PROGRAMA de Ciencia de Datos

Carrera/s: Licenciatura en Informática

Asignatura: Ciencia de Datos

Núcleo al que pertenece: Avanzado

Profesor: Pablo Factorovich

Asignaturas Correlativas: Algoritmos, Matemática III y Probabilidad y Estadística

Objetivos:

- Que les alumnes conozcan los problemas típicos de la Ciencia Datos y cómo esta se relaciona con el Aprendizaje Automático y la Estadística.
- Que les alumnes se familiaricen con la noción de modelo en el contexto de Ciencia de Datos.
- Que les alumnes comprendan algoritmos utilizados para el abordaje de problemas típicos de la Ciencia de Datos como la regresión, clasificación, agrupamiento o reducción de dimensionalidad. Entre estos cabe mencionar la gradiente descendente típico o para muchos datos, resolución de sistemas de ecuaciones normales, regresión polinómica, redes neuronales, árboles de decisión, k promedios y máquinas de vector de soporte, etc.
- Que les alumnes pueden aplicar los algoritmos recién mencionados, identificando sus limitaciones y aplicando técnicas clásicas para superarlas. Que les alumnes puedan evaluar sus modelos y a partir de este análisis encontrar la forma de mejorarlos.

Contenidos mínimos:

- Noción de modelo, del problema de aproximación y de error en la aproximación de un modelo.
- Algoritmos para regresión lineal y polinómica: gradiente descendente, ecuaciones normales y redes neuronales. Adaptación para regresión polinómica
- Algoritmos para clasificación en dos o más clases: regresión logística, redes neuronales, árboles de decisión, máquinas de vector soporte.
- Problema de agrupamiento y k promedios
- Problema de reducción de dimensionalidad y algoritmos para resolverlo.

- Trabajo con grandes volúmenes de datos: minimización para esas circunstancias y map reduce.
- Aplicación práctica de Ciencia de Datos

Carga horaria semanal: 4 hs

Programa analítico:

Unidad 1: Noción de Ciencia de Datos y problemas abordados.

Definiciones de Ciencia de Datos; puntos en común y diferencias con la Estadística y el Aprendizaje Automático. Modelos en el ámbito de Ciencia de Datos. Problemas abordados en el área como: regresión, clasificación, agrupamiento o reducción de dimensionalidad. Clasificación de problemas de aprendizaje en supervisados y no supervisados. Problemas de aproximación: minimización de una función de costos (1 semana)

Unidad 2: Problema de regresión lineal, algoritmos para resolverla y usos en problemas polinomiales.

Modelo unidimensional y supuestos. Algoritmo de gradiente descendente unidimensional y su convergencia. Repaso de álgebra lineal básica: vectores, matrices, producto escalar, vectorial, sistemas lineales, inversa de matrices. Modelo multidimensional y supuestos. Gradiente descendente multidimensional y su convergencia. Mínimos cuadrados por ecuaciones normales. Regresión polinomial. Lenguajes y bibliotecas para trabajo matricial. (2 semanas)

Unidad 3: Problemas de clasificación y regresión logística.

Clasificación. Regresión logística: modelos y supuestos. Frontera. Función de costos. Gradiente descendente para regresión logística y su convergencia. Importancia de verificar el cómputo del gradiente mediante otras técnicas de diferenciación. Clasificación múltiple. Regularización: problemas de sobreajuste (*overfitting*) y aplicación en regresión polinomial y logística. (2 semanas)

Unidad 4: Redes neuronales.

Modelos con hipótesis no lineales. Presentación de la redes neuronales y cómo permiten abordar los límites de la linealidad. Clasificación múltiple con redes neuronales. Backpropagation. Presentar la idea de aprendizaje profundo (1 semana).

Unidad 5: Ajuste de modelos.

Conjunto de entrenamiento, de validación y de prueba. Evaluación de hipótesis: parcialidad (*bias*) y varianza. Curvas de aprendizaje. Análisis de errores y medidas para clases de conjuntos desbalanceados. (1 semana)

Unidad 6: Agrupamiento.

Agrupamiento: algoritmo de k promedios (*k-means*), elección de la cantidad de grupos e inicialización del algoritmo. (1 semana)

Unidad 7: Reducción de dimensionalidad.

Reducción de dimensionalidad: motivación, algoritmos de componentes principales, reconstrucción en términos de las variables originales luego de operar con los componentes hallados, selección apropiada del número de componentes. (media semana)

Unidad 8: Detección de anomalías.

Motivación. Algoritmo basado en distribución normal. Diferencias con estrategias de aprendizaje supervisado. Elección de variables para la detección de anomalías. (media semana).

Unidad 9: Sistemas de recomendación y aprendizaje online.

Motivación. Filtrado colaborativo y factorización de matrices de bajo rango con normalización por promedio. Aprendizaje *online*. (media semana).

Unidad 10: Grandes datos.

Variantes de gradiente descendente para el uso de grandes volúmenes de datos y su convergencia. *Map reduce*. (1 semana).

Unidad 11: Estrategias para la aplicación práctica de Ciencia de Datos.

Secuencia de problemas encadenados para la resolución de uno mayor (*pipelining*). Generación de datos artificiales, Análisis de techo. (1 semana).

Unidad 12: Otros algoritmos de clasificación.

Máquinas de vectores soporte. Árboles de decisión. Bayesiano ingenuo (1 semana).

