

HW3

Given the Kaggle notebook at this link (<https://www.kaggle.com/code/pokekarat/ekarat-hw1>) and the ML workflow provided below, please improve the existing ML's performance by reducing the MSE. Please also provide a detailed explanation of the changes made at each stage from 2 to 7.

Dataset: <https://www.kaggle.com/datasets/sohier/calcofi>

Your Kaggle notebook: [HW-Regression \(Click For More Detail\)](#)

Topic : ปัจจัยที่ส่งผลต่อคุณภาพน้ำทะเล (ความลึก กับ potential density)

All figures are in SVG format (where available).

1. Overview (2-3 sentences)

คุณภาพน้ำทะเลเป็นตัวบ่งชี้สภาพแวดล้อมของทะเล และการเปลี่ยนแปลงที่เกิดขึ้นในมหาสมุทร การเปลี่ยนแปลงของคุณภาพน้ำทะเลส่งผลต่อการจัดเรียงตัวของน้ำในแนวตั้ง น้ำอุ่น และน้ำเย็นจะไม่ผสมกันง่าย ทำให้น้ำทะเลแบ่งออกเป็นชั้นๆ งานนี้ใช้เทคนิคการเรียนรู้ของเครื่องหมายเพื่อช่วยทำนายคุณภาพน้ำทะเลจากข้อมูลที่มีอยู่

2. More specific details (2-3 sentences)

การศึกษาน้ำที่กำหนดให้คุณภาพน้ำทะเล เป็นตัวแปรเป้าหมาย โดยใช้ความลึกของน้ำทะเล (Depth) และค่า potential density (STheta) เป็นตัวแปรอิสระ โดยข้อมูลจะถูกแบ่งออกเป็นชุดฝึกและชุดทดสอบเพื่อสร้างและประเมินโมเดล โมเดลจะเรียนรู้ความสัมพันธ์ระหว่างตัวแปรอิสระกับคุณภาพน้ำทะเล

3. Existing issues (2-3 sentences)

การใช้แรงกด (R_PRES) หรือความลึก (Depth) ร่วมกับความเค็ม (Salinity) ยังไม่สามารถอธิบายโครงสร้างการเปลี่ยนแปลงของคุณภาพน้ำทะเลได้อย่างครบถ้วน ผลงานให้ค่า MSE ยังคงอยู่ในระดับสูง อย่างไรก็ตาม การใช้ความลึกร่วมกับค่า potential density (STheta) ซึ่งสะท้อนการแยกชั้นของน้ำทะเลโดยตรง สามารถลดข้อจำกัดดังกล่าว และให้ผลการทำนายดีขึ้น

4. Motivation (2-3 sentences)

การลดค่า Mean Squared Error (MSE) ของการทำนายคุณภาพน้ำทะเลช่วยเพิ่มความแม่นยำและความน่าเชื่อถือของโมเดล การใช้ตัวแปรความลึก (Depth) และ potential density (STheta) ช่วยให้โมเดลสามารถอธิบายโครงสร้างแนวตั้งของน้ำทะเลได้ดียิ่งขึ้น ดังนั้นจึงมีการนำเทคนิคการเรียนรู้ของเครื่องและการทำ feature engineering มาใช้เพื่อปรับปรุงประสิทธิภาพของโมเดล

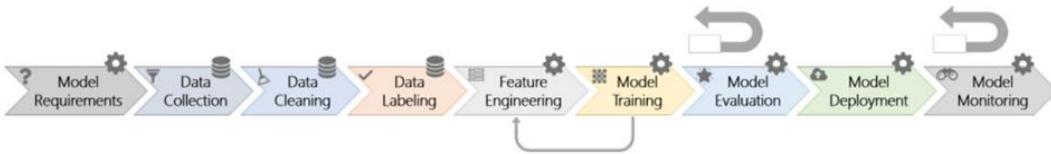
5. Problem statement (Input and Objective)

ข้อมูลนำเข้า (Input):

