Minería de datos y modelización predictiva

Profesor: Lorenzo Escot

Explicando el contenido de esta semana

Hola, soy Loren Escot profesor de econometría en la Facultad de Estudios Estadísticos y encargado de presentaros esta semana la **última de las sesiones del Módulo de Minería de Datos y Modelización Predictiva**

Os presentaré **dos aplicaciones concretas de la modelización predictiva** en las que se utilizan muchas de las técnicas estadísticas de análisis de datos que ya habéis visto en sesiones anteriores. De hecho mi sesión es la última de modelización predictiva antes de comenzar con el módulo de machine learning, y de hecho se encuentra **entre los dos mundos**, el de la **modelización "inferencial" tradicional** (dónde el mejor modelo a emplear en nuestras predicciones es el que se ajusta mejor al tipo de variable dependiente y, sobre todo, a los supuestos que se hagan sobre la distribución del término de error) y la **modelización "algorítmica"** típica del ***machine learning*** (dónde los supuestos del término de error juegan un papel secundario y la competencia entre modelos es por saber cuál predice mejor).

Yo utilizo en mis exposiciones, los modelos de regresión logística y los modelos de regresión lineal con alguna variación. Lo que cuento son dos aplicaciones concretas. Pero todo lo que yo os comente de las dos aplicaciones es también válido para ser aplicados con otros modelos predictivos basados en el aprendizaje automático que comenzaréis a ver cuando termine yo esta semana.

El contenido de esta semana está dividido en dos partes o sesiones, que se corresponden con las dos aplicaciones que vamos a ver: **Scoring o puntuación de riesgo**, y el análisis de **datos espaciales**

Todos los videos, materiales y scripts de Python que tenéis en el la página web están organizados según estas dos aplicaciones: **Scoring** y **Espacia**l.

Recomiendo **comenzar con la aplicación de Scoring**, a lo largo de los primeros días de la semana iré completando la información de los videos y los apuntes en el foro incidiendo en las tres o cuatro ideas fundamentales de la construcción de los modelos de scoring.

A partir del jueves comenzaré a comentar la segunda aplicación, las de los modelos de dependencia espacial

Un saludo y buena semana

Loren Escot

El índice de los videos es el siguiente:

Introducción

- T1: Presentación General del profesor y Contenidos

Videos Scoring

- T2: Introducción a los modelos de Puntuación o Scoring de riesgos
- T3: La Tarjeta de Puntuación de Riesgos
- T4: Definición de la variable objetivo Impago
- T5: La Selección de Variables Predictoras mediante estadísticos de concentración
- T6: Estimación del modelo de predicción de probabilidad de impago
- T7: Diagnosis del modelo(1): la Matriz de Confusión
- T8: Diagnosis del modelo(2): La Curva ROC
- T9: Diagnosis del modelo(3): Sesgos de selección muestral
- T10: Score o puntuación de calidad crediticia
- P1: Práctica de Scoring con Python (parte1)
- P2: Práctica de Scoring con Python (parte2)
- P3: Práctica de Scoring con Python (parte3)

Videos Datos Espaciales

- T11: Introducción a los datos Espaciales: patrones y dependencia espacial
- T12: GIS: sistemas de Información Geográfica
- T13: Cartografías y Mapas de Coropletas
- T14: Cartografías con Python
- T15: Matrices de Pesos Espaciales y Autocorrelación Espacial
- T16: Regresión Espacial con Python

Sesión 1: Scoring o puntuación de riesgo

1. **Introducción al concepto de Riesgo de Crédito** y a los modelos de Valoración o Puntuación del riesgo de crédito
2. Etapas en la construcción de una **tarjeta de puntuación de riesgo** (Scorecard)
    2.1. Selección y depuración de datos de clientes (datos faltantes, datos atípicos,...)
    2.2. Definición de la Vble objetivo y la ventana temporal de observación de impagos
    2.3. Selección de Variables explicativas: tramificación, agrupación, selección (WOE, Information Value, Gini y otros métodos de selección) y transformación WOE de las variables predictoras
    2.4. Estimación y diagnosis de modelos de probabilidad: aplicación con regresión logística
3. El problema del sesgo de **selección muestral y la inferencia de denegados**
4. **Construcción de las Scorecard**, elección del umbral de riesgo óptimo y validación final del modelo

En primer lugar, veremos una aplicación de los modelos predictivos en la "**gestión de riesgos**", más concretamente el **Scoring o puntuación de Riesgo de Crédito**. Estas técnicas son aplicables también al cálculo de las Primas de Riesgo o ; a los modelos para la **detección de Fraude**; los **modelos de fuga de clientes** (Customer Churn) ..... y en general a los modelos de **puntuación de riesgo de evento**.

