Minería de Datos y **Machine Learning**

Módulo II - Aprendizaje Supervisado: Regresión

software

descripción

En los últimos años, los rápidos avances en las tecnologías de recopilación y almacenamiento de datos han dado lugar a conjuntos de datos que son "grandes" en muchos sentidos de la palabra. **La Minería de Datos** está relacionada con el análisis de estos grandes conjuntos de datos con la finalidad de proporcionar ideas, patrones, modelos descriptivos y predictivos que permitan extraer y generar conocimiento para las organizaciones. Algunas de las tareas más comunes son la clasificación, agrupamiento, descubrimiento de reglas de asociación y patrones de respuestas, detección de anomalías, etc.

Para poder realizar estas tareas, la Minería de Datos combina los aspectos teóricos fundamentales de áreas tan diversas como la estadística, las matemáticas, la ingeniería y las ciencias de la información con muchas aplicaciones prácticas y relevantes de la vida real. Dentro de esas disciplinas destaca el aprendizaje automatizado o **machine learning**, que está teniendo profundos efectos en muchas industrias diferentes, desde servicios financieros hasta ventas minoristas y publicidad; convirtiéndose rápidamente en una herramienta fundamental para tomar mejores decisiones en los negocios: decisiones basadas en datos, no en instintos o conjeturas.

Este módulo del curso tiene como objetivo presentar los fundamentos del aprendizaje estadístico, discutir estrategias para evaluar la eficiencia predictiva de los modelos supervisados y presentar los principales modelos de regresión.

publico objetivo

Analistas de datos. Profesionales relacionados a la inteligencia de negocios, investigación de mercados e interesados en el área de Ciencia de Datos. Académicos e investigadores. . Público en general que requiera extraer conocimiento desde diferentes fuentes de información.

logros de aprendizaje

Al finalizar este curso, el participante conocerá los fundamentos del aprendizaje supervisado y aplicará modelos de regresión para inferir y predecir. De manera específica el participante estará en capacidad de:

- Conocer las diferentes técnicas y modelos del aprendizaje estadístico y automatizado, que forman parte del proceso de aprendizaje supervisado y que son esenciales para la minería de datos.
- Aplicar computacionalmente los modelos aprendidos en diferentes campos como la industria, el comercio, la banca, los seguros, biología, etc.
- Comprender el desarrollo de los algoritmos de las principales técnicas de regresión.
- Presentar de manera efectiva los resultados obtenidos.

contacto

+51970227123

info@perustat.com

http://www.perustat.com

fb://perustat

PeruStat
Analytics training, consulting & researching

contenidos

Tema 1 **Fundamentos del Aprendizaje** Supervisado (8 horas)

- Aprendizaje estadístico
  - Ideas generales
  - Aprendizaje estadístico vs. Machine Learning.
  - Aprendizaje estadístico vs. Ciencia de Datos.
  - Principales problemas.
- Predicción e Inferencia
- Modelos paramétricos y no paramétricos.
- Precisión vs. interpretación: Métodos inflexibles y métodos flexibles.
- Compromiso sesgo-varianza.
- Revisión de conceptos y notación básica
  - Vectores aleatorios.
  - Momentos de una distribución.
  - Valor esperado.
  - Matriz de varianza-covarianza y correlaciones.
  - Combinaciones lineales.
  - Distribución normal multivariada.

Tema 2 **Regresión Lineal** Múltiple (8 horas)

- Definición y notación.
- Modelo lineal clásico.
- Estimación: Mínimos Cuadrados Ordinarios y Máxima Verosimilitud.
- Inferencia.
- Evaluación y selección de modelos.
- Evaluación de supuestos y análisis de residuos.
- Predicción.
- Predictores cualitativos.

Tema 3 **Evaluación de la performance** predictiva (8 horas)

- Error de entrenamiento vs. error de prueba.
- Funciones de pérdida.
- Particionamiento del conjunto de datos.
- Técnicas de evaluación por remuestreo
  - LOOCV (Leave Oune Out cross-validation)
  - KCV (k-fold cross-validation)
  - Bootstrap.
- Hiperparámetros.
- Selección de modelos.

Tema 4 **Selección de** Variables (4 horas)

- Selección de Subconjuntos
  - Selección de los mejores subconjuntos
  - Selección Stepwise.
  - Elección del modelo óptimo.
- Métodos de regularización
  - Regresión Ridge
  - Regresión Lasso

Tema 5 **Extensiones al modelo** lineal (4 horas)

- Regresión Polinomial.
- Regresión usando Splines.
- Regresión local.
- GAMs (Modelos Aditivos Lineales)

contacto


+51970227123


info@perustat.com


http://www.perustat.com


fb://perustat

PeruStat
Analytics training, consulting & researching

metodología

La metodología del curso se basa en una combinación de clases teóricas y análisis de casos prácticos en la computadora, con la finalidad de que el participante comprenda la metodología, la motivación, los supuestos, las fortalezas y las debilidades de los métodos tratados en el curso. Cada sección del curso está motivada por un conjunto de datos en particular, de tal forma que el participante gane experiencia trabajando con una amplia variedad de fuentes de datos similares a los que usa en la realidad. Los contenidos están estructurados en 8 sesiones con un total de 32 horas académicas (24 horas cronológicas).

evaluación y asistencia

- **Certificado de asistencia**: Para obtener este certificado debe de mantener un porcentaje mínimo de 75% de asistencia a las clases.
- **Certificado de aprobación**: El 100% de la calificación final se obtiene sobre la base de cuestionarios o listas de ejercicios al final de cada sesión. Para recibir el certificado de aprobación el participantes deben obtener al menos un 70% de los puntos posibles y contar con el porcentaje mínimo de asistencia a clases.

Serán otorgados certificados a nombre de PeruStat Analytics S.A.C. que acredita a los participantes del curso. La certificación que se otorga es excluyente.

materiales

Material preparado por el equipo de capacitación con los contenidos del curso el cuál será entregado a los participantes en medios físicos o digitales.

referencias

Bishop, C. (2011). Pattern Recognition and Machine Learning. Springer.

Clarke, B., Fokoue, E. y Zhang, H. (2009). Principles and Theory for Data Mining and Machine Learning. Springer Verlag.

Gareth, J., Witten, D., Hastie, T. y Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

Hastie, T., Tibshirani, R. y Friedman. J. (2016). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Kuhn, M. y Johnson, K. (2013). Applied Predictive Modeling. Springer Verlag.

Lantz, B. (2015). Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems Packt Publishing; 2 edition .

Larose, D.T. (2006). Data Mining Methods and Models. Wiley Interscience.

Ledolter, J. (2013). Data Mining and Business Analytics with R. John Wiley & Sons.

Nisbet, R., Elder IV, J. y Miner, G. (2009). Handbook of Statistical Analysis and Data Mining Applications. Academic Press.

Putler, D. y Krider, R. E. (2012). Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R. Chapman and Hall/CRC.

Shmuelil, G., Bruce, P., Yahav, I., Patel, N. y Lichtendahl, K. (2017). Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. John Wiley & Sons.

Véliz, C. (2018). Aprendizaje Automático. Análisis para la Minería de Datos y Big Data. Departamento Académico de Ciencias. Pontificia Universidad Católica del Perú.

Wu J. y Coggeshall, S. (2012). Foundations of Predictive Anaytics. Chapman and Hall/CRC.

contacto

+51970227123

info@perustat.com

http://www.perustat.com

fb://perustat

PeruStat
Analytics training, consulting & researching