FACULTAD DE CIENCIAS AGRARIAS Y FORESTALES
PROSECRETARÍA DE POSGRADO

Curso de Posgrado

Análisis de Grandes Bases de Datos

Carácter:

Curso de Posgrado acreditable a Carreras de Posgrado

Docente responsable

PhD. Pablo Daniel Reeb
Profesor Adjunto
Cátedra Bioestadística y Modelos Multivariados y Diseño de Experimentos
Facultad de Ciencias Agrarias – UNCo

Docentes Intervinientes

PhD. Pablo Daniel Reeb
Dictado clases teórico-prácticas
Carga Horaria: 16 horas
Formato: Taller
Temas: Métodos basados en árboles – Support Vector Machines – Aprendizaje No Supervisado

Mg. Guillermo Sabino
Dictado clases teórico-prácticas
Carga Horaria: 16 horas
Formato: Taller
Temas: Statistical Learning – Métodos de Remuestreo – Selección de Modelos Lineales

Fundamentación de la Propuesta

La tecnología para la adquisición, almacenamiento y administración de datos ha motivado y posibilitado la creación de grandes bases de datos, las cuales crecen en dos direcciones: mayor cantidad de observaciones y mayor cantidad de variables o atributos. Estos crecientes volúmenes de datos, que se encuentran en muy variadas y diversas actividades humanas, tienen asociados el problema de almacenamiento y gestión de bases de datos y la obtención de información-conocimiento a partir de ellos (Romero, 2009).

El tratamiento estadístico de las grandes bases de datos engloba un conjunto de técnicas que constituyen, junto a otras, la estrategia de análisis conocida como KDD: Descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases (Fayyad et al., 1996)), cuyo objetivo es obtener información de alto nivel a partir de una gran base de datos de bajo nivel.

Statistical Learning se refiere a un conjunto de herramientas para el modelado y la comprensión de datos complejos. Es una zona de reciente desarrollo en estadística y se combina paralelamente con la informática y, en particular, con Machine Learning, el cual está constituido por algoritmos que tienen la particularidad de detectar patrones en distintos contextos de bases de datos para luego reproducirlos, es decir, algoritmos con la capacidad de “aprender”. Este campo abarca muchos métodos tales como Regresiones Lasso, árboles de clasificación y regresión, Boosting y Support Vector Machine, entre otros.

Con el auge de los problemas de "Big Data", Statistical Learning se ha convertido en un campo muy innovador en muchas áreas científicas, como así también en marketing, finanzas, y otras disciplinas de negocios. Actualmente existe una gran demanda de personas con habilidades en Statistical Learning.

Hoy en día, el uso de herramientas para extracción de información de grandes bases de datos tiene amplia difusión en casi todas las disciplinas científicas. Para el abordaje de los contenidos de este curso se abordará el libro “An Introduction to Statistical Learning: With Applications in R”.

Objetivos

General: Lograr una visión global de las implicancias de trabajar con grandes bases de datos y conocer los problemas asociados con la misma desde la óptica del Statistical Learning.

Específicos:

1) Conocer el amplio abanico de técnicas disponibles para el análisis de grandes bases de datos.

2) Aprender las bondades e implicancias de las técnicas de remuestreo.

3) Adquirir destrezas en la implementación de Cross-Validation.

4) Aplicar las múltiples técnicas estadísticas estudiadas a datos reales que permitan comparar las ventajas y desventajas de su implementación.

Contenidos (Programa Analítico + Bibliografía)

Statistical Learning. ¿Qué significa Statistical Learning? ¿Por qué estimamos una función? ¿Cómo estimamos una función? Una solución intermedia entre la precisión de la predicción y la capacidad de interpretación del modelo. Aprendizaje Supervisado vs. No Supervisado. Regresión vs. Clasificación. Evaluación de la precisión de un modelo. Medidas de calidad del ajuste. Equilibrio entre sesgo y variabilidad. Casos de respuesta cualitativa. Ejercicios.

