INTELIGENCIA ARTIFICIAL

Homework 1: Regresión lineal

Descripción breve

En el presente documento se presenta la explicación de los resultados obtenidos durante la ejecución de la práctica de Regresión Lineal.

Sebastián Pedrosa Granados y Germán Alejo Domínguez

Tabla de contenido

1. Introducción

A la hora de realizar un algoritmo de Machine Learning debemos comenzar por comprender la información que vamos a emplear para realizar tanto las fases de entrenamiento como de testeo del algoritmo. En el presente ejercicio se ha empleado un conjunto de 252 estimaciones del porcentaje de grasa corporal de personas basándonos en las siguientes 14 métricas corporales:

1. Densidad determinada por pesaje bajo el agua
2. Años
3. Peso en libras
4. Altura en pulgadas
5. Circunferencia del cuello (cm)
6. Circunferencia del pecho (cm)
7. Circunferencia del abdomen (cm)
8. Circunferencia de la cadera (cm)
9. Circunferencia del muslo (cm)
10. Circunferencia de la rodilla (cm)
11. Circunferencia del tobillo (cm)
12. Circunferencia del bíceps extendido (cm)
13. Circunferencia del antebrazo (cm)
14. Circunferencia de la muñeca (cm)
15. Porcentaje de grasa corporal

Ésta última métrica (15) corresponde precisamente a la estimación de la grasa corporal y constituye el objetivo de nuestro algoritmo: predecir en base a las otras 14 métricas el índice de grasa corporal de un sujeto mediante regresión lineal empleando la ecuación normal y el descenso del gradiente para hallar óptimos viables que nos permitan minimizar el error de la predicción.

Para realizar esto vamos a emplear la herramienta Octave para la programación, entrenamiento y testeo del modelo.

2. Modelo de regresión lineal según la ecuación normal

En este apartado se nos solicita realizar un modelo empleando la ecuación normal para resolver las thetas (θ) que nos permitirán predecir los índices de grasa corporal.

Ecuación normal:

$$\theta = (X^T X)^{-1} X^T y$$

2.1. Empleando el conjunto completo como test y training

Para comenzar se nos indica que debemos resolver el problema empleando el conjunto completo de datos tanto para entrenar al modelo como para probarlo. En este caso teníamos dos formas de llegar al objetivo: mediante el conjunto completo de datos realizando una regresión multivariable con el conjunto completo y realizando una regresión univariable por cada atributo, listando los errores cometidos de menor a mayor y seleccionando los 5 mejores atributos dados en el modelo anterior para realizar una regresión multivariable con este subconjunto.

En el primer caso podemos observar el siguiente error tras calcular las thetas oportunas: 0.48020

Mientras que en el segundo caso podemos ver que la lista de los errores cometidos corresponde a la siguiente:

Tabla de errores	
Atributo	Error
1. Densidad determinada por pesaje bajo el agua	0.35666
2. Años	6.55079
3. Peso en libras	5.31805
4. Altura en pulgadas	6.85683
5. Circunferencia del cuello (cm)	5.96289
6. Circunferencia del pecho (cm)	4.86356
7. Circunferencia del abdomen (cm)	3.91631
8. Circunferencia de la cadera (cm)	5.25048
9. Circunferencia del muslo (cm)	5.59295
10. Circunferencia de la rodilla (cm)	5.70790
11. Circunferencia del tobillo (cm)	6.57382
12. Circunferencia del bíceps extendido (cm)	5.93953
13. Circunferencia del antebrazo (cm)	6.42251
14. Circunferencia de la muñeca (cm)	6.38031

Los mejores resultados nos los dan los atributos 1, 7, 6, 8 y 3, con estos creamos un subconjunto y realizamos una regresión multivariable de la que obtenemos el error: 5.3180

2.1.1. Conclusión

Dados los resultados que arrojan ambos modelos, podemos afirmar que realizar una regresión multivariable sobre el conjunto completo de datos resulta más eficiente que realizar una regresión univariable por cada atributo y luego una multivariable sobre el conjunto de los mejores.

2.2. Empleando 30% del conjunto como test y el 70% como training

En este apartado se nos pide repetir la operativa anterior realizando un holdout aleatorio manteniendo un conjunto para training con el 70% del conjunto completo y un conjunto para test del 30% del completo. Este experimento se realiza con la intención de averiguar si merece la pena realizar una separación entre el conjunto de entrenamiento y el conjunto de prueba de nuestro modelo.

En este apartado volvemos a tener las mismas maneras de realizar el modelo por lo que en un primer lugar vamos a obtener del modelo multivariable con el conjunto completo de datos de training un error de 0.42734 al realizar la predicción empleando el conjunto de test.

Por otro lado, si realizamos la regresión univariable sobre todos los atributos del conjunto de datos de forma individual obtendremos la siguiente tabla de errores:

Tabla de errores	
Atributo	Error
1. Densidad determinada por pesaje bajo el agua	0.35666
2. Años	0.35409
3. Peso en libras	6.89913
4. Altura en pulgadas	5.51098
5. Circunferencia del cuello (cm)	7.07744
6. Circunferencia del pecho (cm)	6.18691
7. Circunferencia del abdomen (cm)	5.19969
8. Circunferencia de la cadera (cm)	3.95314
9. Circunferencia del muslo (cm)	5.46940
10. Circunferencia de la rodilla (cm)	5.65474
11. Circunferencia del tobillo (cm)	5.56737
12. Circunferencia del bíceps extendido (cm)	6.59515
13. Circunferencia del antebrazo (cm)	5.93762
14. Circunferencia de la muñeca (cm)	6.52058

De estos obtenemos que los mejores resultados los arrojan los atributos 1, 2, 8, 7 y 9 obteniendo un error en la predicción de 3.7963.

