Machine learning

Davide Abete, Fabrizio Cominetti e Agazzi Ruben

January 5, 2022

Contents

Il progetto consiste nell'analisi di vari modelli di regressione applicati ad un dataset contenente le informazioni ambientali riferite alla temperatura globale a partire dal Gennaio 1850 al Novembre 2015.

1 Introduzione

Il cambiamento climatico è la più grande minaccia del nostro tempo. Nell'ultimo secolo la temperatura globale è sempre più aumentata.Lo scopo del progetto consiste nel realizzare e analizzare vari modelli di regressione per prevedere la temperatura media globale del mese successivo, sulla base dei dati rilevati durante la mensilità precedente.La scelta di questo progetto è stata influenzata dalle tematiche riguardanti il cambiamento climatico, un tema in costante ascesa di interesse e di importanza sempre più rilevante. Per effettuare questa previsione abbiamo usato un dataset contenente varie misure di temperatura media, rilevate ogni mese a partire dal Gennaio 1750 al Novembre 2015.

2 Dataset

Il Dataset utilizzato è "Climate Change: Earth Surface Temperature Data".Il Dataset è stato ottenuto da Kaggle, che è una versione ripulita del Dataset formito dalla "Berkley Earth", un'organizzazione non-profit indipendente specializzatta in "Environmental Data Science".

Fra i vari Dataset forniti su Kaggle è stato scelto quello riguardante le temperature globali.

2.1 Descrizione del Dataset

Il Dataset è composto da 3192 righe e da 9 colonne. Le colonne sono le seguenti:

- dt: Data di rilevamento dei dati
- landAverageTemperature: temperatura media globale del terreno espressa in gradi Celsius.
- LandAverageTemperatureUncertainity: Intervallo di confidenza al 95% intorno alla media della temperatura globale
- LandMaxTemperature: media della temperatura massima globale del terreno espressa in gradi Celsius
- LandMaxTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura massima globale
- LandMinTemperature: media della temperatura minima globale del terreno espressa in gradi Celsius
- LandMinTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura minima globale
- LandAndOceanAverageTemperature: Temperatura media globale del terrestre e oceanica espressa in gradi Celsius.
- LandAndOceanAverageTemperatureUncertainty: Intervallo di confidenza al 95% intorno alla media della temperatura media terrestre e oceanica globale.

2.2 Esplorazione dei dati

Per la fase di esplorazione dei dati è stata utilizzata la piattaforma Knime, in particolare è stato usato il

nodo Statistics, attraverso il quale sono stati rilevati i seguenti indici e numero di missing values.

Colonna	Media	STD	Missing V.
landAvgTemp	8.3747	4.3813	12
LandAvgTempUnc	0.9385	1.0964	12
LandMaxTemp	14.3506	4.3096	1200
LandMaxTempUnc	0.4798	0.5832	1200
LandMinTemp	2.7436	4.1558	1200
LandMinTempUnc	0.4318	0.4458	1200
LndOcnAvgTemp	15.2126	1.2741	1200
LndOcnAvgTempUnc	0.1285	0.0736	1200

La fase di esplorazione ha permesso di identificare la variabile target che ci interessa prevedere, ovvero *landAverageTemperature* del mese successivo, e le colonne restanti che fungono da attributi esplicativi. La variabile target selezionata è di tipo numerico, continua e assume valori fra $[-2.080, 19.021]$ con 3 cifre decimali.

3 Pre-Processing

Durante la fase di data exploration si è potuto osservare che quasi tutte le colonne hanno 1200 valori mancanti, in quanto le osservazioni effettuate prima del Gennaio 1850 non sono stati rilevati tali dati.

3.1 Missing Values

Per risolvere questa problematica si è deciso di rimuovere le righe che presentano missing values, eliminando quindi 1200 righe e riducendo l'intervallo di tempo dei valori che ora partono dal Gennaio 1850 invece che dal Gennaio 1750. Per eliminare le righe che presentano missing values è stato utilizzato il nodo di Knime chiamato *«Missing Values»*

3.2 Data Augmentation

Per poter allenare i vari classificatori ci serve il dato relativo alla temperatura media terrestre del mese successivo. Per fare questo abbiamo ordinato decrescente i dati in base alla colonna della data utilizzando il nodo *«Sorter»* e in seguito abbiamo utilizzato il nodo *«Lag Column»* per creare una nuova colonna contenente la temperatura media del mese successivo.

3.3 Selezione delle Variabili

Per selezionare le variabili da utilizzare per l'apprendimento dei classificatori è stato utilizzato un filtro di correlazione in modo da eliminare attributi ridondanti. Il valore soglia di correlazione scelto è di 0.9. Alla fine del processo si feature selection sono state tenute 6 colonne su 10. Per effettuare questa operazione sono stati utilizzati i nodi *«Linear Correlation»* e *«Correlation Filter»*

4 Modelli di Regressione