Métodos de Remuestreo. Validación Cruzada. Validación a partir de una aproximación. Validación Cruzada Leave-One-Out. Validación Cruzada k-Fold. Equilibrio entre el sesgo y la varianza en Validación Cruzada k-Fold. Validación Cruzada en Problemas de Clasificación. Bootstrap. Ejercicios.

Selección de Modelos Lineales**.** Selección de subconjuntos. Mejor selección. Selección paso a paso. Elección del Modelo óptimo. Shrinkage Methods. Ridge Regression. Regresiones Lasso. Selección del parámetro Tuning. Métodos de reducción de dimensión. Principal Components Regression (PCR). Partial Least Squares (PLS). Consideraciones en Alta Dimensionalidad. ¿Cuál es el problema con la Alta Dimensionalidad? Regresión en Alta Dimensionalidad. Interpretando resultados en Alta Dimensionalidad. Ejercicios.

Métodos basados en árboles. Introducción a los árboles de decisión. Árboles de decisión. Árboles de Clasificación. Árboles vs. Modelos Lineales. Ventajas y desventajas de los Árboles. Bagging, Random Forests, Boosting. Ejercicios.

Support Vector Machines. Clasificador de margen máximo. ¿Qué es un Hiperplano? Clasificación utilizando la separación de Hiperplanos. Construcción de los márgenes máximos. Análisis de casos no separables. Introducción a Support Vector Classifiers. Detalles y características. Support Vector Machines (SVM). Clasificación con límites de decisión no lineales. Aplicaciones. SVMs con más de dos clases. Clasificación Uno contra Uno. Clasificación Uno contra todos. Relación con Regresión Logística. Ejercicios.

***Aprendizaje No Supervisado*.** Los desafíos del Aprendizaje No Supervisado. Análisis de Componentes Principales. Análisis de Cluster. K-Means Clustering. Cluster Jerárquicos. Detalles de aplicación en el uso de Cluster. Ejercicios.

BIBLIOGRAFIA DE REFERENCIA

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition. Springer-Verlag.

Metodología:

Clases durante cuatro días con una carga horaria presencial de 32 horas distribuida diariamente de 9 a 13 hs. y de 14 a 18 hs. Este curso es de índole teórico-práctica. Incluye clases teóricas y laboratorios de análisis de datos usando el software libre R. La carga horaria total se completa con 13 horas de trabajo no presencial.

Laboratorio de computación

Una porción sustancial del tiempo del curso se dedica a los laboratorios de análisis de datos. Cada estudiante deberá disponer de una computadora durante el horario de clase. Es esencial que las computadoras cuenten con el siguiente software instalado en su última versión: R (http://cran.r-project.org/) y RStudio (http://www.rstudio.com/)

Duración total: 45 horas.

Cronograma

LUNES	Mañana	Statistical Learning – Métodos de Remuestreo
	Tarde	Laboratorio R
MARTES	Mañana	Selección de Modelos Lineales
	Tarde	Laboratorio R
MIÉRCOLES	Mañana	Métodos basados en árboles – Support Vector Machines.
	Tarde	Laboratorio R
JUEVES	Mañana	Aprendizaje No Supervisado
	Tarde	Laboratorio R

	Clases teóricas
	Clases Prácticas

9- <u>Evaluación</u>: (Explicitar condiciones para la aprobación del curso)

Para obtener la acreditación del curso se requerirá:

a) Haber asistido a por lo menos el 80% de las clases programadas (Se otorga certificado de **asistencia**)

b) La acreditación del curso se hará a través de un informe escrito grupal, de hasta tres asistentes (Se otorga certificado de **aprobación**).

10 - <u>Cupo de alumnos para el dictado</u>: Mínimo 10 alumnos y máximo 20 alumnos

- Destinatarios:

Todos aquellos investigadores que estén interesados en el uso de métodos estadísticos modernos para el modelado y predicción de datos. Este grupo incluye científicos, analistas de datos, pero también para aquellas personas que estén vinculados con campos no cuantitativos como ciencias sociales. Se espera que los participantes tengan acreditado como mínimo un curso de estadística.