**UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS**
**CIENCIAS DE LA COMPUTACIÓN**

**MACHINE LEARNING**
**Laboratorio Regresión Logística**
**(Primer Semestre del 2019)**

Objetivos de aprendizaje:

- Entrenar modelos de regresión logística.

---

# 1. Actividad en Weka

**Predicción de diabetes (40 minutos)**. Se desea predecir la probabilidad que un paciente tenga diabetes. Los datos de los pacientes se encuentran en el archivo `diabetes.arff`. Se le pide que entrene un modelo usando el algoritmo `SimpleLogistic` usando los parámetros por defecto de algoritmo. Una vez obtenido el modelo, realice una predicción utilizando los siguientes datos:

- `preg` = 6
- `plas` = 148
- `pres` = 72
- `skin` = 35
- `insu` = 0
- `mass` = 33.6
- `pedi` = 0.627
- `age` = 50

Para realizar la predicción calcule el valor de la salida $x$ del modelo y obtenga la probabilidad usando la siguiente fórmula:

$$p(claseX) = \frac{1}{1 + e^{-x}}$$

**Preguntas de discusión**

- Describa al modelo. ¿Es un buen modelo? ¿Cómo verifica si es un buen modelo?
- Use realice el entrenamiento usando el algoritmo `Logistic`. ¿Qué diferencias encuentra en la salida? ¿Existe diferencia en la predicción?
- Compare la predicción realizada usando el algoritmo `Naïve Bayes`. Recuerde que los valores numéricos se manejan generalmente asumiendo que tienen una distribución de probabilidad normal o gaussiana.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

## 2. Actividad en RapidMiner

**Regresión Logística usando el conjunto de datos iris (20 minutos)** Se desea encontrar el modelo más óptimo para predecir el conjunto de datos denominado `iris.csv`. Deberá usar la técnica de validación cruzada para entrenar el modelo. Utilice el componente `Logistic Regression` para este ejercicio.

**Preguntas de discusión**

- ¿Cómo hace el algoritmo para poder predecir 3 clases?
- ¿Los datos son consistente con los algoritmos ejecutados en Weka?

## 3. Actividad en Python

**Predicción de la fertilidad (40 minutos)** En el archivo `Fertility_diagnosis.csv` se encuentran los datos de 100 voluntarios que se han realizado una prueba de fertilidad. En base a los datos de este archivo se desea predecir la probabilidad de infertilidad de determinadas instancias. Para lo cual se le pida que aplique el modelo de regresión logística. En el archivo `Fertility_diagnosis - descripcion.pdf` encontrará la descripción detallada de cada campo. La descripción completa de este conjunto de datos la puede encontrar en la URL `https://archive.ics.uci.edu/ml/datasets/Fertility`.

Use la técnica de la validación cruzada. Incluya dentro de los algoritmos a probar el algoritmo `LogisticRegression`. Vea el siguiente programa para que sepa la librería a importar y los comandos a usar para el entrenamiento.

```
1 from sklearn.linear_model import LogisticRegression
2
3 regressionlogistica = LogisticRegression()
4 regressionlogistica.fit(X_entrenamiento, y_entrenamiento)
```

**Preguntas de discusión**

- ¿Cuál es la eficiencia del modelo?
- ¿Qué atributos seleccionó para entrenar el modelo? ¿Cómo justifica su elección?
- ¿El valor de este modelo algorítmico es consistente con el modelo obtenido en Weka?
- ¿Existe alguna diferencia si es que normalizamos los atributos?

Monterrico, 16 de mayo de 2019.