Revisaremos las diferentes fases del análisis de riesgos, los métodos de estimación y diagnosis de los modelos de probabilidad; y repasaremos aspectos fundamentales a tener en cuenta en este tipo de análisis como son el **sesgo de selección muestral**, la **tramificación de variables continuas, la transformación WOE de variables** y los métodos de selección de variables explicativas utilizando cr**iterios de concentración**.

Os he dejado la siguiente documentación en el campus virtual:

- *Apuntes de Scoring de Riesgo De Crédito.pdf*: Son unos apuntes que he preparado con los contenidos de la sesión (como los he preparado yo os diré que están muy bien, por lo que se agradecen comentarios para su mejora)
- *CreditScoring_BigDataUCM.pdf:* Es la presentación de diapositivas que utilizaré en los videos
- Naeem Siddiqi.pdf: Es un manual de referencia básico, no muy técnico, para que el que lo necesite profundice en todo el proceso de construcción de tarjetas de puntuación

Detrás de estos modelos de puntuación del riesgo está la **regresión logística**, y como ya la habéis visto anteriormente yo hago sólo un recordatorio de estos modelos.

Así que me centro más en la **metodología de la construcción de los modelos de puntuación de riesgos**. Se trata de encontrar el mejor modelo para **separar a buenos y malos clientes**, a clientes que si les concedemos un préstamo nos lo devolverá sin problemas y a clientes que serán malos pagadores (incurrirán en **impago**). No queremos conceder préstamos a malos clientes, por eso se utiliza un modelo de puntuación del riesgo de crédito, para intentar predecir si un cliente al que tengo que conceder o no conceder un préstamo será o no buen pagador.

En los apuntes y los videos nos centraremos en la **puntuación de riesgo de nuevos clientes** (**scoring de admisión**). Individuos que nos solicitan un préstamo pero de los que no tenemos ninguna información, sólo la que le preguntemos a ellos en el momento para decidir si le concedemos o no el crédito.

Como digo, detrás de la puntuación del riesgo hay un modelo de regresión, un modelo de probabilidad, la logística, o cualquier modelo que sirva para estimar probabilidades de que suceda algo: en nuestro caso el impago (también valdrían árboles, redes, etc… o cualquiera de los denominados **modelos de clasificación** binaria)

Dicho modelo de probabilidad utilizará una serie de variables sobre el nuevo cliente que ayuden a predecir su probabilidad de impago.

Una cuestión importante es que normalmente de los nuevos clientes no tengo información, por eso tengo que preguntarle directamente al nuevo potencial cliente por las variables que necesito para evaluar su riesgo. Por eso es fundamental elegir **pocas variables predictivas que ayuden mucho a separar a los buenos de los malos clientes**.

Os presento en este sentido los **WoE y el Information Value**, que son una medida basada en la concentración de los malos clientes en determinadas categorías de las variables explicativas. Estos criterios de concentración ayudan a hacer una selección inicial de variables candidatas a formar parte del modelo de predicción de la probabilidad de impago.

hay más…. pero ya iremos hablando de ello durante la semana

Sesión 2. Datos Espaciales y Modelos de Econometría Espacial

1. **Introducción a los Sistemas de Información Geográfica (GIS) y representación de datos georeferenciados**
    - 1.1. Mapas, escalas, proyecciones y sistemas de coordenadas
    - 1.2. Referencias: trabajar con cartografías (shapes) con Geopandas
    - 1.3. Tipos de datos Espaciales o georeferenciados: puntos, líneas y polígonos
    - 1.4. Ejemplo: los mapas de corpletas con geopandas (Choropleth maps)
2. **Midiendo la Dependencia Espacial**
    - 2.1. La Ley de Tobler o "Primera Ley de la Geografía"
    - 2.2. Heterogeneidad Espacial vs Dependencia Espacial
    - 2.3. Econometría Espacial: Dependencia Espacial en modelos Econométricos
    - 2.4. La autocorrelación Espacial
    - 2.5. Matriz de Vecindad: contigüidad y distancias
    - 2.6. Variables con Retardo Espacial
3. **Análisis Exploratorio de Datos Espaciales**
    - 3.1. Midiendo la dependencia Espacial: estadísticos locales y globales
    - 3.2. ***I*** de Moran, ***C*** de Geary, ***G(d)*** de Getis y Ord.
    - 3.3. El gráfico de Moran y contrastes de Significatividad de ausencia de correlación espacial
    - 3.4. Agrupación espacial y mapas LISA (Local Indicator of Spatial Association)
4. **Modelos de Econometría Espacial: Especificación y diagnosis**
    - 4.1. Modelo de Retardo Espacial
    - 4.2. Modelo de Error Espacial
    - 4.3. Modelo de Durbin
    - 4.4. Otras especifcaciones
    - 4.5. Contraste de Hipótesis entre Diferentes Especificaciones
    - 4.6. Diagnosis de los Errores
    - 4.7. Efectos directos y efectos indirectos

En la segunda de las aplicaciones dejamos los "riesgos" y nos adentraremos en el apasionante mundo de la "estadística espacial". Veremos qué tienen de particular los **datos espaciales**, qué son los **Sistemas de Información Geográfica** (GIS), y presentaremos algunas herramientas para la elaboración de mapas y el análisis de la distribución geográfica de datos espaciales. A partir de ahí nos centraremos en el análisis de la **autocorrelación espacial** y en los **modelos de econometría espacial.**

Os dejo también la documentación sobre esta segunda de las aplicaciones dedicada a la estadística espacial, que luego será comentada en los videos.