Bibliografía obligatoria:

- Trevor Hastie, Robert Tibshirani y Jerome Friedman. 2009. The Elements of Statistical Learning (2nd ed.). Springer.

Organización de las clases:

Cada semana se realiza un desarrollo teórico inicial y una clase práctica de fijación de contenidos que incluye ejercitación y consultas. Las clases siguen las unidades temáticas en orden.

Guías prácticas

Objetivos de la Práctica 1 - Ciencia de Datos y problemas abordados: l
Introducir la noción de Ciencia de Datos, sus relaciones con otras disciplinas y los problemas abordados por esta.

Objetivos de la Práctica 2 - Problema de regresión lineal, algoritmos para resolverla y usos en problemas polinomiales:
Trabajar los supuestos de los modelos lineales y los algoritmos para resolverlos. Introducir los problemas que estos pueden presentar en la práctica y trabajar las técnicas para abordarlos. Presentar la necesidad de modelar polinomial algunas situaciones y cómo se pueden modificar los mismos algoritmos destinados a modelos lineales para resolver los primeros .

Objetivos de la Práctica 3 - Problemas de clasificación y regresión logística:
Introducir el problema de clasificación y la noción de frontera, así como la regresión logística como algoritmo para resolver este problema. Presentar los problemas que puede presentar en su uso práctico este algoritmo (principalmente el sobreajuste) y cómo se lo puede abordar. Mostrar que su aplicación a dos clases puede extenderse exitosamente a un mayor número de clases.

Objetivos de la Práctica 4 - Redes Neuronales:
Introducir esta técnica usada para regresiones y clasificación y mostrar que puede usarse para resolver problemas no lineales.

Objetivos de la Práctica 5 - Ajuste de modelos:
Introducir la técnica de diseño de algoritmos golosos para problemas de optimización. Identificar problemas que puedan ser resueltos óptimamente por esta técnica: subestructura óptima y optimización local que lleve a la optimización global. Comparar

esta técnica con la de programación dinámica: los requisitos, los algoritmos resultantes y sus complejidades.

Objetivos de la Práctica 6 - Agrupamiento:
Introducir el problema, el algoritmo típico para el problema y la cuestión de la cantidad adecuada de grupos.

Objetivos de la Práctica 7 - Reducción de dimensionalidad.

Introducir el problema, el algoritmo típico para el problema y la cuestión de las componentes adecuadas. Trabajar la reconstrucción de la solución hallada en términos de las variables originales

Objetivos de la Práctica 8 - Detección de anomalías.

Introducir el problema, el algoritmo típico para el problema y la cuestión de las variables apropiadas para la detección.

Objetivos de la Práctica 9 - Sistemas de recomendación.

Introducir el problema y el algoritmo típico para el problema.

Objetivos de la Práctica 10 - Grandes datos.

Trabajar sobre las variantes de gradiente descendente para el uso de grandes volúmenes de datos y los problemas de convergencia que presenta. Ejercitar la idea de *Map reduce*.

Objetivos de la Práctica 11 - Estrategias para la aplicación práctica de Ciencia de Datos.

Trabajar sobre la ejemplos reales la secuencia de problemas encadenados para la resolución de uno mayor (*pipelining*). , la generación de datos artificiales y el análisis de techo.

Objetivos de la Práctica 12 - Otros algoritmos de clasificación.

Aplicar en ejemplos las máquinas de vectores soporte, los árboles de decisión y el algoritmo bayesiano ingenuo y comparar resultados.

Modalidad de evaluación:

Los mecanismos de evaluación en modalidades libre y presencial de esta asignatura están reglamentados según los siguientes artículos del Régimen de estudios de la UNQ (Res. CS 201/18)

En la modalidad de libre, se evaluarán los contenidos de la asignatura con un examen escrito, un examen oral e instancias de evaluación similares a las realizadas en la modalidad presencial.

CRONOGRAMA TENTATIVO

Semana	Tema/unidad	Actividad*				Evaluación
		Teórico	Práctico			
			Res Prob.	Lab.	Otros Especificar	
1	Noción de Ciencia de Datos y problemas abordados	x	x	X		
2	Problema de regresión lineal, algoritmos para resolverla y usos en problemas polinomiales	x	x	X		
3	Problema de regresión lineal, algoritmos para resolverla y usos en problemas polinomiales	x	x	X		
4	Problemas de clasificación y regresión logística	x	x	X		
5	Problemas de clasificación y regresión logística	x	x	X		
6	Redes neuronales	x	x	X		
7	Ajuste de modelos	x	x	X		
8	Parcial		x	X		Parcial
9	Agrupamiento	x	x	X		

10	Reducción de dimensionalidad / Detección de anomalías	x	x	X		
11	Sistemas de recomendación y aprendizaje online	x	x	X		
12	Grandes datos	x	x	X		
13	Estrategias prácticas para la aplicación de Ciencia de Datos	x	x	X		
14	Otros algoritmos de Clasificación	x	x	X		
15	Presentación de evaluación y desarrollo		x	X		TP
16	Consultas y entrega de TP		x	X		TP

***INDIQUE CON UNA CRUZ LA MODALIDAD**

Firmado digitalmente por: SUAREZ Mariana Alejandra
Directora del Departamento de Ciencia y Tecnologla
UNIVERSIDAD NACIONAL DE QUILMES