Nongjie Feng
Prof. Amir Jafari
Intro to Data Mining
06 December 2021

The Greatest Lakes Individual Final Report

Introduction

The five Great Lakes of North America are some of the most prominent and essential bodies of water on our planet. Our goal for this project is to utilize data mining techniques and algorithms to study the physical properties of the Great Lakes. We build classification model to determine whether they reach maximum ice coverage with respect to maximum ice cover average threshold. We developed a KNN model to predict which lake a scientific team is most likely examining based on observed surface temperature, ice concentration, and a given lake's physical properties. In this part, I applied linear regression and logistic regression mostly focusing on the relationship between surface temperature and ice percentage.

Algorithms

Linear regression is a very straightforward simple linear approach for predicting a quantitative response Y on the basis of a predictor variable X. We interpret coefficient as the average effect on Y of a one unit increase in variable X, holding all other predictors fixed. Linear Regression fits a linear model with coefficients beta to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.
Multiple linear regression model form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Rather than modeling this response Y directly, logistic regression models the probability that Y belongs to a particular category. In logistic regression, we use the logistic function and we use a method called maximum likelihood to fit the model. If p is a probability, then p/(1-p) is the corresponding odds; the logit of the probability is the logarithm of the odds. The logistic function will always produce an S-shaped curve. In a logistic regression model, increasing X by one unit changes the log odds by coefficient beta. Since the relationship between logistic function and X in is not a straight line, the amount that function changes due to a one-unit change in X will depend on the current value of X.
Logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

log-odds:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Likelihood function:

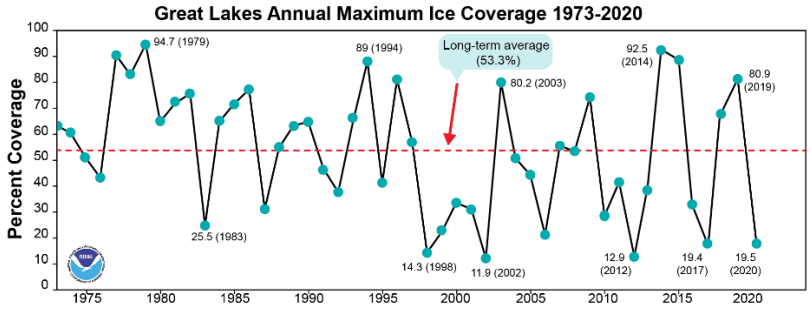
$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Experiment setup

Rows in the dataset with missing values are excluded. The Lake variable is a categorical variable, and we need to encode the labels with values between 0 and (n-1) classes: Lake Erie:0, Lake Huron:1, Lake Michigan:2, Lake Ontario:3, Lake Superior:4. We use all the variables excluding time variables year, day and id. The ice percentage is our target variable. Split the data into a training set and a testing set. We will train our model on the training set and then use the test set to evaluate the model. Rescale input variables using standardization or normalization will make more reliable predictions. Set parameter 'normalize' true and the regressors X will be normalized before regression by subtracting the mean and dividing by the L2-norm. There are three common evaluation metrics for regression problems. Mean Absolute Error (MAE) is the mean of the absolute value of the errors; Mean Squared Error (MSE) is the mean of the squared errors and is more popular than MAE, because MSE punishes larger errors, which tends to be useful in the real world; Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors and is also popular because RMSE is interpretable in the Y units.

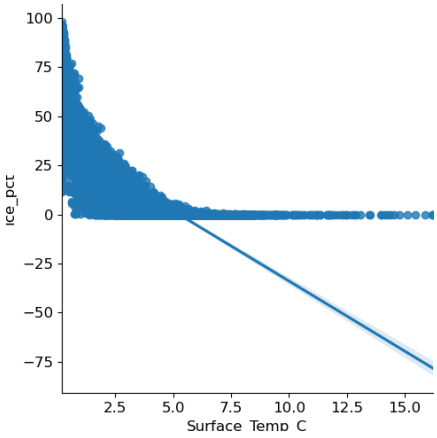
NOAA scientists project the long-term maximum Great Lakes ice cover average for 1973 to 2020 is 53.3%. We will be building classification model to determine whether they reach maximum ice coverage with respect to maximum ice cover average threshold. In order to map this probability value to a discrete class (true/false), we select a threshold value of 0.533. Mathematically, it can be expressed as: When p larger than 0.533 and class is 1, they reach maximum ice coverage; When p less than 0.533 and class is 0, they do not reach maximum ice coverage. We change the categorical variable Lake to a dummy variable in the logit model. 'MinMaxScaler' scales all the data features in the range [0, 1]. Next, we train the logistic model, we use default 'C' equal to 1 and threshold 0.533. Then, we can increase C to 100 and fit a more flexible model. We can also use more regularized model than the default value of C=1, by setting C=0.01. We use the 'predict_proba' method to predict and it gives the probabilities for the target variable (0 and 1). We use confusion matrix for summarizing the performance of a classification algorithm. Four types of outcomes are possible while evaluating a classification model performance. True Positives (TP): True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class. True Negatives (TN): True Negatives occur when we predict an observation does not belong to a certain class and the observation actually does not belong to that class. False Positives (FP): False Positives occur when we predict an observation belongs to a certain class but the observation actually does not belong to that class. This type of error is called Type I error. False Negatives (FN): False Negatives occur when we predict an observation does not belong to a certain class but the observation actually belongs to that class. This is a very serious error and it is called Type II error. Classification report is another way to evaluate the classification model performance. It displays the precision, recall, f1 and support scores for the model. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in actual class. F1 Score is the weighted average of Precision and Recall. Another tool to measure the classification model performance visually is Receiver Operating Characteristic (ROC) Curve. The ROC Curve plots the True Positive Rate (TPR) against the False