- *Econometría Espacial_BigData_Python.pdf*: es el archivo pdf con la presentación que utilizo en los videos
- *Bibliografía Espacial7z.7z* : archivo (comprimido con 7zip) que contiene un artículo de Ignacio Alonso Fernández-Coppel que a mi me ayudó en su día a entender que eran los GIS y los diferentes sistemas de proyección. También hay un artículo breve que describe los principales modelos de econometría espacial y un excelente manual para el análisis de datos espaciales

Los datos espaciales son datos que incluyen un índice o referencia de su localización. En los videos hablo primero de forma muy resumida de los diferentes

sistemas de información Geográfica GIS, sistemas que son necesarios para hacer una representación de los datos espaciales geo-referenciados...Y ya más adelante me meto con el concepto de la autocorrelación o dependencia espacial, fundamental en todo el análisis de datos espaciales.

A modo de resumen, los modelos de regresión espacial son modelos de regresión con una variable dependiente y un conjunto de variables independientes o explicativas, entre las que se incluye como una variable explicativa más esa dependencia espacial, es decir, la situación en un entorno cercano o próximo a cada objeto espacial.... ya los vamos comentando.

Scripts de Python

En ambos casos tanto la sesión 1 (Scoring) como la sesión 2 (Espacial) primero se presentarán los principales conceptos desde un punto de vista teórico, y después ejecutamos aplicaciones prácticas de ambos temas en Python.

En Scoring os dejo un único script (un jupiter notebook), listo para abrir y ejecutar. También os dejo el conjunto de datos *germancredit.csv* que se utiliza en la práctica y que tendréis que cargar para poder realizarla.

En los tres videos en los que os cuento esta práctica de scoring con Python os muestro también como he creado un entorno de desarrollo específico para esta práctica en el que he instalado entre otras la librería **optimalbinning**, que es la librería que utilizaré para aplicar toda la metodología que os he presentado en las clases teóricas.

Para la segunda **parte sobre Estadística Espacial** os dejo, por un aparte un fichero comprimido *cartografíasPython.7z* Es importante que los descomprimáis respetando la estructura de carpetas, una de **cartografías** y otra de **datos**, porque luego hago referencia a ellas dentro de las prácticas. Os he dejado también 2 scripts (también cuadernos de jupiter)

- *1_IntroRepresentación Espacial*: donde os muestro como cargar cartografías y realizar mapas de coropletas fácilmente utilizando **geopandas**
- *2_Regresión Espacial*: aquí os muestro como cuantificar la auto correlación espacial y como incluirla en vuestros modelos predictivos (yo me limito a verlo en un modelo de regresión lineal sencillo) utilizando **Pysal**

Creo que los videos ayudarán, y siempre podemos ir utilizando el Foro para ir solucionando dudas y cuestiones que os surjan.

La tarea de evaluación es un cuestionario tipo test.

Se trata de un cuestionario de repaso, así que **os recomiendo que lo vayáis haciendo a medida que vais estudiando los videos y la documentación**. No os preocupéis (mucho) si os equivocáis a la hora de contestar el cuestionario. Tenéis dos oportunidades para contestar y mandar a evaluar el examen. Mientras no mandéis a corregir el examen podéis entrar y salir las veces que queráis y dejar guardada las respuestas que ya tengáis claras, cuando volváis a entrar se **os dirá que volvéis al último intento, en realidad es a la última vez que grabasteis el cuestionario**.

Hasta que no mandáis a corregir el cuestionario no agotáis vuestra primera vuelta. Una vez corregida la primera vez, se os indican los fallos. Entonces tenéis otra segunda oportunidad para corregir las respuestas incorrectas y mandarla a corregir de nuevo

No hay tiempo (bueno la fecha límite de entrega que especifique el coordinador de master). Así que pensarlo bien y si no terminái, grabar la sesión, que siempre podrás volver a entrar al cuestionario y continuar contestando (se respetan las respuestas que hubierais contestado ya en la última vez que grabasteis).

Para hacer dos de los cuestionarios necesitaréis cargar una serie de datos que os hemos dejado en un Excel aparte

- *DatosPractica_Scoring.xls*, necesitaréis importar estos datos para poder contestar a esas dos preguntas.