Ciencia de datos

Instructor: Victor Muñiz Sánchez

(victor_m@cimat.mx)

Asistente: Víctor Gómez Espinosa

(victor.gomez@cimat.mx)

Descripción: En este curso se mostrarán métodos básicos de aprendizaje máquina y reconocimiento estadístico de patrones para el análisis de datos multivariados.

Objetivo: Mostrar los métodos básicos de aprendizaje supervisado, no supervisado, y métodos de visualización para datos en alta dimensión. Se hará especial énfasis en el uso computacional y aplicaciones en ciencia de datos.

Horarios: Clases regulares los martes y jueves de 11 a 12:30 hrs. Sesiones especiales y de ayudantía se programarán según las necesidades del curso en un horario de 14:30 a 17 hrs.

Temario

1. Métodos de visualización y reducción de dimensión
   - *a*) Técnicas básicas de visualización
   - *b*) Métodos de proyección y reducción de dimensión
   - *c*) Métodos basados en componentes principales
2. Métodos de aprendizaje no supervisado
   - *a*) El concepto de disimilaridad
   - *b*) Clustering
     - clustering jerárquico
     - K-medias y métodos relacionados
   - *c*) Aprendizaje de variedades (manifold learning)
     - Kernel PCA

- Multi-dimensional Scaling (MDS) e Isomap
- t-distributed Stochastic Neighbor Embedding (t-SNE)
- Spectral embeddings y clustering espectral

3. Métodos de aprendizaje supervisado

   a) Teoría de decisión estadística

   b) Análisis discriminante lineal y cuadrático

   c) Regresión logística

   d) Redes neuronales

      1) Hiperplanos separadores y el algoritmo perceptron
      2) Redes neuronales multicapa

   e) Hiperplanos separadores óptimos y Máquinas de Soporte Vectorial

   f) Selección de modelos y regularización

   g) Modelos aditivos y métodos relacionados

      - Arboles de decisión
      - Boosting
      - Random Forest

Pre-requisitos: Modelos estadísticos (inferencia y regresión), álgebra lineal, cálculo en varias variables, fundamentos de optimización (deseable), conocimientos de programación estructurada y orientada a objetos (deseable).

Software: se usará `python` principalmente. Se usará también `ggobi` para algunas visualizaciones (`http://www.ggobi.org/`)

Evaluación:

- Tareas (50 %)
- Proyecto final (50 %)

Tareas, reportes y proyecto:

- Todas las tareas, así como reportes de proyecto se harán en LaTex. Para esto, se pueden usar las dos plantillas/ejemplo que se proporcionan: `tarea_plantilla.tex` y `ejemplo_proy.tex`. Se entregarán impresos en formato PDF.
- Calificación de tareas. Se hará en escala de 0 a 100: $100 \times e^{0.15t}$, donde t son los días de retraso.
- Los archivos que se requieran entregar (incluyendo código), se nombrarán de forma: `tareaNUM_nombre.pdf`, `tareaNUM_nombre.ipynb` etc...
- Atención especial en la *forma* del reporte: claro, bien estructuarado, citas y referencias correctas, etcétera. Ten en mente que tu reporte lo leerá alguien mas, así que debe ser claro en todos los sentidos.

Referencias:

- T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning. Data mining, inference and prediction.* 2nd. edition. Springer, 2009.
- A. J. Izenman. *Modern Multivariate Statistical Techniques. Regression, classification and manifold learning.* Springer, 2013.
- R.O. Duda, P. Hart, D.G. Stork. *Pattern classification.* 2nd. edition. Wiley, 2000.
- Otras que se irán mencionando y muchos papers.