Positive Rate (FPR) at various levels. The Geometric Mean or G-Mean is a metric for imbalanced classification that, if optimized, will seek a balance between the sensitivity and the specificity. One approach would be to test the model with each threshold returned from the call 'roc_auc_score' and select the threshold with the largest G-Mean value. Given that we have already calculated the Sensitivity (TPR) and the complement to the Specificity when we calculated the ROC Curve, we can calculate the G-Mean for each threshold directly. ROC-AUC is a single number summary of classifier performance. The higher the value, the better the classifier. In addition, we can use tkinter to build a GUI to calculate different test-set accuracy for different hyperparameter C.



Results

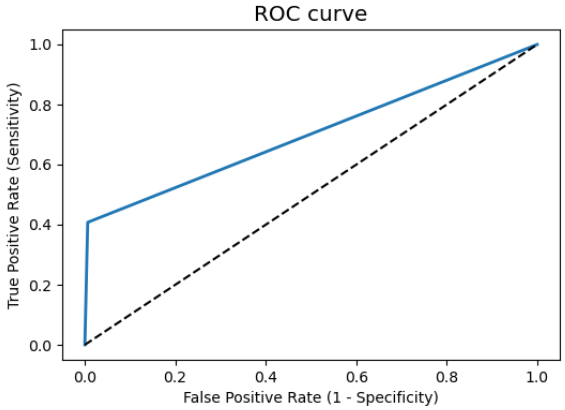
From the line plot between surface temperature and ice percentage, we can see that there is a negative relationship between two variables. The lower the temperature, the greater the percentage of ice. The MSE of test and train set is around 320 and the R-square is around 0.44. The intercept of the model is 15, which means when the surface temperature is 0, there is 15 percent of ice cover. Holding all other features fixed, a 1 unit decrease in surface temperature. Ice percentage is associated with an increase of 15.

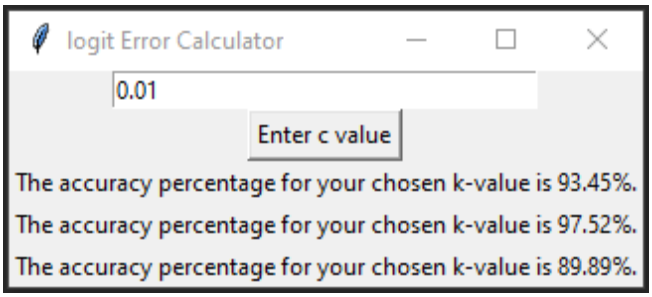


	Coefficient
Elevation_meters	1.531980e+14
Length_km	-2.505699e+13
Breadth_km	8.195978e+13
Avg_Depth_meters	2.239198e+13
Max_Depth_meters	-5.039058e+13
Volume_km3	-2.284854e+14
Water_Area_km2	2.197783e+14
Land_Drain_Area_km2	-7.097686e+13
Total_Area_km2	-1.567553e+14
Shore_Length_km	-1.610179e+12
Retention_Time_years	-3.299119e+13
Lake	2.036360e+14
Surface_Temp_C	-1.574068e+01

```
Test set evaluation:
-----------------------------------
MAE: 13.185771710850881
MSE: 324.37812761736313
RMSE: 18.010500482145495
R2 Square 0.4365849176327412
# ---------------------------------
Train set evaluation:
-----------------------------------
MAE: 13.127703228346917
MSE: 323.97822626116726
RMSE: 17.999395163759456
R2 Square 0.45196139369443256
```

