

## Tarea 4

Profesor: Andrés Meza

Auxiliar: Javiera Ahumada, Víctor Navarro

Fecha entrega: 10 junio (antes de las 23:59 horas)

- En esta tarea se corregirán sólo 2 problemas.
  - La tarea debe subida a u-cursos como un solo archivo en formato .ipynb, .html o .pdf antes de la fecha y hora indicada.
  - **No se aceptarán atrasos ni se recibirán archivos en los correos del profesor o los auxiliares.**
  - La solución de los problemas debe contener breves comentarios y explicaciones que faciliten su comprensión, especialmente de los códigos empleados.
- P1.** Usando los datos en *apogee.fits* grafique  $[\alpha/M]$  versus  $[Fe/H]$ . Tenga cuidado con los datos marcados con -9999 porque significa que el dato en esa celda no existe. Aplique el método de *Gaussian Mixing* implementado en scikit-learn o cualquier otro paquete para encontrar el mejor ajuste bi-dimensional a estos datos. Determine los valores de los parámetros del modelo que maximiza la "likelihood". Determine el número de Gaussianas que mejor ajusta los datos.
- P2.** Escriba un programa que implemente el método KDE con un kernel de Epanechnikov. Úselo para determinar el mejor ajuste para la distribución de temperatura efectiva  $T_{eff}$  en los datos *apogee.fits*. Utilice el MISE para determinar el  $h_{opt}$ .
- P3.** Utilice el método de K-means implementado en scikit-learn o cualquier otro paquete para determinar el número de clusters en la distribución de estrellas en un gráfico de  $[\alpha/M]$  versus  $[Fe/H]$ . Divida la muestra en  $k = 2, 3, 4$  y 5 grupos. ¿Cuántos grupos entregan la mejor representación de los datos? Explore distintos valores para la semilla inicial de los centros (promedios) de cada cluster. Compare estas distintas realizaciones calculando la distancia relativa entre los centroides de cada grupo. ¿Cuántas iteraciones requiere el método para converger?