2.2.1. Conclusión

Dados los resultados arrojados por este segundo experimento podemos ver que el método multivariable sigue siendo sustancialmente mejor que la realización de una regresión univariable con ranking de atributos.

2.3. Conclusión general

Como conclusión global, tras analizar las conclusiones de los apartados anteriores, podemos afirmar que realizar un holdout que divida el conjunto principal en dos subconjuntos, test y training, mejora inequívocamente la eficiencia del modelo a la hora de predecir y que el uso de la regresión multivariable cuando se tienen conjuntos con diferentes métricas para la predicción resulta bastante más práctico y eficiente que emplear métodos más complejos como elaborar un ranking de los 5 mejores atributos de forma univariable y realizar a posteriori una regresión multivariable.

3. Modelo de regresión empleando el descenso del gradiente

Para este apartado vamos a emplear el método de descenso del gradiente para obtener las thetas óptimas para el modelo empleando el holdout realizado en el apartado 2.2 del presente documento.

Método de descenso del gradiente:

$$\theta = \theta - \alpha \frac{1}{m} \left(X^T * (X * \theta - y) \right)$$

3.1. Empleando el conjunto completo de datos

En este primer apartado se nos pide que empleemos el conjunto completo de los datos de training para realizar la regresión. En primer lugar realizamos una normalización de los datos de entrenamiento debido a que se encuentran en rangos muy amplios y esto puede resultar en un modelo que no se ajuste correctamente a la información de las muestras

Ecuación para la normalización de los datos:

$$X = \frac{X - M_e}{\sigma}$$

Una vez realizado el modelo obtenemos la siguiente gráfica:

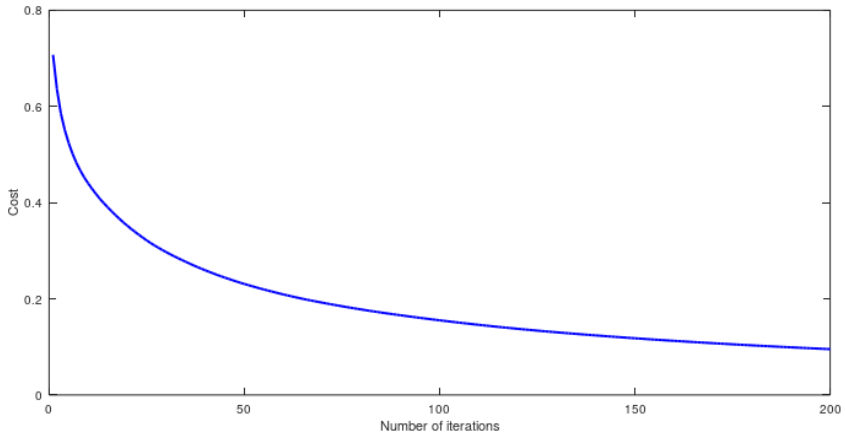

Ilustración 1 Gráfica descenso de gradiente conjunto completo

En esta gráfica podemos observar como con un parámetro Alpha de 0.03 y 200 iteraciones el algoritmo es capaz de encontrar el mínimo global para las thetas del modelo.

3.2. Empleando el conjunto formado por los 5 mejores

En este caso vamos a realizar un modelo empleando también el método de descenso del gradiente, pero empleando en esta ocasión el conjunto de los mejores atributos creado en el apartado 2.2. Para realizar este modelo también vamos a necesitar normalizar los datos ya que la información contenida en sigue encontrándose en rangos muy dispares, por lo que, para ajustar lo máximo posible la recta a los datos, es necesario situarlos en un rango mas acotado. Para la normalización emplearemos la formula mostrada en el apartado 3.1.

Una vez construyamos el modelo podremos obtener la siguiente gráfica:

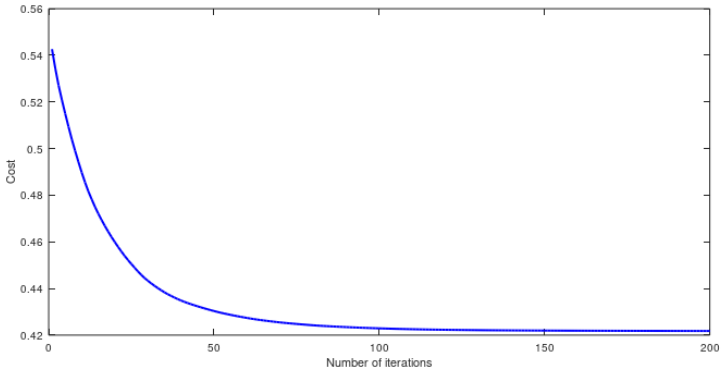

Ilustración 2 Descenso del gradiente con los 5 mejores atributos

En esta gráfica podemos observar como en este caso el óptimo se encuentra de forma más rápida, en menos de 200 iteraciones el algoritmo se estanca hallando un mínimo global con un parámetro Alpha de 0.3, lo que supone una velocidad de aprendizaje bastante elevada.