กำหนดให้มีชุดข้อมูลฝึก D ซึ่งประกอบด้วยตัวแปรอิสระ ได้แก่ ความลึกของน้ำทะเล (Depth) และค่า potential density (STheta) พร้อมกับค่าคุณภาพน้ำทะเลเป็นตัวแปรตาม และชุดข้อมูลทดสอบ B ที่มีโครงสร้างข้อมูลเดียวกัน

วัตถุประสงค์ (Objective):

กำหนดให้มีชุดข้อมูลฝึก D และชุดข้อมูลทดสอบ B วัตถุประสงค์ของงานนี้คือการใช้ข้อมูลจาก D เพื่อค้นหาสมมติฐาน H ที่ทำให้ค่า Mean Squared Error (MSE) ระหว่างค่าคุณภาพน้ำทะเลจริง y ในชุดข้อมูล B และค่าที่โมเดลทำนาย H(x_B) มีค่าน้อยที่สุด



2. Data Collection

ข้อมูลที่ใช้ในการศึกษานี้ได้มาจากชุดข้อมูลของ CalCOFI (California Cooperative Oceanic Fisheries Investigations) ซึ่งเป็นชุดข้อมูลสมมุติศาสตร์ที่มีความต่อเนื่องยาวนานตั้งแต่ปี ค.ศ. 1949 จนถึงปัจจุบัน ครอบคลุมสถานีตรวจวัดมากกว่า 50,000 สถานี บริเวณชายฝั่งตอนใต้และตอนกลางของรัฐแคลิฟอร์เนีย การศึกษานี้เลือกใช้ข้อมูลทางกายภาพของน้ำทะเล ได้แก่ อุณหภูมิ น้ำทะเล (Temperature), ความลึกของน้ำทะเล (Depth) และค่า potential density (STheta)

3. Data Cleaning

จัดการค่าที่ขาดหาย ค่าผิดปกติ และตรวจสอบความถูกต้องของหน่วยวัด เพื่อให้ข้อมูลมีคุณภาพและพร้อมสำหรับการวิเคราะห์ ด้วยโมเดลการเรียนรู้ของเครื่อง ด้วยวิธี forward fill คือ การนำค่าก่อนหน้าที่ไม่หาย มาใส่แทนค่า空缺 ไปที่หายไป

4. Data Labeling

กำหนดให้อุณหภูมน้ำทะเลเป็นตัวแปรเป้าหมาย (y) โดยมีค่าความลึก (Dept) เป็นตัวแปร X1 และ potential density (STheta) เป็นตัวแปร X2

5. Feature Engineering

เป็นขั้นตอนการปรับปรุง, สร้าง หรือเลือกตัวแปรอิสระ (X) เพื่อให้โมเดลสามารถเรียนรู้ความสัมพันธ์กับอุณหภูมน้ำทะเลได้ดีขึ้น ซึ่งผลโดยตรงต่อการลดค่า MSE ของโมเดล

1) การใช้ความลึกของน้ำทะเล (Depth)

อุณหภูมน้ำทะเลมีการเปลี่ยนแปลงตามความลึกอย่างชัดเจน โดยเฉพาะบริเวณชั้นผิวน้ำ ความลึกจึงเป็นตัวแปรสำคัญที่ช่วยอธิบายการลดลงของอุณหภูมิตามแนวตั้ง อย่างไรก็ตาม ความสัมพันธ์ดังกล่าวมักไม่เป็นเชิงเส้นตรง

2) การใช้ Potential Density (STheta)

ค่า potential density เป็นตัวแปรที่สะท้อนโครงสร้างการแยกชั้นของน้ำทะเล ซึ่งมีความสัมพันธ์กับอุณหภูมิและการผสมของมวลน้ำ ที่มีความหนาแน่นต่ำมักมีอุณหภูมิสูงกว่า ขณะที่น้ำที่มีความหนาแน่นสูงมักมีอุณหภูมิต่ำกว่า ดังนั้นการใช้ potential density เป็นตัวแปรอิสระช่วยให้โมเดลเข้าใจโครงสร้างทางกายภาพของน้ำทะเลได้ดีขึ้น

```

# Extract 2 columns 'T_degC', 'Salinity' for pure and better showing
bottle_df = bottle[['T_degC', 'Salinity', 'R_PRES']]

# And called again
bottle_df.columns = ['Temperature', 'Salinity', 'Pressure']

+ Code + Markdown

bottle_df = bottle_df[:500]      # lets take limit for speed regression calculating
bottle_df.head()


