1213/10

# Formulario de Aprobación Curso de Posgrado 2011.

## Asignatura: Aplicaciones de Data Mining

**Profesor de la asignatura** [1]**:** Ing, Magister, Gustavo Denicolay, Universidad de Buenos Aires

**Profesor Responsable Local** [1]**:** Prof. Alejandro Vaisman, Gr. 4, INCO, UdelaR

**Instituto ó Unidad:** Computación
**Departamento ó Area:**

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

**Fecha de inicio y finalización: A definir 1er Semestre de 2011.**
**Horario y Salón**: a confirmar

**Horas Presenciales: 44**

**Créditos**: **10**

**Público objetivo y Cupos:** Profesionales y académicos universitarios, principalmente de las áreas de Informática, ciencias económicas y ciencias biológicas.

**Objetivos:** Introducir las principales técnicas de data mining, y aplicarlas a la resolución de casos concretos de análisis de datos y predicción de comportamiento. Capacitar en forma práctica y teórica a los alumnos para que puedan detectar y resolver problemas de data mining y construir modelos aplicando una adecuada combinación de técnicas, incluyendo sociación, clasificación, clustering, y regresión.

**Conocimientos previos exigidos:** Ninguno.

**Conocimientos previos recomendados:** Nociones de probabilidad y estadística, bases de datos relacionales y algoritmos.

**Metodología de enseñanza:**

Se dictará el curso en dos partes: en la primer semana se impartirán diversos conceptos, y se introducirán técnicas de pre-proceamiento de datos. Se plantará un problema real, y se definirá un trabajo a realizar en las semanas subsiguientes. Este trabajo consistirá en la preparación de un set de datos para realizar un análisis poterior. Luego se realizarán dos clases de consulta, y a continuación se dictará clases durante una semana, nuevamente, introduciendo técnicas de predicción. Finalmente se planteará un problema sobre el set de datos preparado anteriormente, referido a la aplicación de las técnicas explicadas en la clase. El curso se complementa con horas de estudio asistido para poder llevar a cabo exitosamente el proceso de enseñanza-aprendizaje. Se estima 1.5 hs de estudio por cada hora presencial **(66hs).** Adicionalmente se consideran 10hs presenciales de consultas con el profesor local, tareas de recopilación de datos, etc.).

**Forma de evaluación:**
Trabajo domiciliario en dos partes: una a entregar luego de la primera semana de clases, referido a preproceso de dato. La segunda, a entregar luego de la finalización del dictado de clases. El trabajo consitirá en aplicar las técnicas vistas en clase, **a casos reales con datos reales,** y evaluar los resultados obtenidos, contrastándolos contra la realidad. **Total estimado de horas de trabajo: 30**

**Temario:**

- El proceso de KDD (3hs).

- Análisis de Asociaciones y el problema de la canasta del mercado. Algoritmos (4hs)
- Extensiones a las Reglas de Asociación (2hs)
  - Atributos categóricos
  - Atributos numéricos
  - Jerarquías
- Patrones Secuenciales (2hs)
- Técnicas de Preprocesamiento de Datos. Su importancia en el proceo de KDD (4hs)
- Aprendizaje no Supervisado. Introducción (3hs)
- Algoritmos de Clustering: k- means, Jerárquico.(4h)
- Modelos Predictivos y Aprendizaje Supervisado.(4hs)
- Algoritmos de Clasificación. Arboles de Decisión (4hs)
- Regresión lineal (4hs)
- Aplicación de Técnnicas a problemas reales: segmentación, predicción de comportamiento de clientes, etc. Casos de estudio. (10hs).

**Bibliografía:**

***Papers***

Agrawal R.., Imielinsky, T., Swami, A. Mining Association Rules between Sets of Items in Large Databases, SIGMOD 1993, 207-216.

Agrawal R., Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases. , VLDB 1994, pp. 485-499

Agrawal R., Srikant, R. . Mining Sequential Patterns, ICDE 1995, pp. 3-14.

Brin, S., Motwani , R., Silverstein, C. Beyond Market Baskets: Generalizing Association Rules to Correlations

Fawcett, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Technical Report HPL-2003-4, HP Labs, 2003

Garofalakis, M, Rastogi, R., Shim, K. Mining Sequential Patterns with regular expressions constrains. IEEE/TDKE, Vol. 14 n. 3., pp. 530-552.

Flach, P. Putting Things in Order, On the fundamental role of ranking in classification and probability estimation. , 18th European Conference on Machine Learning, 2007.

Japkowicz, N. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. AAAI Workshop, Technical Report WS-00-0

Oates, T. and Jensen, D. Large Datasets Lead to Overly Complex Models: an Explanation and a Solution. pp. 294. Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp 294 - 298, 1998

Provost, F. Domingos, P. Tree Induction for Probability-based Ranking. Machine Learning, 52,3, September 2003, pp 199-215, 2003

Salzberg, S. On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, 1, 317-327 (1997)

Srikant R, Agrawal, R. . Mining Generalized Association Rules. VLDB 1995, pp. 407-419

Srikant, R. Agrawal R. . Mining Sequential Patterns : generalization and performance improvements EDBT 1996, pp. 3-17

*Libros*

Tan Pang-Ning, Kumar Vipin , Steinbach Michael. Introduction to Data Mining. Addison-Wesley, ISBN 0321321367, Mayo 2005

Adamo, J.M. Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms. Springer, ISBN 0387950486, Dic 2000

Pyle, D. Data Preparation for Data Mining. Morgan Kaufmann Publishers, 1999