML, cuaderno Python. Pormenores de Python, anaconda y cómo crear un ambiente de trabajo con una versión específica de Python. Opciones para plotly dentro de Jupyter.

5:00 - 11:10 Carga de librerías y lectura de los datos de Vinos. Evaluación inicial de la estructura del dataframe. Comparación con la descripción de las mismas.

11:10 - 16:30 Esquema de trabajo y pasos para comprobar adecuación de los datos. Repaso del flujo de depuración para ordenar ideas.

16.30 - 18.55 Tipos de variables con ***.info()***

18.55 - 20:30 Número de valores distintos de las variables con ***.nunique()***. Identificación de posibles factores o variables no realmente numéricas.

20:30 - 23.40 Histograma de CalifProductor. Cruce con variable objetivo. Decisiones.

23.40 - 24:55 Conversión a factores o variable categórica de variables numéricas con menos de 10 valores distintos. Aplicación de condiciones con pandas. Argumento ***.loc()***

24:55 - 29:55 Tabla descriptiva de las variables con ***.describe()***.

29:55 - 38:10 Inspección gráfica de las variables. Funciones interesantes para pintar el dataset completo distinguiendo por tipo de variables.

38:10 - 41:40 Corrección de errores detectados. Arreglando Etiqueta. Filosofía ***apply(lambda)***, funciones ***str***, ***toupper()***. Reordenar categorías de una variable categórica con ***reorder_categories()***.

41:40 - 42:50 Corrección de valores fuera de rango en Azúcar. Funciones ***replace***, ***inplace***.

42:50 - 44:50 Variable **Alcohol**. Función ***between*** y condiciones por columnas con .loc().

44:50 - 49:20 Variable **Clasificación**. Valor ? y evaluación de la incidencia de clasificación desconocida…posibles patrones. Generación de una categoría adicional para este grupo. Cambio nombres categorías de Región.

49:20 - 51:20 Separación del archivo y extracción (por precaución) de las variables objetivo que son "sagradas" de cara al tratamiento masivo de outliers y missings.

Depuración_2

0 - 7:30 Estudio de valores ***atípicos/outliers***. Repaso del flujo de depuración y filosofía de tratamiento de outliers.

7:30 - 16:30 Distintos criterios en función de la simetría. Posibilidades de gestión. Miss, winsor. Función **gestiona_outliers().** Aplicación general a todas las variables de tipo numérico del input en modo ***'check'***. Evaluación de la incidencia en las salidas.

16:30 - 20:00 Aplicación de la función en modo ***'winsor'***. Evaluación del proceso de winsor. Unión de las variables categóricas al archivo input winsorizado.

20:00 - 29:10 Estudio de los valores **perdidos/missings.** Alternativas de tratamiento: eliminación a priori, no hacer nada, imputación de los valores. Tipos de imputaciones.

29:10 - 34:38 Evaluación de la incidencia por columnas/variables y por filas/registros (creación de la variable prop_missings).

34:38 - 38:20 Patrones de coexistencia de missings entre variables. Escenarios MCAR, MAR, NMAR. Librería **missigno.**

38:20 - 38:50 Eliminación porl lista a priori, evaluación de la pérdida de registros.

38:50 - 42:50 Imputación de los missings. Sklearn y distintos métodos. Definición de los imputadores para continuas/categóricas. Imputaciones simples y por modelos multivariantes. Separación en variables numéricas y categóricas.

42:50 - 47:55 Aplicación de imputaciones simples por media, mediana. Evaluación de los resultados.

47:55- 49:43 Imputaciones por modelos multivariantes. KNN y cadenas de markov con regresión.

49:43 - 50:50 Imputaciones para categóricas. Creación del archivo final imputado con ***concat()*** y limpio. Evaluación de resultados.

50:50 - 56:50 Creación del archivo final depurado y distintas opciones para el guardado de datos sin pérdida de características. HDFS, parquet...

56:50 - 1:00:01 Resumen y reflexiones finales.

CONCEPTOS DE REGRESIÓN LINEAL. RELACIONES Y TRANSFORMACIONES DE VARIABLE EN REGRESIÓN

Conceptos de regresión lineal

0 - 2:40 Introducción a modelos de predicción para una variable continua. Especificación del modelo lineal. Fórmula matemática y predicciones.