```

	Temperature	Salinity	Pressure
0	10.50	33.440	0
1	10.46	33.440	8
2	10.46	33.437	10
3	10.45	33.420	19
4	10.45	33.421	20

```

from sklearn.preprocessing import PolynomialFeatures

poly_df = PolynomialFeatures(degree = 6)
transform_poly = poly_df.fit_transform(X_train)

linreg2 = LinearRegression()
linreg2.fit(transform_poly,y_train)

polynomial_predict = linreg2.predict(transform_poly)


```

```

rmse = np.sqrt(mean_squared_error(y_train,polynomial_predict))
r2 = r2_score(y_train, polynomial_predict)
print("RMSE Score for Test set: " +"{:.2}".format(rmse))
print("R2 Score for Test set: " +"{:.2}".format(r2))

RMSE Score for Test set: 0.64
R2 Score for Test set: 0.95

```

```

# Extract 2 columns 'T_degC', 'Salinity' for pure and better showing
bottle_df = bottle[['T_degC', 'Depthm', 'STheta']]
# And called again
bottle_df.columns = ['Temperature', 'Dept', 'Potential_density']

+ Code + Markdown

bottle_df = bottle_df[:500]      # lets take limit for speed regression calculating
bottle_df.head()


```

	Temperature	Dept	Potential_density
0	10.50	0	25.649
1	10.46	8	25.656
2	10.46	10	25.654
3	10.45	19	25.643
4	10.45	20	25.643

```

from sklearn.preprocessing import PolynomialFeatures

poly_df = PolynomialFeatures(degree = 3)
transform_poly = poly_df.fit_transform(X_train)

linreg2 = LinearRegression()
linreg2.fit(transform_poly,y_train)

polynomial_predict = linreg2.predict(transform_poly)


```

```

rmse = np.sqrt(mean_squared_error(y_train,polynomial_predict))
r2 = r2_score(y_train, polynomial_predict)
print("RMSE Score for Test set: " +"{:.2}".format(rmse))
print("R2 Score for Test set: " +"{:.2}".format(r2))

RMSE Score for Test set: 0.44
R2 Score for Test set: 0.98

```

จากภาพจะเห็นว่า การเลือก feature ที่สังกัดกระบวนการทางกายภาพของน้ำทะเลโดยตรง เช่น ความลึกและ potential density ช่วยให้โมเดลทำงานอยุ่นหกมิน้ำทะเลได้แม่นยำขึ้นอย่างชัดเจน ซึ่งเป็นตัวอย่างของการทำ feature engineering ที่ช่วยลดค่า RMSE

6. Model Training

ฝึกโมเดลด้วยชุดข้อมูลฝึก โดยใช้เทคนิคการเรียนรู้ของเครื่องผ่านไลบรารีในภาษา Python คือ Sklearn เพื่อเรียนรู้ ความสัมพันธ์ระหว่างตัวแปรเชิงเส้นและอุณหภูมิน้ำทะเล ในรูปแบบของ linear regression และ polynomial regression

7. Model Evaluation

ประเมินประสิทธิภาพของโมเดลด้วยชุดข้อมูลทดสอบ โดยใช้ค่า Mean Squared Error (MSE) หรือ Root Mean Squared Error (RMSE) เป็นตัวชี้วัด ค่า MSE หรือ RMSE ที่ต่อแสดงถึงความสามารถของโมเดลในการทำงานอยุ่นหกมิน้ำทะเลได้อย่างแม่นยำ

สมาชิกกลุ่ม

- 1.6810422009 น.ส.สุพิชญา ขุนทรรษ
- 2.6810422012 น.ส.ปุณยนุช ตันบี
- 3.6810422021 น.ส.เพ็ชร์ พรวาทย์ไชย
- 4.6810422025 น.ส.จิริษยา สีราย