

Reconocimiento de patrones y aprendizaje automático

Tarea 1: Regresión

Fecha de entrega: Sábado 13 de noviembre de 2021

Profesor: Andrés Aldana Gonzáles
Ayudante: Felipe Navarrete Córdova

1. Ejercicios

1. Contesta las siguientes preguntas (1 pt).

- a) ¿Qué tipo de problemas se pueden resolver mediante regresión?
- b) ¿La siguiente ecuación es un modelo de regresión lineal?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

- c) En regresión lineal, ¿Cómo se define e interpreta el coeficiente R^2 ?
- d) ¿Es conveniente utilizar el valor R^2 en regresión no lineal? En caso contrario, ¿qué métrica o valor se puede utilizar?
- e) ¿Qué es el sobre entrenamiento en una regresión?
- f) Supongamos que los resultados de una regresión muestran que el coeficiente de regresión es $R^2 = 1$. ¿Podría este modelo estar sobre entrenado? Justifica tu respuesta.

2. **Peso y Altura (5 pts).** El archivo weight-height .csv contiene los datos de peso (libras) y estatura (pulgadas) de hombres y mujeres. Realice los siguientes ejercicios:

- a) Realice una gráfica de puntos peso (eje y) vs altura (eje x) que muestre la relación entre ambas variables.
- b) Calcule el coeficiente de correlación de Pearson entre peso y altura.
- c) Realice y grafique una regresión lineal para calcular el peso de una persona en función de su altura. Proporcione los coeficientes de la regresión.
- d) Calcule e interprete el coeficiente de determinación R^2 .
- e) Grafique un histograma de los valores residuales de la regresión.
- f) ¿Cuál es el error esperado al hacer una predicción con este modelo?
- g) Calcule y grafique un intervalo de confianza de la regresión con un valor de confianza de 90 %.

- h) Dada una predicción del peso en función de la altura con este modelo, ¿cuál es la probabilidad de que el error entre el valor observado y la predicción sea menor a 20 libras? Justifique su respuesta.
3. **Regresión no lineal (5 pts).** El archivo **data.csv** contiene dos columnas, una con la variable independiente X y la otra con la variable dependiente Y . Realice los siguientes ejercicios.
- a) Realice una gráfica de puntos Y vs X que muestre la relación entre ambas variables.
 - b) Implemente una forma para identificar y eliminar outliers en este conjunto de datos.
 - c) Proponga y grafique un modelo de regresión que describa la relación entre las variables X y Y . Proporcione los coeficientes de la regresión. Sugerencia: La relación entre X y Y NO ES LINEAL y no es logarítmica.
 - d) Obtén el coeficiente de determinación R^2 e interpreta.
 - e) Grafique el histograma de residuales para este modelo
 - f) ¿Cuál es el error esperado al hacer una predicción con este modelo de regresión?
 - g) ¿Cuál es la probabilidad de que una predicción realizada con este modelo tenga un error mayor a 500?

2. Software recomendado

Por la simplicidad, facilidad de uso, y compatibilidad se recomienda utilizar *python* como language de programación y las siguientes bibliotecas:

- numpy
- pandas
- scikit-learn
- scipy (*optimize.curve_fit*)

3. Entregables

La tarea se debe entregar en un notebook de Jupyter con los resultados y las gráficas correctamente discutidos. Los archivos de datos deben estar en el mismo nivel de directorio del notebook para facilitar la ejecución de los programas.

4. Bibliografía recomendada

- Christopher M. Bishop - Pattern Recognition and Machine Learning-Springer (2011) - Chapter 3 - Linear Models for Regression