2:40 - 7:20 Estimación por mínimos cuadrados ordinarios (OLS). Matriz de diseño. Correlación entre predictores. Colinealidad y problemas que supone en los modelos lineales.

7:20 - 10:20 Tratamiento de variables categóricas en modelos de regresión. Variables indicador o variables Dummy. Ejemplo con el data FEV. Rectas paralelas por cada categoría. Especificación del modelo evitando colinealidad.

10:20 - 12:20 Inclusión de interacciones en los modelos de regresión. Distintas pendientes de las variables continuas en grupos de las variables categóricas.

12:20 - 14:00 Posibles problemas en regresiones. Categorías minoritarias. Evaluación del modelo. Descomposición de la varianza. Sumas de cuadrados total, explicada y residual.

14:00 - 17:10 Salida esperada del modelo de regresión. Summary e interpretación de distintos contrastes sobre modelo y parámetros. Coeficiente de determinación (r2) y r2 ajustado.

17:10 - 21:10 Evaluación de modelos en minería de datos. Problemas con la potencia de los contrastes estadísticos y métodos alternativos de remuestreo. Training-test y validación cruzada como "inferencia 3.0".

21:10 - 22:50 Ejemplo de comparación sesgo-varianza por validación cruzada de modelos con y sin cierta variable. Como evaluar modelos bajo este paradigma.

22:50 - 24:20 Otros criterios de comparación de modelos. AIC y BIC.

Relaciones y transformación de variables

24:20 - 29:15 Introducción a la práctica. Lectura de datos y comprobación de la adecuación de tipos de variables. Arreglos iniciales. Reordenación de categorías de las variables categóricas.

29:15 - 30:15 Descriptivos de las variables y comprobación de distribuciones.

30:15 - 34:20 Visualización de las relaciones entre pares de variables continuas. Gráficos tipo rejilla. Evaluación de las posibles relaciones con la variable objetivo de las variables continuas. Scatterplot y sentido de la influencia. Alternativas de modelización y qué pasa con tantos valores nulos de la variable objetivo.

34:20 - 36:55 Concepto de variables aleatorias de control y sus utilidades.

36:55 - 44:25 Paquete pandas_profiling para estudio descriptivo automático con interesantes salidas uni y bivariante.

44:25 - 45:45 Decisiones para la modelización excluyendo los vinos que no se venden. Implementación. Creación del archivo vinos_compra.

45:45 - 47:15 Creación del input de predictores y separación de la variable objetivo. Razón de ser e implementación.

47:15 - 52:10 Ranking de efectos a priori por V de Cramer. Función Cramer_v, implementación y aplicación a una variable y al input completo con la filosofía apply-lambda.

52:10 - 55:50 Visualización de relaciones a priori con la variable objetivo. Variables categóricas frente a la variable objetivo. Boxplot paralelo, violín, striplot.. Evaluación a nivel marginal. Adelantando sentido de los efectos en relación a la frecuencia relativa a priori.

55:50 - 1:04:00 Transformaciones de variables para aumentar linealidad o relación con la variable objetivo. Super función transformación automática y su filosofía de actuación. Mejores transformaciones para todas las variables de manera sencilla. Aplicación a nivel univariante y para todos los predictores continuas.

1:04:00 - 1:08:15 Creación del dataset con transformaciones y guardado del mismo para posterior explotación. Posibles problemas con el index de los archivos en presencia de filtros!! Concepto de reset_index. [Nota: Cuidado aquí que me dejé varObjCont sin su correspondiente reset_index!!!]

1:08:15 - 1:09:30 Evaluación de las transformaciones. Hay ganancia?

1:09:30 - 1:14:50 Discretización, tramificación o binning de variables. Tipos y su código asociado. Ventajas e inconvenientes. Binning por árboles de decisión. Ejemplo con la variable Azúcar.

1:14:50 - 1:15:50 Resumen de relaciones a priori y transformaciones.

REGRESIÓN LINEAL

1 - 11:50 min: Introducción al doble paradigma de ajuste de modelos en Python. Matrices explícitas vs. Interfaz fórmula.

11:50 - 15:09 min: Particiones trainig-test en python.

15:09 - 26:10 min: Ejemplo de modelización con filosofía formula-data. Salida esperada de regresión lineal e interpretación.

26:10 - 33:35 min: Modelo completo de referencia con fórmula. Evaluación y posibilidades para modelado manual.

