APRENDIZAJE DE MÁQUINA - PARCIAL 1

Fecha: septiembre 6 de 2024. Duración: 3 horas

REGLAMENTO DE ESTUDIANTES de la PONTIFICIA UNIVERSIDAD JAVERIANA

Faltas disciplinarias graves y gravísimas:

Artículo 123 inciso (d): El fraude en actividades, trabajos y evaluaciones académicos y la posesión o utilización de material no autorizado en los mismos. (Falta disciplinaria grave)

Artículo 124 inciso (b): Todas las modalidades de plagio. (Falta disciplinaria gravísima)

Artículo 124 inciso (e): La suplantación en una evaluación académica, en exámenes preparatorios, en trabajos de grado y tesis. (Falta disciplinaria gravísima)

Sanciones Disciplinarias:

Artículo 127: Las faltas graves serán sancionadas con amonestación escrita con cargo a la hoja de vida del estudiante y la imposición de matrícula condicional durante el tiempo necesario para cumplir la condición.

Artículo 128: Adicional a la sanción disciplinaria, el fraude en actividades, trabajos y evaluaciones académicos se sancionará académicamente con la pérdida de la asignatura, la cual será calificada con nota definitiva de cero punto cero (0.0).

NO SE PERMITE INTERCAMBIO DE INFORMACIÓN CON PERSONA ALGUNA DURANTE EL DESARROLLO DEL PARCIAL. SOLO PUEDE INTERACTUAR CON EL PROFESOR. PARA COMUNICARSE CON EL PROFESOR, ENVÍELE UN MENSAJE PRIVADO POR TEAMS.

IMPORTANTE

- Debe generar un notebook para cada punto, responda las preguntas planteadas dentro de una celda de texto en el notebook. Nombre cada notebook así: "**Punto X – nombre del estudiante**". Antes de que se cumpla el plazo límite, debe cargar los 3 notebooks en el buzón "Parcial 1" configurado en la plataforma de BS. Si se cargan después del plazo, el parcial se califica sobre 4.0 si es en los siguientes 30 minutos. Después de estos 30 minutos ya no se recibirán parciales.
- Para el entrenamiento y prueba de los modelos, **solo puede utilizar los comandos vistos en clase**.
- Para todos los comandos que involucren una selección aleatoria, por ejemplo, dividir el dataset entre entrenamiento y prueba, o seleccionar registros de un dataset, fije siempre la semilla con los **últimos 3 números de su cédula** de ciudadanía.
- Para evaluar el desempeño de sus modelos, sobre los datasets de prueba, utilice en problemas de clasificación**: accuracy, y F-Score** para cada clase; en problemas de regresión, **RMSE y R2**.
- Si hay datos faltantes simplemente elimine los registros donde se presenten estos casos.

PUNTO 1 (20). K-NEAREST NEIGHBORS

Características biomecánicas de los pacientes ortopédicos: Clasificación de pacientes basada en seis características.

Objetivo: Clasificación de los pacientes como pertenecientes a una de dos categorías: Normal o Anormal (Hernia discal o Espondilolistesis).

Dataset: Biomecánica-Ortopedia.csv

Descripciones de los campos:

Cada paciente está representado en el conjunto de datos por seis atributos biomecánicos derivados de la forma y orientación de la pelvis y la columna lumbar (cada uno es una columna):

1. pelvic_incidence: incidencia pélvica
2. pelvic_tilt numeric: inclinación pélvica
3. lumbar_lordosis_angle: ángulo de lordosis lumbar
4. sacral_slope: inclinación sacra
5. pelvic_radius: radio pélvico
6. degree_spondylolisthesis: grado de espondilolistesis
7. class: clasificación del paciente

Preguntas:

1. (6) Encuentre el mejor valor de K desde el punto de vista de ACCURACY. Evalúe el desempeño del modelo con dicho valor.
2. (7) Plantee un valor de K que lleve a un overfitting. Explique porqué se da el overfitting.
3. (7) Plantee un valor de K que lleve a un underfitting. Explique porqué se da el underfitting.

PUNTO 2 (12). MODELO NAIVE BAYES

Influencia de características demográficas sobre el ingreso

Objetivo: Predicción del ingreso de una persona con base en variables categóricas

Dataset: Censo.csv (seleccione 10.000 registros aleatoriamente, tome solo las variables categóricas)

Descripciones de los campos:

1. age: Edad de la persona
2. workclass: Clase de trabajo que tiene la persona
3. fnlwgt: Indicador numérico de la proporción de la población que la persona representa
4. education: Nivel de educación
5. education_num: número de años de estudio
6. marital_status: Estado civil de la persona
7. occupation: Area en que la persona trabaja
8. relationship: Tipo de relación que tiene la persona
9. race: Raza de la persona
10. sex: Sexo de la persona
11. capital_gain: Ganancias de capital obtenidas
12. capital_loss: Pérdidas de capital
13. hours_per_week: Número medio de horas de trabajo a la semana
14. native_country: País de origen
15. salary: Nivel de ingresos

Preguntas:

1. (12) Dentro de su modelo, ensaye 3 valores del umbral de probabilidad, para decidir si una persona tiene un nivel de ingreso menor o mayor a 50 mil dólares. ¿Cuál valor ofrece el mejor desempeño?

PUNTO 3 (18). ANÁLISIS DE REGRESIÓN / REGULARIZACIÓN

Precios de diamantes

Objetivo: Predicción del precio de un diamante con base en atributos físicos del mismo

Dataset: Diamantes.csv (seleccione 10.000 registros aleatoriamente)

APRENDIZAJE DE MÁQUINA - PARCIAL 1

Fecha: septiembre 6 de 2024. Duración: 3 horas

Pontificia Universidad JAVERIANA Colombia

Descripción de los campos:

1. Índice: número consecutivo
2. Precio: precio en dólares estadounidenses
3. Quilates: peso del diamante
4. Talla: calidad de la talla (**0:** regular, **1:** buena, **2:** ideal, **3:** superior, **4:** muy buena)
5. Color: color del diamante (**0:** D, **1:** E, **2:** F, **3:** G, **4:** H, **5:** I, **6:** J)
6. Claridad: medida de la claridad del diamante (**0:** I1, **1:** IF, **2:** SI1, **3:** SI2, **4:** VS1, **5:** VS2, **6:** VVS1, **7:** VVS2)
7. Profundidad: porcentaje de profundidad total
8. Tabla: anchura de la parte superior del diamante en relación con el punto más ancho
9. X: longitud en mm
10. Y: anchura en mm
11. Z: profundidad en mm

Preguntas:

1. (6) Desarrolle un modelo de regresión lineal múltiple – utilizando el método de **stepwise** – para estimar el precio de un diamante. Evalúe su desempeño.
2. (12) Desarrolle un modelo de regresión lineal múltiple usando regularización **ELASTIC NET**. Optimice los parámetros correspondientes. Evalúe su desempeño. Escriba la ecuación final de la función objetivo a minimizar (con los parámetros optimizados). ¿Tiene un mejor desempeño que el modelo sin regularización?