The logistic regression model accuracy score is 0.9345. So, the model does a very good job in predicting whether or not it will reach maximum ice coverage in the Great Lakes. The training-set accuracy score is 0.9365 while the test-set accuracy is 0.9345. These two values are quite comparable. So, there is no question of overfitting. We can see that, C=100 results in higher accuracy. Training-set accuracy score is 0.9771 and test-set accuracy is 0.9752. So, we can conclude that a more complex model should perform better. When C=0.01, Training-set accuracy score is 0.8964 and test-set accuracy is 0.8989. Both the training and test set accuracy decreased. We can see that our model accuracy score is 0.9345 but null accuracy score is 0.9076. So, we can conclude that our Logistic Regression model is doing a very good job in predicting the class labels. True Positives(TP) is 1582; True Negatives(TN) is 73; False Positives(FP) is 10; False Negatives(FN) is 106. Precision is 0.9937 and it identifies the proportion of correctly predicted positive outcomes. It is more concerned with the positive class than the negative class.; Sensitivity is 0.9372 and it identifies the proportion of correctly predicted actual positives; Specificity is 0.8795. ROC-AUC of our model is 0.7008 and we can conclude that our classifier does a fair job in predicting whether it will reach maximum ice coverage or not.





Summary

From the result, a small number of observations predict that there will be maximum ice coverage. Majority of observations predict that there will be no maximum ice coverage. Logistic regression is a process of modeling the probability of a discrete outcome given input variables. Usually, we use logistic regression as a binary outcome, which fits our project very well. Also, logistic regression is considered safe and robust, according to what we learnt from class. The ROC-AUC is not close to 1, the G-Mean method suggest that set threshold near to 1 may get better results.

The linear regression has a fair R-square score. From the plot in linear regression result, the dataset contains a lot of data points that surface temperature is larger than 5. However, the minimum ice percentage is 0. These points affect the overall performance of the linear model. We can delete these observations, but it will lose a lot of data. We may also try some other models like lasso regression and elastic net.

Dealing with missing data is the most common and important problem in modeling. If we have more time, we can try different ways to deal with missing data. For example, the EM algorithm is an alternative to Newton–Raphson or the method of scoring for computing MLE in cases where the complications in calculating the MLE are due to incomplete observation and data are MAR, missing at random, with separate parameters for observation and the missing data mechanism, so the missing data mechanism can be ignored.

Percentage of the code that you found or copied from the internet: (120-18)/(120+76)*100 = 52%

Works Cited

scikit-learn. 2021. *sklearn.linear_model.LinearRegression*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> [Accessed 6 December 2021].

scikit-learn. 2021. *sklearn.linear_model.LinearRegression*. [online] Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> [Accessed 6 December 2021].

En.wikipedia.org. 2021. *Logit - Wikipedia*. [online] Available at: https://en.wikipedia.org/wiki/Logit [Accessed 6 December 2021].

Hastie, T., Tibshirani, R. and Friedman, J., n.d. *The elements of statistical learning*.

James, G., Witten, D., Hastie, T. and Tibshirani, R., n.d. *An introduction to statistical learning*.

Kaggle.com. 2021. *Linear Regression House price prediction*. [online] Available at: <https://www.kaggle.com/faressayah/linear-regression-house-price-prediction> [Accessed 6 December 2021].

Kaggle.com. 2021. *Logistic Regression Classifier Tutorial*. [online] Available at: <https://www.kaggle.com/prashant111/logistic-regression-classifier-tutorial> [Accessed 6 December 2021].

Kaggle.com. 2021. *Heart Disease Prediction using Logistic Regression*. [online] Available at: <https://www.kaggle.com/neisha/heart-disease-prediction-using-logistic-regression> [Accessed 6 December 2021].

pandas?, A., Troy, A. and Zwinck, J., 2021. *Any way to get mappings of a label encoder in Python pandas?*. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/42196589/any-way-to-get-mappings-of-a-label-encoder-in-python-pandas> [Accessed 6 December 2021].

Research.noaa.gov. 2021. *NOAA projects 30-percent maximum Great Lakes ice cover for 2021 winter - Welcome to NOAA Research*. [online] Available at: <https://research.noaa.gov/article/ArtMID/587/ArticleID/2706/NOAA-projects-30-percent-average-Great-Lakes-ice-cover-for-2021-winter> [Accessed 6 December 2021].

Brownlee, J., 2021. *A Gentle Introduction to Threshold-Moving for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/> [Accessed 6 December 2021].