33:35 - 36:58 min: Modelo completo de referencia con el paradigma matriz explícita de diseño.

36:58 - 40:22 min: Importancia de variables en modelo.

40:22 - 43:10 min: Proceso backward manual (paradigma formula)

43:10 - 47:30 min: Evaluación de modelos en el conjunto de test.

47:30 - 57:01 min: Evaluación de modelos por validación cruzada repetida. Pormenores.

57:01 - 59:30 min: Modelo con interacción.

59:30 - 1:02:16 min: Problemas de especificación en interacciones y recodificación de variables en medio del proceso de modelización. Consideración de alguna variable tramificada.

1:02:16 - 1:09:40 min: Función general para validación cruzada repetida y comparativa final de todos los modelos manuales. Elección del mejor modelo.

1:09:40 - 1:16:00 min: Ajuste del modelo final en datos completos e interpretación de los parámetros de la regresión lineal.

REGRESIÓN LOGÍSTICA

Conceptos teóricos

0 - 5:45 Concepto general de predicción en clasificación bienaria. Relaciones con modelo lineal. Filosofía de predicción probabilística. Concepto de función de enlace o link. Funciones de distribución.

5:45 - 9:00 Expresión matemática de las probabiliades en regresión logística. Razón de probabilidades u Odds Ratio. Expresión lineal del logit frente a los predictores. Ilustración de la función logística y salidas esperadas de la regresión logística. Probabilidades y clases estimadas.

9:00 - 11:00 Conceptos de odds y odds ratio. Relaciones con las probabilidades. Dualidad en la interpretación.

11:00 - 16:00 Interpretación de los parámetros y magia matemática para obtener los odds ratio desde los parámetros beta del modelo. Caso binario, caso categórico y caso continuo.

16:00 - 19:15 Estimación de parámetros en regresión logística. Diferencia con regresión lineal y sus famosos mínimos cuadrados. Estimación por Máxima Verosimilitud. Métodos de estimación, métodos numéricos iterativos.

19:15 - 22:50 Evaluación del modelo logístico. Contrastes sobre modelo y parámetros en el modelo logístico. Relación con el caso lineal. Pseudo-R2 de McFadden como intento de cálculo de una especie de coeficiente de determinación del caso lineal en regresión logística. Escalas de medida.

22:55 - 27:30 Evaluación de modelos de clasificación. Matriz de confusión del modelo y métricas asociadas. Accuracy, sensibilidad, especificidad, precisision y recall.

27:30 - 30:35 Concepto de curva ROC y evaluación antes de la generación del factor de predicciones o la decisión del punto de corte de la probabiliad estimada. Utilidades como métrica para la comparación de modelos.

30:35 - 33:10 Técnicas de remuestreo. Validación cruzada. Punto de corte óptimo para la probabilidad estimada. Distintas estrategias. Clasificación desbalanceada y sus peligros.

33:10 - 33:50 Extensión del modelo logístico binario al caso multinomial o multiclase para la predicción de una variable categórica con más de dos niveles.

Práctica con Python

0 - 3:00 Introducción al problema. Carga de NuestrasFunciones. Lectura de datos depurados. Comprobación de cosas y reordenación de categorías de las variables categóricas.

3:00 - 4:40 Variables aleatorias de control. Creación del input de predictores y variables objetivo por separado.

4:40 - 12:30 Relaciones a priori con la variable objetivo. Ranking por V de Cramer. Visualización gráfica, mosaicos y boxplot/histograma por niveles de la objetivo.

12:30 - 15:30 Transformación de variables continuas para maximizar relaciones con la objetivo. Función mejor_transformacion. Creación del dataset completo con transformación y guardado del mismo. Evaluación.

15:30 - 16:50 Inspección inicial de la variable objetivo binaria. Frecuencia relativa a priori de las clases. Desbalanceo. Adelantando capacidades del modelo en relación al desbalanceo.

16:50 - 19:20 Esquema training-test. Particiones. Distintos esquemas: formula-data / matriz explícita X – variable objetivo.

19:20 - 23:05 Paradigma fórmula-data. Función ols_fórmula y función logit de statmodels.api. Ajuste del modelo y summary. Problemas de convergencia y

evaluación de cosas raras en los modelos mediante el análisis de los errores de estimación de los parámetros. Warnings de statmodels.

