Universidad de Santiago de Chile
Facultad de Ingeniería
Departamento de Ingeniería Informática

Asignatura : Taller de minería de datos avanzada
Programa : Magister en Ingeniería Informática
Profesor **:** Max Chacón Pacheco
Ayudante **:** Felipe-Andrés Bello Robles
Fecha Entrega Oral **:** 7 de junio de 2022
Fecha Entrega Escrito **:** 14 de junio de 2022

TALLER 5: Máquinas de Vectores soporte (SVM)

Objetivos:

- Realizar un proceso de selección de características.
- En base al ranking de características realizar modelamiento con SVM buscando el principio de parsimonia.
- Comprender las diferencias entre un kernel lineal y un RBF.
- Comprender el efecto de los parámetros C, y gamma según corresponda.
- Realizar método de búsqueda en grilla de parámetros y mediante la curva ROC encontrar el mejor clasificador
- Comparar con métodos de clasificación anteriores.

Aspectos importantes a considerar: Para obtener los resultados y cumplir los objetivos del laboratorio, se debe tener en cuenta los siguientes puntos:

- Utilizar la primera base de datos asignada (Taller 1) que corresponde a un problema de clasificación.
- Utilizar "R" http://www.r-project.org/ y su librería "e1071"
- Realizar una comparación la literatura, de manera de establecer la efectividad del método a la resolución del problema, incluyendo ventajas y desventajas de éste.

Escrito: Se debe elaborar un *Artículo* de máximo 6 páginas, según el formato:
https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines

	Puntos a evaluar	**Porcentaje**
Estructura del Artículo	Presentación, ortografía y redacción	5%
	Abstract e Introducción	10%
	Métodos (explicación del funcionamiento)	20%
	Resultados	20%
	Discusión	25%

	Conclusiones	20%

Observaciones:

Consultas al mail Felipe.bello@usach.cl

El trabajo debe ser presentado de forma oral (50%) y escrita (informe 50%) en horario de clases el día 7 de junio y 14 de junio de 2022. Disponen de 10 minutos de exposición y 5 para contestar preguntas de la comisión.

En el proceso de selección de características, la idea es ir simplificando el modelo hasta encontrar un punto de equilibrio donde obtener el modelo más simple con el menor error asociado (parsimonia). Para ello pueden usar una librería que integrar funciones del software WEKA, donde se puede obtener la importancia del atributo desde diferentes enfoques. https://cran.r-project.org/web/packages/RWeka/RWeka.pdf

Por ejemplo:

```
require(RWeka)

data(iris)

ranking<-InfoGainAttributeEval(Species ~ . , data = iris)

print(ranking)
```

```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.6982615    0.3855963    1.4180030    1.3784027
```

En los resultados el ranking queda:

Petal.Length, Petal.Width, sepal.length, sepal.width.

Al comparar estos resultados con los presentados en el práctico de “random forest” pueden ver las similitudes (importancia: precisión y Gini)

La información de las bases de datos se encuentra en la página:

https://cran.r-project.org/web/packages/e1071/e1071.pdf

Nota Final: Promedio simple de las experiencias.