Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
Computación: Ciencia y Tecnología del Mundo Digital

Tarea Grande 2

Profesores Vicente Domínguez – Luis Ramírez

Anunciada: 01 de octubre de 2018

Indicaciones

- Fecha de Entrega: 16 de octubre de 2018
- Debes entregar tu tarea en tu repositorio GitHub privado asignado para esta evaluación.
- Cada hora o fracción de atraso descuenta 0,5 puntos de la nota que obtengas.
- La tarea es *en parejas*. La copia será evaluada con nota 1 en el la tarea, además de las sanciones disciplinarias correspondientes.

Objetivo

El objetivo de esta tarea es que aprendas a:

- Utlizar Python para preprocesar un set de datos.
- Analizar un set de datos para elegir las mejores características para solucionar un problema.
- Aprender a usar el paquete `scikit-learn` de Python.
- Medir el error de un modelo de predicción.
- Visualizar los resultados obtenidos mediante `Altair` de Python

Descripción de los datos

Archivos

En tu repositorio de la tarea, encontrarás un archivo denominado `winequality.csv`. Este fichero contiene información sobre 3918 vinos, como un identificador único, su calidad y datos sobre su composición química.

En esta tarea, tu grupo deberá entrenar modelos que sean capaces de clasificar los vinos según sus cualidades para predecir la calidad de éste.

Set de datos

En el dataset, cada fila corresponde a un vino y posee los siguientes atributos por columna:

- `id`: Numérico - Identificador del vino.
- `acidez fija`: Numérico - Ácidos naturales del vino.
- `acidez volátil`: Numérico- Ácidos formados durante la fermentación.
- `ácido cítrico`: Numérico - Antioxidante presente en el vino.
- `azúcar residual`: Numérico - Azúcar restante luego de la fermentación.
- `cloruros`: Numérico - Cantidad de cloruro presente en el vino.
- `dióxido sulfúrico libre`: Numérico - Compontente que entrega propiedades antioxidantes y antisépticas.
- `dióxido sulfúrico total`: Numérico - Suma de dióxido sulfúrico libre y combinado.
- `densidad`: Numérico - Densidad del vino.
- `pH`: Numérico - El vino es ácido por lo que el pH se encuentra en el rango de 0 a 7.
- `sulfatos`: Numérico - Utilizados para el tratamiento de la uva.
- `alcohol`: Numérico - Grado de alcohol del vino.
- `calidad`: Numérico - Calidad del vino representada en una escala de 0 a 10.

Parte 1: Preprocesamiento de los datos

Los datos entregados tienen algunos valores faltantes, los cuales se representan como `'N/A'`. Antes de entrenar un modelo deben arreglar estos datos, quedando a su criterio cómo lo hacen.
Además, deberán **normalizar** los datos de modo que se obtengan mejores resultados.
Finalmente, en esta tarea se espera que los datos sean manipulados por medio de la librería **pandas**, por lo que su uso será obligatorio tanto en el preprocesamiento como en el procesamiento mismo de los datos.

Parte 2: Modelos

A modo de predecir posibles datos no etiquetados (es decir, que no tienen una calidad asociada), deberás entrenar modelos de predicción, los cuales son:

K-Nearest Neighbors (KNN)

Método que sirve para clasificar una instancia según la clase mas popular entre sus vecinos más cercanos.

Árboles de Decisión

Estructura de árbol formada por nodos que representan reglas de desición.
Para ambos modelos, deberás probar con 5 parámetros diferentes. Para el caso de KNN, el parámetro será la cantidad de vecinos a revisar, mientras que en los árboles de decisión deberán probar con 5 combinaciones distintas de profundidad máxima del árbol y tamaño de las hojas.

Parte 3: Evaluación de clasificadores

Luego de entrenar tus modelos, deberás probarlos, por lo que será necesario dividir los datos para poder entrenar y testear sus modelos. La sugerencia es usar una división aleatoria de $70\,\%/30\,\%$ entre entrenamiento y *testing*.
Se espera que testeen sus modelos usando este subconjunto de sus datos y sean capaces de argumentar cuál de sus modelos hace mejor el trabajo de clasificación a partir de esos resultados.

Parte 4: Clustering

En esta parte, deberás **reducir la dimensionalidad** de tus datos por medio de *Principal Component Analysis (PCA)*. La idea de hacer esto es poder visualizar tu dataset en **2** dimensiones.
Luego de aplicar la transformación, se debe visualizar el dataset por medio de la librería **Altair**.
Finalmente, deberás aplicar **K-Means** y **Meanshift** para identificar clusters, los cuales deberás graficar de modo que se puedan ver los diferentes clusters y la calidad de cada vino simultáneamente.
Cada uno de estos algoritmos requiere de un parámetro, por lo que deberás graficar ambos para 5 parámetros distintos a tu elección.

Parte 5: Informe

Además de lo pedido anteriormente, deberán responder las siguientes preguntas con respecto a la tarea:

1. ¿Cómo abordaste los datos faltantes? ¿En qué podría afectar a tu tarea el abordar esta parte de manera distinta?
2. ¿Cuál es la matriz de confusión de cada modelo? Incluye imágenes de cada matriz.
3. Considere que un vino es **bueno** cuando su calidad es mayor o igual a 6.

   ¿Cuál es el *precision* y el *recall* de cada uno de los modelos entrenados? ¿Que miden estas métricas? ¿Existe alguna otra métrica que las agrupe a ámbas?
4. En la sección de clustering, ¿Se formaron clusters aparentes? De ser así, explique qué representa cada cluster. De lo contrario, explica por qué no se formaron clusters.
5. Muestre la métrica *silhouette score* para cada algoritmo de clustering. ¿Qué es lo que mide esta métrica? Argumente cuales son los mejores parámetros para cada algoritmo.

Bonus

Para el bonus, además deberás entrenar un modelo de **SVM**, explicando cómo funciona y visualizando sus resultados.

Formato de entrega

La entrega de esta tarea se hará por medio del repositorio GitHub privado de tu grupo, donde deberán entregar un archivo `.ipynb` con todo el desarrollo de la tarea y con las preguntas del informa respondidas donde corresponden. (Por ejemplo: Como abordar los datos faltantes debería ir sobre el bloque de código donde se aborda este problema.)