23:05 - 27:00 Diagnostico de problemas. Posibilidades de tratamiento. Falta de datos en alguna clase. Recodificación de categorías. Evaluación del nuevo modelo. Análisis de mejoras.

27:00 - 28:15 Importancia de las variables en modelo y proceso de modelización manual hacia delante.

28:15 - 36:30 Métricas de ajuste en training-test. Ajuste del modelo con sklearn. Matrices explicitas de diseño y particularidades sobre tipos de entrada. Ravel(). Evaluación de todo tipo de métricas. Curva roc y distintas posibilidades para pintarla.

36:30 - 42:05 Proceso manual forward. Interacciones y su interpretación.

42:05 - 45:30 Comparación por validación cruzada repetida. Función cross_val_log. Lista de fórmulas y aplicación masiva a todos los modelos. Evaluación numérica y presentación de boxplots sesgo-varianza de la validación cruzada repetida. Wide to long y creación del data para la visualización.

45:30 - 47:30 Elección del modelo final y distintas posibilidades. Parsimonia vs. Capacidad predictiva.

47:30 - 52:30 Búsqueda del punto de corte óptimo para la probabilidad estimada. Casos accuracy y youden. Relación con la frecuencia a priori de evento. Predicciones como probabilidades con predict. Visualización de las probabilidades estimadas por cada clase. Ojímetro para el pto de corte. Función curva roc.

52:30 - 56:45 Matriz de confusión de las distintas soluciones para distintos puntos de corte. Punto 0.5 y punto de corte óptimo. Ventajas e inconvenientes. Correcciones cuando youden se pasa de movimiento.

56:45 - 1:01:05 Ajuste del modelo en datos completos. Interpretación de los parámetros. Cálculo de los Ors del modelo como exponenciales de los parámetros. Conclusiones del mejor modelo manual por parsimonia.

SELECCIÓN DE VARIABLES EN REGRESIÓN

0 - 1:35 Introduicción a la selección de variables.

1:35 - 5:58 Selección secuencial o clásica de variables. Distintas direcciones y particularidades.

5:58 - 8:08 Selección de variables por regresión Lasso. Restricción sobre el valor de los parámetros y su utilidad como selector de variables.

8:08 - 13:23 Práctica de selección de variables en el conjunto de datos de vinos. Preliminares. Carga de funciones, lectura del archivo todo_cont con transformaciones.

13:23 - 15:40 Generar la matriz explícita de diseño con get_dummies. Adición de contante. Elección de las referencias. Comprobación.

15:40 - 17:14 Modelo manual ganador de regresión lineal.

17:14 - 18:28 Selección secuencial de variables con *mlxtend*. Opciones sobre direcciones y configuración de parámetros disponibles.

18:28 - 23:10 Pruebas de selección de variables con mejor configuración. Evaluación de los resultados obtenidos.

23:10 - 27:10 Configuración por parsimonia. Resultados. Configuración con un número fijo de predictores o efectos.

27:10 - 29:00 comparación por validación cruzada repetida para selección de variables. Función cross_val_selecVar y sus particularidades.

29:00 - 31:50 Generación masiva de interacciones entre todas las variables. Polinomial_features. Aplicación al dataset reducido. Eliminación de interacciones sin sentido.

31:50 - 35:45 Selección de variables con interacciones. Modelos complejos.

35:45 - 38:45 Regresión Lasso. Filosofía e implementación con Python. Solamente variables y transformaciones.

38:45 - 40:10 Regresión lasso con las interacciones entre variables. Evaluación de las interacciones interesantes. Posibilidades.

40:10 - 42:15 Validación cruzada de modelos con interacciones. Comparativa.

42:15 - 44:30 Comparativa final y decisiones.

44:30 - 46:37 Resumen final y alternativas más allá para la selección de variables en Python.

SERIES TEMPORALES

Conceptos teóricos

0 - 1:15 Introducción.

1:15 - 4:00 Qué es una serie temporal? Ejemplos de representación. Particularidades.

4:00 - 10:19 Supuestos para el análisis de series temporales. Estacionariedad. Procesos estocásticos estacionarios en sentido débil. Filosofía modelos de series como un filtro. Condiciones.

10:19 - 14:47 Ruido blanco. Ejemplo serie estacionaria en media. Ejemplo de serie con componentes visibles. Bondades de las series estacionarias de cara a la predicción.

