FICHA TEMATICA
CURSO DE POSGRADO

Fecha: 29-03-2019

Título del Curso de Posgrado: Fundamentos de Machine Learning

Docente: Ariel José Berenstein.
E- mail:arieljberenstein@gmail.com
Tel: 11 6462 4658

Docente coordinador: Nicolás Palopoli
E-mail: nicopalo@gmail.com
Tel: 11 6747 4682

Destinatarios:
Graduados y estudiantes avanzados de Bioinformática, Biotecnología, Informática, Programación, Automatización, Ingenierías y carreras afines.

Carga horaria: 30 hs

Lugar de Realización: UNQ

Metodología:

Teórico:	■
Práctico:	☐
Teórico-práctico:	☐

Modalidad:

Virtual:	☐
Presencial:	■

Evaluación Final: Examen escrito.

Matrícula deseada: Se espera un mínimo de 10 y un máximo 30 alumnos.

ANEXO 1
CONTENIDOS Y BIBLIOGRAFÍA

Título del Curso:

Fundamentos de Machine Learning

Fundamentación:

Las áreas de biología computacional, bioinformática y disciplinas afines se encuentran en constante crecimiento y demanda de profesionales capaces de entender, manipular y modelar los grandes volúmenes de datos que se generan y hacen públicos a diario. El estudio de conjuntos de datos grandes y complejos requiere de herramientas computacionales aptas para integrar y cuestionar esos datos en forma inteligente, capaces de trascender su descripción para proporcionar modelos predictivos que transformen los datos biológicos emergentes en nuevo conocimiento sobre la biología. En este sentido, el presente curso de posgrado busca sentar bases fundacionales y sólidas de estadística inferencial y aprendizaje automático (machine learning) que permitan incorporar dichas técnicas al análisis de datos biológicos. El mismo está dirigido tanto a graduados como a estudiantes avanzados de grado. Al finalizar el curso, los participantes podrán comprender las técnicas estadísticas de evaluación de modelos de aprendizaje automático y reconocer los algoritmos básicos de clasificación, para implementarlos posteriormente en su lenguaje de preferencia y área de interés.

Objetivo general:

Ofrecer bases sólidas para la comprensión, implementación y comparación de modelos de machine learning, motivando su uso en el área de biología computacional y bioinformática en particular.

Objetivos específicos:

- Que los alumnos incorporen las bases de estadística inferencial necesarias para evaluar y comparar modelos de aprendizaje automático.
- Que entiendan la interrelación entre los conceptos de sobreajuste, generalización, regularización, varianza y sesgo de los modelos de aprendizaje automático.
- Que se familiaricen con las técnicas computacionales adecuadas para entrenar y validar modelos.
- Que puedan reconocer y juzgar la aplicación de métodos de aprendizaje automático para una mejor comprensión de la literatura científica actual.
- Que incorporen herramientas fundacionales para abordar su desarrollo en el área de aprendizaje automático, con la confianza necesaria para su aplicación a las distintas áreas temáticas de interés.

Unidades/módulos:

1) Introducción general.
Aprendizaje supervisado vs no supervisado. Aprendizaje por refuerzos. Motivación y aplicaciones en bioinformática.

2)Tópicos de estadística inferencial.
Bases de estadistica inferencial. Estadística, por qué ? Distribución de probab, distrib acumulada. Ejemplos. Test de hipótesis. Error de tipo I y error de tipo II. Potencia.
Test paramétricos vs no paramétricos.
Test de interés: Shapiro-Wilk test, Fisher Exact Test, Mann-Witney-U test, Kolmogorov-Smirnov Test.
Corrección por testeo múltiple.
Bonferroni. FDR.
Ejemplos de aplicación en bioinformática. Expresión diferencial. Gene Ontology.

3) Aprendizaje supervisado. Conceptos Generales
Modelos de aprendizaje supervisado: regresión y clasificación.
Clasificación binaria vs clasificación de múltiple categorías. Umbral de decisión en modelos de clasificación. Predictor ideal. Predictor aleatorio.
Generalización. Overfitting. Bias-Variance trade-off. Regularización.
Parámetros del modelo, ajuste vía máxima verosimilitud.
Hiperparámetros. Métodos de remuestreo. Validación cruzada.
Setup de validación. Estimación del error.
Desbalance de clases.
Mejores métodos vs mejores variables.

4) Modelos Lineales generalizados:
Regresión logística univariada. Formulación del modelo univariado y multivariado. Ajuste de parámetros. Ejemplos.
Regularización L1 y L2.
Hiperparámetro de regularización. Selección vía validación cruzada.
Regresión logística para vector de respuesta multiclase.

5) Modelos no lineales. Parte 1
Árboles de decisión. Idea intuitiva. Estratificación del espacio de variables.
Criterio de corte de ramas del árbol. Pureza: Gini Impurity. Cross entropy. Greedy approach para construcción del modelo. Overfitting.
Pruning del árbol para mejorar su poder de generalización.
Hiperparámetros. Seteo vía validación cruzada.
Árboles vs Modelos lineales.
Pros-Cons de los modelos de árbol.

6) Evaluación y comparación de modelos.
Evaluación de un clasificador. Medidas de desempeño. Dependencia con el umbral de decisión. Medidas independientes de umbral de decisión. Curvas ROC y PR. Aŕea bajo la curvas AUC. Distribución estadística de AUC. Intervalo de confianza. Varianza en función del tamaño muestral. Comparación estadística de clasificadores. DeLong Test.

7) Modelos no lineales. Parte 2: Modelos de ensamble.
Bagging. Idea intuitiva. Bagging trees. Formulación del modelo. Clasificación por mayoría de votos. Hiperparámetros.
Estimación del error vía out of bag.
Importancia de variables. Efecto de co-linealidades en las variables.
Random Forest. Idea y diferenciación respecto a bagging. Hiperparámetros. Extensos Casos de aplicación.
Boosting. Idea intuitiva. Naturaleza secuencial. Boosting trees. Hiperparámetros. XGBoost. Cualidades, ventajas-desventajas. Hiperparámetros.

Bibliografía:

Obligatoria:
1)Statistical Modeling and Machine Learning for Molecular Biology. Alan M. Moses University of Toronto, Canada. CHAPMAN & HALL/CRC Mathematical and Computational Biology Series. 2017.
2) An Introduction to Statistical Learning with Applications in R. Gareth James Daniela Witten Trevor Hastie Robert Tibshirani. Springer. 2015

Optativa:
3) The Elements of Statistical Learning Data Mining, Inference, and Prediction. Trevor Hastie Robert Tibshirani Jerome Friedman. Second Edition. Springer
4) Bioinformatics Algorithms: An Active Learning Approach 2nd Edition, Vol. I Phillip Compeau & Pavel Pevzner. Active Learning Publishers 2015.
5) Bioinformatics Algorithms: An Active Learning Approach 2nd Edition, Vol. II Phillip Compeau & Pavel Pevzner. Active Learning Publishers 2015.

Modo de evaluación:

Examen escrito.

Cronograma de clases:

Dos encuentros por semana (días no consecutivos), durante un total de 2 semanas.
Día 1, Mañana (9-13hs). Unidades 1 y 2. Docentes: A.Berenstein, N.Palopoli.
Día 2, Mañana (9-13hs). Unidades 3 y 4. Docentes: A.Berenstein, N.Palopoli.
Día 3, Mañana (9-13hs). Unidades 5 y 6. Docentes: A. Berenstein, N.Palopoli.
Día 4, Mañana (9-13hs). Unidad 7 + Revisión. Docentes: A.Berenstein, N.Palopoli.