UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS
CIENCIAS DE LA COMPUTACIÓN

MACHINE LEARNING
Laboratorio 2
(Primer Semestre del 2019)

Objetivos de aprendizaje:

- Entrenar clasificadores KNN.

---

1. Actividad en Weka

El algoritmo `IB1` fue presentado en el artículo *Instance-based learning algorithms* propuesto por David W. Aha, Dennis Kibler y Marc K. Albert [1]. El funcionamiento de este algoritmo es explicado en la sección 2, del artículo (principalmente en las sub-secciones 2.1, 2.2, 2.3 y 2.4).

Lectura (30 minutos). Lea las sub-secciones 2.1, 2.2, 2.3 y 2.4 del artículo [1] para entender como funciona el algoritmo `IB1`.

Cross-validation (10 minutos) Usando los conjuntos de datos `iris`, `ionosphere` y `diabetes`, entrene el algoritmo `IBk` usando la opción de prueba `Cross-validation` y luego anote la métrica F_1 en la tabla 1 usando diferentes valores para el parámetro *KNN*. Use los valores de *KNN* que se indican en la tabla 1 en la variable k.

Tabla 1: Casos de prueba para la opción Cross-validation.

Algoritmo	Conjunto de Datos	F_1 k=1	F_1 k=3	F_1 k=5	F_1 k=7	F_1 k=9
IBk	iris.arff					
IBk	ionosphere.arff					
IBk	diabetes.arff					

2. Actividad en RapidMiner

Usando datos de prueba (10 minutos) Se desea encontrar el modelo más óptimo para predecir el conjunto de datos denominado `Adult`. La descripción completa de este conjunto de datos la puede encontrar en la URL `https://archive.ics.uci.edu/ml/datasets/Adult`. Deberá usar como datos de entrenamiento el archivo `adult.csv` y como datos de prueba el archivo `adult_test.csv`. Ambos archivos se encuentran en el aula virtual. Utilice el algoritmo `k-NN`.

Preguntas de discusión

- ¿Los datos son consistente con los algoritmos ejecutados en Weka? Compare los valores de la precisión.

Validación cruzada (10 minutos) Se desea encontrar el modelo más óptimo para predecir la clase del conjunto de datos denominado `Car Evaluation` usando la técnica de la validación cruzada. La descripción completa de este conjunto de datos la puede encontrar en la URL `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`. El conjunto de datos se encuentra en el aula virtual. Utilice tanto el algoritmo `k-NN`.

3. Actividad en Python

Validación cruzada (15 minutos) Se desea encontrar el modelo más óptimo para predecir el conjunto de datos denominado `Fertility`. La descripción completa de este conjunto de datos la puede encontrar en la URL `https://archive.ics.uci.edu/ml/datasets/Fertility`. Deberá usar como datos de entrenamiento el archivo `fertility_Diagnosis.csv` usando la técnica de la validación cruzada. El archivo se encuentra en el aula virtual. Incluya dentro de los algoritmos a probar el algoritmo `KNeighborsClassifier`. Vea el siguiente programa para que sepa la librería a importar y los comandos a usar para el entrenamiento.

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors)
knn.fit(X_entrenamiento, y_entrenamiento)
```

Preguntas de discusión

- ¿Qué algoritmo consigue el mejor modelo? ¿Cómo justifica su elección?
- ¿Qué atributos seleccionó para entrenar el modelo? ¿Cómo justifica su elección?

Pando, 30 de abril de 2019.

Referencias

[1] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.