14:47 - 25:40 Componentes de las series temporales. Tendencia, estacionalidad y componente irregular. Descomposiciones "inocentes" aditiva y multiplicativa y su relación con la presencia de heterocedasticidad. Ejemplo de descomposición, cálculo de tendencia por medias móviles. Análisis residual.

25:40 - 32:10 Transformaciones en series temporales. Hacia la estacionariedad. Diferenciaciones y concepto de serie en diferencias. Diferenciaciones estacionales para acercarnos a la estacionariedad de la serie. Ejemplo de camino hacia la serie estacionaria con pasajeros de avión. Estabilización de varianza y diferenciaciones regular y estacional.

32:10 - 40:33 Métodos de suavizado exponencial. Concepto filosófico y el porqué se llaman exponenciales. Suavizado simple y su formulación matemática. Interpretación del parámetro Alpha. Predicciones esperadas por un suavizado simple.

40:33 - 43:08 Suavizado doble o de Holt para series con tendencia. Nuevo parámetro de pendiente beta. Predicciones esperadas tipo recta.

43:08 - 46:30 Suavizado de Holt-Winters para series con estacionalidad. Particularidades y nuevo parámetro gamma. Modelos aditivo y multiplicativo y sus diferencias. Predicciones esperadas por un modelo con estacionalidad, flexibilidad para "curvar" las predicciones.

46:30 - 50:40 Cuantísima fórmula!! Nos centramos en la filosofía y aspectos prácticos para el análisis. Modus operandi en la práctica.

50:40 - 58:08 Concepto de autocorrelación simple y parcial y su importancia en el análisis de series temporales. Correlogramas simple y parcial para identificar series estacionarias y órdenes de los modelos ARIMA.

58:08 - 1:04:34 Modelo Autoregresivo (AR). Memoria larga. Operador de retardo para la formulación, polinomio característico y su relación con la estacionariedad de las series temporales. Ejemplo de modelo AR(1) y su patrón de desaparición de correlaciones en ACF y PACF. Modelo AR(2) no estacionario. Filosofía de extensión de órdenes.

1:04:34 - 1:08:16 Modelo de Medias Móviles (MA). Memoria corta. Patrones en ACF y PACF. Dualidad Ar-MA en la desaparición de correlaciones con los retardos. Formulación y polinomio característico en la parte de los errores.

1:08:16 - 1:14:20 Modelo Mixto ARMA. Conceptos de órdenes p (AR) y q (MA). Ejemplo y visualización de ACF y PACF con superposiciones. Interpretación general de correlogramas y trucos para observación.

1:14:20 - 1:23:32 Procesos integrados. Modelos ARIMA. Estacionariedad a la entrada al modelo. Diferenciaciones en lo regular (d) y en lo estacional (D). El Modelo ARIMA(p,d,q). El modelo estacional SARIMA(p,d,q)(P,D,Q)s. Ejemplo de particularización de fórmulas.

1:23:32 - 1:27:32 Metodología Box-Jenkings de estudio de series temporales. Flujo de análisis. Esquema. Enlace a texto bibliográfico.

Práctica con Python

0 - 4:35 Introducción y esquema de trabajo.

4:35 - 7:40 Conexión Python – R. Importar paquetes o datasets.

7:40 - 9:30 Lectura de datos desde csv. Arreglos para crear serie temporal en Python. Juego con el index fecha.

9:30 - 12:00 Estudio descriptivo de la serie. Visualización de la evolución temporal. Conclusiones sobre los datos. Componentes de la serie: tendencia, estacionalidad y heterocedasticidad y transformación logarítmica.

12:00 - 15:15 Contraste de estacionariedad. Test de Dickey Fuller. Hipótesis nula y conclusiones. Función definida para este test.

15:15 - 24:05 Descomposición de la serie de tipo aditivo y multiplicativo. Extracción de componentes. Resultados de la descomposición en cuanto a estacionariedad.

24:05 - 25:55 Seasonal_plot. Utilidades para la evaluación de la estacionalidad.

25:55 - 30:15 Ir hacia la serie estacional. Logaritmos y diferenciaciones regular y estacional. Evaluación de los residuos en cuanto a la estacionariedad.

30:15 - 37:10 Funciones de autocorrelación simple y parcial. Relación con la estacionariedad de la serie y con los órdenes AR y MA. Test para las autocorrelaciones de los residuos. Test de Ljung-box.

