

Formulario de Aprobación Curso de Posgrado

Asignatura: Árboles de Clasificación y Regresión y Bosques Aleatorios

(Si el nombre contiene siglas deberán ser aclaradas)

Profesor de la asignatura ¹: Prof. Jean-Michel Poggi, Professeur des Universités, (Univ Paris Descartes & LMO, Univ Paris-Sud, Orsay).

(título, nombre, grado o cargo, Instituto o Institución)

Profesor Responsable Local ¹: Dr. Mathias Bourel, Grado 3, IMERL.

(título, nombre, grado, Instituto)

Instituto ó Unidad: Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia".

Departamento ó Area:

¹ Agregar CV si el curso se dicta por primera vez.

(Si el profesor de la asignatura no es docente de la Facultad se deberá designar un responsable local)

Horas Presenciales: 24

(se deberán discriminar las mismas en el ítem Metodología de enseñanza)

Nº de Créditos: 5

(de acuerdo a la definición de la UdelaR, un crédito equivale a 15 horas de dedicación del estudiante según se detalla en el ítem metodología de la enseñanza)

Público objetivo y Cupos: Estudiantes de la Maestría en Ingeniería Matemática, Maestría en Matemática, Ingenieros y otros estudiantes interesados. Sin cupo.

(si corresponde, se indicará el número de plazas, mínimo y máximo y los criterios de selección. Asimismo, se adjuntará en nota aparte los fundamentos de los cupos propuestos. Si no existe indicación particular para el cupo máximo, el criterio general será el orden de inscripción en el Depto. de Posgrado, hasta completar el cupo asignado)

Objetivos:

Introducir los aspectos metodológicos de CART & Random Forest así como algunas modernas contribuciones. Aplicar dichas técnicas a conjuntos de datos reales e interpretar los resultados obtenidos. Acercar el estudiante al empleo de los paquetes para el análisis estadístico de datos disponibles en el ambiente de desarrollo de software libre R (<http://www.r-project.org/>).

Conocimientos previos exigidos: introducción a la probabilidad y estadística, álgebra lineal, cálculo matricial.

Conocimientos previos recomendados: teoría de la inferencia estadística

Metodología de enseñanza:

(comprende una descripción de la metodología de enseñanza y de las horas dedicadas por el estudiante a la asignatura, distribuidas en horas presenciales -de clase práctica, teórico, laboratorio, consulta, etc.- y no presenciales de trabajo personal del estudiante)

- Horas clase (teórico): 24
- Horas clase (práctico): 18
- Horas clase (laboratorio): 6
- Horas consulta: 1
- Horas evaluación: 0

- o Subtotal horas presenciales: 25
- Horas estudio: 30
- Horas resolución ejercicios/prácticos: 15
- Horas proyecto final/monografía: 15
 - o Total de horas de dedicación del estudiante: 75

Forma de evaluación: Proyecto final.

Temario:

1. **Árboles de clasificación y regresión.** Introducción. Árboles y splits. Construcción: árbol maximal y podado. Un resultado teórico desde el punto de vista de selección de modelos. Extensiones y variantes. Árboles en práctica usando el paquete de R rpart.
2. **Medidas de influencia para árboles y otros tópicos.** Medidas de influencia para árboles. Medidas de influencia y estabilidad para Graphical Models..
3. **Bosques Aleatorios, Importancia y selección de variables.** Introducción. Árboles, Bagging y Random Forests (RF). Extensiones y variantes, algunos resultados teóricos. Out-of-bag error y medida de importancia de variables. Selección de variables usando RF. RF en práctica usando los paquetes para R randomForest y VSURF.
4. **Una aplicación: RF-Based Approach for Physiological Functional.**
5. **Una aplicación a datos de polución**
6. **Bosques aleatorios para grandes volúmenes de datos**

(Temario original en inglés)

1. **CART trees.** Introduction. CART trees and splits. Construction: maximal tree and pruning. A typical theoretical result from the model selection viewpoint. Extensions and variants. CART in practice using the R package rpart
2. **Influence measures for CART and related topics.** Influence measures for CART. Influence Measures and Stability for Graphical Models
3. **Random Forests, Variable importance and Variable Selection.** Introduction. Trees, Bagging and Random Forests (RF). Extensions and variants, some theoretical results. Out-of-bag error and variable importance measure. Variable Selection using RF. RF in practice using the R packages randomForest and VSURF
4. **An application: RF-Based Approach for Physiological Functional.** Variable Selection: Towards Driver's Stress Level Classification. Introduction and motivation. Physiological functional variables and wavelets. Block variables importance measure. Functional variable selection using RF. Driver's Stress Level Classification
5. **An application to pollution data**
6. **Random Forests for Big Data**

Bibliografía:

(título del libro-nombre del autor-editorial-ISBN-fecha de edición)

- Breiman, L., Friedman, J., Olshen, R. et Stone, C. [1984]. Classification and Regression Trees . Chapman & Hall, New York.



Facultad de Ingeniería Comisión Académica de Posgrado

- Bar-Hen, A., Gey, S. et Poggi, J.-M. [2015]. Influence measures for CART classification trees. *Journal of Classification*, 32 (1), 21–45.
- Therneau, T., Atkinson, B. et Ripley, B. [2015]. rpart : Recursive Partitioning and Regression Trees
- Bar-Hen, A., Gey, S. et Poggi, J.-M. [2015]. Influence measures for CART classification trees. *Journal of Classification*, 32 (1), 21–45.
- A. Bar-Hen, J.-M. Poggi [2016] Influence Measures and Stability for Graphical Models. *Journal of Multivariate Analysis*, Vol. 47, 145-154
- Breiman, L. [2001]. Random forests. *Machine learning* , 45 (1), 5–32.
- Genuer, R., Poggi, J.-M. et Tuleau-Malot, C. [2010]. Variable selection using random forests. *Pattern Recognition Letters* , 31 (14), 2225–2236.
- Genuer, R., Poggi, J.-M. et Tuleau-Malot, C. [2015]. Vsurf : An R package for variable selection using random forests. *The R Journal* , 7 (2), 19–33.
- Gregorutti, B., Michel, B. et Saint-Pierre, P. [2015]. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90, 15–35.
- El Haouij N., Poggi J.-M., Ghozi R., Sevestre Ghalila S., Jaïdane M [2016], Random Forest-Based Approach for Physiological Functional Variable Selection: Towards Driver's Stress Level Classification, HAL, hal.archives-ouvertes.fr/hal-01426752v1
- M. Bobbia, F.-X. Jollois, J.-M. Poggi, B. Portier [2011] Quantifying local and background contributions to PM10 concentrations in Haute-Normandie, using random forests, *Environmetrics*, 22, 758-768
- Genuer, R., Poggi, J.-M., Tuleau-Malot, C. et Villa-Vialaneix, N. [2017]. Random Forests for Big Data. ArXiv e-prints, HAL, hal.archives-ouvertes.fr/hal-01233923v1, Submitted to Big Data Research



5/ CUATRO

**Facultad de Ingeniería
Comisión Académica de Posgrado**

Datos del curso

Fecha de inicio y finalización: Segunda mitad de febrero

Horario y Salón: A definir
