Minería de datos y modelización predictiva

‘Índice de contenidos’

Autor: Guillermo Villarino

Actualizado Diciembre 2024

Contenido

PARTE I: DEPURACIÓN DE DATOS Y MODELOS DE REGRESIÓN.

Esta primera parte se centra en la preparación del conjunto de datos de cara a la modelización predictiva de tipo supervisada y la presentación de los modelos clásicos de predicción para variables continuas mediante regresión lineal y variable dicotómica mediante regresión logística.

INTRODUCCIÓN A LA MINERÍA DE DATOS. CONCEPTOS INICIALES

Presentación de definiciones de conceptos fundamentales en minería de datos en lo relativo a modelos de predicción. Sesgo-varianza de las estimaciones, diferencias con estadística clásica, esquema de validación por remuestreo y sobreajuste.

DEPURACIÓN DE DATOS

Procedimiento de evaluación del conjunto de datos y características de las variables implicadas. Tipos, valores fuera de rango, missings no declarados, outliers y missings. Técnicas para el estudio y depuración del conjunto de datos. Gestión de outliers, imputaciones simples y multivariantes.

RELACIONES Y TRANSFORMACIONES DE VARIABLE EN REGRESIÓN

Estudio de las relaciones entre las variables independientes o predictores y la variable objetivo de la modelización. Cruces de variables y estudio bivarinate. Distintas técnicas visuales y estadísticas para establecer relaciones marginales.

REGRESIÓN LINEAL

Modelos lineales de regresión. Concepto de modelo predictivo en general y particularización matemática al caso lineal. Estimación de parámetros del modelo por mínimos cuadrados ordinarios (OLS) y sus características. Colinealidad y tratamiento de variables de tipo categórico. Diferencias en código Python entre el ajuste mediante interfaz fórmula y ajuste mediante matrices de diseño explícitas. Presentación de las librerías de statmodels y sklearn para el ajuste de modelos de regresión, ventajas e inconvenientes. Esquema de modelización manual y evaluación por training-test y por validación cruzada repetida. Elección de modelo final e interpretación de los parámetros de la regresión.

REGRESIÓN LOGÍSTICA

Modelo clásico de clasificación binaria mediante regresión logística. Introducción e idea filosófica para estimación probabilística. Concepto de función de enlace y alternativas comunes. Estimación de parámetros por Máxima Verosimilitud y sus diferencias con Mínimos Cuadrados. Interpretación de parámetros. Concepto de odds y odds ratio. Métodos para la evaluación de la bondad de ajuste del modelo logístico. Pseudo R2 de Mc Fadden, matriz de confusión y sus métricas asociadas (accuracy, sensibilidad, especificidad, precisión, recall…). Concepto de curva ROC.

SELECCIÓN DE VARIABLES EN REGRESIÓN

Métodos automáticos de selección de variables para la modelización predictiva. Selección secuencial hacia delante (forward) y hacia atrás (backward), diferentes opciones en Python. Pequeña introducción a la Regresión Lasso y su poder como selector automático de variables relevantes de cara al modelo, diferentes opciones con efectos transformados e interacciones.

PARTE II: SERIES TEMPORALES Y MODELOS NO SUPERVISADOS

En la segunda parte se abordan los métodos de estudio y predicción de series temporales univariantes clásicos como introducción a los estudios temporales y las dos principales técnicas no supervisadas (no hay variable objetivo específica) para reducción de dimensiones y segmentación de los datos.

SERIES TEMPORALES

Concepto de serie temporal y su relación con los procesos estocásticos. Concepto de estacionariedad y su importancia. Componentes de series temporales. Estudio

descriptivo y descomposiciones aditiva y multiplicativa. Test de estacionariedad de Dickey Fuller. Métodos de suavizado exponencial (simple, doble y Holt Winters) particularidades. Estudio de funciones de autocorrelación simple y parcial y patrones autoregresivos (AR) y de medias móviles (MA) para la identificación y proposición de órdenes en modelos ARIMA. Operador de retardo B, polinomio característico y su relación con la estacionariedad del proceso. Modelos ARIMA manuales y automáticos. Metodología Box-Jenkins para el estudio de series temporales univariantes.

REDUCCIÓN DE DIMENSIONES. PCA Y FA

Características de modelos no supervisados. Reducción de dimensionalidad del conjunto de datos y sus posibles objetivos. Interpretación vs. Preproceso de cara a posterior explotación. Dos métodos para la reducción de dimensiones Análisis de Componentes Principales (PCA) y Análisis Factorial (FA). Diferencias en modelo matemático y particularidades. Evaluación de la adecuación muestral como proceso a priori para valorar la calidad de las soluciones. PCA y métodos para evaluación del número de componentes a retener en la solución. Matrices de cargas y puntuaciones y sus utilidades. Estimación de cargas y descomposición en valores singulares. Propiedades. Evaluación de soluciones, gráfico de cargas, gráfico de puntuaciones y biplot. Interpretaciones en el plano de las dos primeras componentes. Análisis factorial. Modelo matemático, distintas estimaciones de las cargas y rotaciones de la solución. Comunalidad y unicidad. Ventajas interpretativas.

SEGMENTACIÓN DE LA POBLACIÓN. CLUSTERING

Agrupación no supervisada o segmentación del conjunto de datos. Características y objetivos. Métodos de clustering jerárquico y no jerárquico. Concepto de linkage, dendograma y métodos numéricos y gráficos para evaluación de características de los grupos formados. Centroides y biplot o proyecciones. Métricas de evaluación de la solución a falta de variable "supervisora": variabilidad interna y silueta. Evaluación del número de grupos óptimo para la solución. Solución final como factor que indica el grupo predicho para cada registro.