37:10 - 39:00 Métodos de suavizado exponencial: Simple, Doble y estacional de Holt-Winters. Serie logarítmica. Ventanas de training-test.

39:00 - 42:00 Super función para la evaluación de los modelos de series temporales en el esquema training test. Test de autocorrelación y errores en el test en un único gráfico. Función eval_model().

42:00 - 45:15 suavizado exponencial simple. Particularidades y ajuste en Python. Evaluación del modelo. Características del parámetro alfa.

45:15 - 46:20 Suavizado doble de Holt. Series con tendencia. Ajuste y predicciones. Evaluación.

46:20 - 51:20 Suavizado estacional de Holt Winters aditivo y multiplicativo. Ajuste e interpretación de soluciones. Serie no logarítmica y comparativa de los errores con suavizados.

51:20 - 52:45 Gráficos de residuos (autocorrelaciones) para los modelos de Holt-Winters. Interpretación.

52:45 - 56:15 Modelos ARIMA. Función resid_check(). Características. Función eval_modelAarima(). Proposición de órdenes AR y MA. ACF y PACF para la serie doblemente diferenciada.

56:15 - 59:52 Modelos ARIMA manuales. Funciones de ajuste y parámetros. Acceso a los residuos y evaluación. Primer modelo tentativo y evaluación.

59:52 - 1:01:45 Modelo ARIMA 2. Ambas partes AR. Evaluación y comparativa. Modelo ARIMA 3. Partes ARMA. Evaluación y comparativa global.

1:01:45 - 1:05:40 Modelo ARIMA automático. Particularidades y posibilidades paramétricas. Solución para el problema y comparativa con los anteriores.

1:05:40 - 1:07:40 Resumen final.

REDUCCIÓN DE DIMENSIONES. PCA Y FA

Conceptos teóricos

0 - 1:20 Introducción a técnicas no supervisadas. Ausencia de variable objetivo y finalidades.

1:20 - 4:20 Dos utilidades clásicas del PCA. Interpretación y preproceso de predictores para modelo supervisado.

4:20 - 5:20 Diferencias entre PCA (componentes en función de las variables originales) y FA (variables originales como expresión lineal en los factores)

5:20 - 6:40 Salidas esperadas de la reducción de dimensionalidad. Cargas (variables vs. Componentes/factores → Interpretación de componentes/factores en relación a las variables originales) y Saturaciones (registros vs. Componentes/factores → Interpretación de la posición de los registros en el espacio de componentes/factores y Solución del proceso para explotación posterior)

6:40 - 9:08 Solución matemática del PCA. Descomposición en valores singulares y características de las componentes.

9:08 - 13:17 Criterios para la retención de Componentes. Varianza, Kaiser, sedimentación…

13:17 - 21:30 Comprobación de la adecuación muestral como proceso de evaluación a priori de cara a una reducción de dimensionalidad. Matriz de correlaciones, su determinante, Test de esferidad de Bartlett y KMO

21:30 - 26:06 Interpretación de soluciones. Gráfico de cargas, gráfico de saturaciones y biplot.

26:06 - 31:40 Ejemplo interpretación biplot archivo mtcars.

31:40 - 35:20 Mejora de interpretabilidad del PCA. Modelo FA y sus posibilidades de optimización para la estimación de cargas.

35:20 - 36:40 Variabilidad explicada. Comunalidades y unicidades. Interpretación por variable.

36:40 - 40:10 Distintas soluciones para FA, concepto de rotación de Factores y sus ventajas.

Práctica con Python

40:10 - 41:36 Introducción a las posibilidades del PCA y FA con distintas librerías de Python. Pros y contras de cada una.

41:36 - 43:40 Esquema de trabajo. Carga de librerías necesarias. Lectura de los datos de cities. Objetivos específicos.

43:40 - 47:30 Evaluación de la adecuación muestral de los datos. Correlaciones, Bartlett y KMO. Qué pasa con Área?

47:30 - 49:05 Escalado de los datos.

49:05 - 51:05 Ajuste del PCA con sklearn. Solución del PCA, matriz de puntuaciones registros vs. Componentes. Argumento n_components, reducción "real" de la dimensionalidad.

