

Aprendizaje de máquina en bioinformática

Grupo de Bioinformática y Biocomputación

Facultad de Ingeniería

Universidad del Valle

| | | | |
|---------------------|----------------------|-----------------|----|
| Código: | 750XXX | Créditos: | 3 |
| Tipo de asignatura: | Electiva Profesional | Habitable: | No |
| Prerrequisito: | Ninguno | Validable: | Si |
| Tipo de programa: | Pregrado | Horas / Semana: | 3 |

JUSTIFICACIÓN

Aprendizaje de máquina, una rama de la inteligencia artificial, trata de las técnicas que permiten que los computadores aprendan. Está basado en datos y busca optimizar un criterio de desempeño, para lograr describir la información contenida en ellos. Tiene importancia especial en bioinformática, área que puede ser considerada como la aplicación de tecnologías de información a la biología molecular. La cantidad de datos biológicos almacenados en bases de datos crece exponencialmente, y la mayoría de ellos está públicamente disponible. En varios campos biológicos como genómica, proteómica y biología de sistemas, se aplican con éxito técnicas de aprendizaje de máquina. Desde el enfoque informático, las técnicas de aprendizaje se pueden clasificar en grupos grandes, entre ellos los que son objetivos de esta asignación: Clasificación y agrupamiento, selección de características, análisis de secuencias biológicas e inferencia filogenética.

La asignación propuesta tiene un enfoque teórico—práctico. En la parte teórica se introducen los conceptos de ambos campos involucrados, se explican las principales técnicas de aprendizaje de máquina y se señalan los problemas bioinformáticos, donde se pueden aplicar exitosamente estas técnicas. La parte práctica consiste en la solución de problemas ejemplares, basado en un software de aprendizaje de máquina (por ejemplo BioMall), donde el estudiante, basándose en datos biológicos, no sólo aplica las técnicas, sino que está motivado a sacar conclusiones de tipo informático y biológico.

OBJETIVOS

Abordar aspectos teóricos y prácticos de las principales áreas temáticas de aprendizaje de máquina, aplicadas a problemas de bioinformática.

Al final del curso, el estudiante está en capacidad de:

- Entender el contexto biológico en que se aplican los modelos y algoritmos de aprendizaje de máquinas.
- Conocer herramientas de aprendizaje de máquinas, con módulos bio. (*RapidMiner, Weka*)
- Modelar el problema biológico de tal forma que se pueda obtener el mejor rendimiento.

CONTENIDO

Introducción a aprendizaje de máquina y bioinformática.

- Proteómica: Predicción de estructuras de proteína

Introducción a herramientas de aplicación del aprendizaje de máquina.

- Weka
- Rapid Miner

Introducción a importantes áreas temáticas de aprendizaje de máquina en bioinformática.

Aplicaciones bioinformáticas de los métodos de aprendizaje de máquina.

- Genómica: Predicción de Genes

- Clasificación y agrupamiento
- Selección de características
- Análisis de secuencias biológicas
- Inferencia filogenética

METODOLOGÍA

La asignatura requiere una participación activa de los estudiantes y de trabajo en grupo. Se harán talleres prácticos sobre una herramienta de aprendizaje de máquina, seleccionada para problemas específicos en las áreas de genómica y proteómica. Los estudiantes leerán al menos un artículo por temas para discutirlos en la sesión semimagistral del profesor.

EVALUACIÓN

- 30 % Pruebas Individuales
- 40 % Proyecto de curso
- 30 % Controles de lecturas (Ensayos, pruebas cortas)

BIBLIOGRAFÍA

- Aik Choon Tan, David Gilbert . Machine Learning and its Application to Bioinformatics: An Overview. August 31, 2001
- Ian H. Witten, Eibe Frank . Data Mining, Practical Machine Learning Tools and Techniques, Second Edition , Elsevier Academic Press. 2005.
- Krane, Dan E. & Raymer, Michael L. 2002. Fundamental Concepts of Bioinformatics. Ed. Benjamin Cummings.
- Baldi P. & Brunak S.; 2001. Bioinformatics: the machine learning approach. Second edition, MIT Press, Cambridge, MA, 2001.

- Jones, Neil C. & Pevzner, Pavel A.; 2004 An Introduction to Bioinformatics Algorithms (Computational Molecular Biology)
- Durbin, Richard; Eddy, Sean R.; Krogh Anders; Mitchison, Graeme; 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press. Cambridge, New York.