51:05 - 57:40 Componentes a retener. Screeplot y distintas visualizaciones con psynlig. Supergráfico completo para selección de componentes. Interpretación de cargas 1 dimensión y 2 dimensiones. Gráfico de puntuaciones con nombre de las ciudades y plotly.

57:40 - 1:00:50 Biplot e interpretación completa. Otras visualizaciones.

1:00:50 - 1:05:50 Ejemplo con el paquete pca. Un proceso muy completo. Acceso a distintos elementos de la solución PCA. Cargas, componentes, saturaciones (esta es la solución de cara a una explotación posterior). Biplot bonito!

1:05:50 - 1:10:00 Análisis Factorial AF. Librerías necesarias y ajuste de modelo. Configuración de parámetros. Rotaciones. Varianza de los factores, comunalidades y unicidades por variable. Cargas y puntuaciones.

1:10:00 - 1:12:45 Biplot para FA. Evaluación de la rotación varimax. Abrir la posibilidad de correlación entre los factores. Rotación oblicua Promax. Alienación mucho más clara.

CLUSTERING

Conceptos teóricos

0 - 4:10 Filosofía del clustering. No supervisión del proceso. Subjetividad y distintos criterios que fijar a priori. Distancias, evaluación de la bondad y algoritmo de optimización.

4:10 - 7:05 Tipos de distancias y sus particularidades.

7:05 - 11:20 Tipos de Clustering. Jerárquico vs. Particional.

11:20 - 19:16 Algoritmo k-means. Filosofía y particularidades. Ilustración del proceso de optimización con un ejemplo de juguete.

19:16 - 29:00 Clustering jerárquico. Características y tipos. Inconvenientes. Ejemplo del proceso jerárquico con un ejemplo de juguete. Concepto de linkage como extensión de la distancia cuando tenemos grupos. Tipos de linkage.

29:00 - 31:15 Idea del método híbrido hierarchical k-means. Reflexiones finales. Otros métodos de clustering en sklearn.

Práctica con python

0 - 1:50 Introducción y esquema de trabajo

1:50 - 6:50 Datos simulados. Función make_blobs y representación a priori.

6:50 - 11:20 Escalado de datos. Clustering jerárquico y pruebas con distintos linkage. Función plot_dendogram.

11:20 - 18:50 Métricas de evaluación de clustering. Evaluando el clustering jerárquico.

18:50 - 24:30 Visualización y comparación con la verdad verdadera. Cálculo de centroides bajo esquema jerárquico. Matriz de confusión.

24:30 - 30:20 Clustering k-means. Ajuste, acceso a components de la salida. Evaluación y comparación con la verdad verdadera. Matriz de confusión.

30:20 - 35:50 Super función para evaluar el número de grupos por k-means, sree_plot_kmeans.

35:50 - 38:30 Ejemplo de países.csv. Observación y arreglos sobre el dataset e introducción al problema. Escalado de datos.

38:30 - 41:00 Clustering jerárquico con distintos tipos de linkage. Decisiones sobre el número de grupos. Obtención de cluster_labels y evaluación del modelo jerárquico.

41:00 - 46:50 Visualización de resultados. Proyecciones sobre pares de variables en el plano. Proyección en 3D. Evaluación de las características de los grupos. Centroides.

46:50 - 49:30 Clustering k-means. Elección del número de grupos en k-means. Ajuste y acceso a componentes de la solución. cluster_labels de sitintas formas. Evaluación de métricas. Visualizaciones.

49:30 - 52:00 Comparativa del acuerdo entre los grupos formados por el jerárquico y por el k-means.

52:00 - 59:00 Obtención de datos actuales. Scrapping al Banco Mundial mediante la api wbdata. Características y posibilidades. Buscador de indicadores. Selección de fechas, países e indicadores y descarga del dataset.

59:00 - 59:50 Eliminación de valores perdidos. Evaluación de la pérdida de registros.

59:50 - 1:01:12 Clustering k-means y evaluación del número de grupos. Comparativa con los datos anteriores de países. Ajuste y acceso a componentes de la solución. Evaluación de métricas.

1:01:12 - 1:03:20 Visualizaciones. Proyecciones sobre pares de variables y en 3D. Conclusiones.

1:03:20 - 1:06:03 Visualización en formato biplot con componentes principales. Ajuste de un PCA y acceso a las dos primeras componentes. Creación del nuevo dataset para proyectar. Biplot y pequeñas conclusiones.