hw2_solutions

January 18, 2024

Problem 1 Suppose the true model is

$$y = 1(\beta x - \varepsilon \geq 0),$$

where $1(\cdot)$ is the indicator function, ε has the logistic distribution with CDF $F(\varepsilon) = \left(1 + e^{-\varepsilon}\right)^{-1}$, x is a binary variable taking values in $\{-1, 1\}$ with $P(x = 1) = p \in (0, 1)$, ε is independent of x, and $\beta > 0$.

(a) Show that the log odds ratio $\log \frac{P(y=1|x)}{P(y=0|x)}$ is linear in β, and therefore the true model for y is a logistic regression.

(b) Suppose you have access to β and therefore can use the true logistic classifier

$$f(x) = 1\left[\log \frac{P(y = 1|x)}{P(y = 0|x)} \geq 0\right].$$

Calculate its 0-1 risk

$$R_{01}(f) = \mathbb{E}1\left[y \neq f(x)\right].$$

(c) Suppose now you use a dubious classifier that just assigns labels randomly (without even looking at x). Show that this classifier is never better than the true classifier in terms of the 0-1 risk.

Solution (a) We have

$$\begin{aligned} P(y = 1|x) &= P(\varepsilon \leq \beta x|x) = \left(1 + e^{-\beta x}\right)^{-1} = \frac{e^{\beta x}}{1 + e^{\beta x}}, \\ P(y = 0|x) &= \frac{1}{1 + e^{\beta x}}. \end{aligned}$$

Therefore, the log odds ratio is

$$\log \frac{P(y = 1|x)}{P(y = 0|x)} = \log e^{\beta x} = \beta x.$$

(b) We have

$$R_{01}(f) = P(y \neq f(x)) = P(y = 1 \text{ and } f(x) = 0) + P(y = 0 \text{ and } f(x) = 1).$$

Notice that $f(1) = 1$ and $f(-1) = 0$. Then

$$P(y = 1 \text{ and } f(x) = 0) = P(\varepsilon \leq \beta x \text{ and } x = -1) = P(\varepsilon \leq -\beta)P(x = -1) = (1 - p) \cdot \frac{1}{1 + e^{\beta}}.$$

Similarly,

$$P(y = 0 \text{ and } f(x) = 1) = P(\varepsilon > \beta)P(x = 1) = p \cdot \frac{1}{1 + e^{\beta}}.$$

Therefore, the 0-1 risk is

$$R_{01}(f) = \frac{1}{1 + e^{\beta}}.$$

(c) The dubious classifier has the 0-1 risk equal to 0.5 (i.e. 50% chance of misclassification). Clearly, $R_{01}(f) \leq 0.5$ since $\beta > 0$.

Problem 2 Suppose in the testing sample there is an equal number of positive- and negative-labeled observations.

(a) For a trivial classifier that always predicts "positive' ', what is the value of accuracy? precision? recall? F1 score?

(b) For a trivial classifier that always predicts "negative' ', what is the value of accuracy? precision? recall? F1 score? (Some of these may be undefined).

(c) For a trivial classifier that assigns labels at random, what is the value of accuracy? precision? recall? F1 score? (Assume that in the testing sample, the classifier misclassified exactly 50% of positive-labeled observations and 50% of negative-labeled observations).

Solution (a) Accuracy is 0.5, precision is 0.5, recall is 1, F1 score is 2/3.

(b) Accuracy is 0.5, precision is undefined (0/0), recall is 0, F1 score is undefined.

(c) Accuracy is 0.5, precision is 0.5, recall is 0.5, F1 score is 0.5.

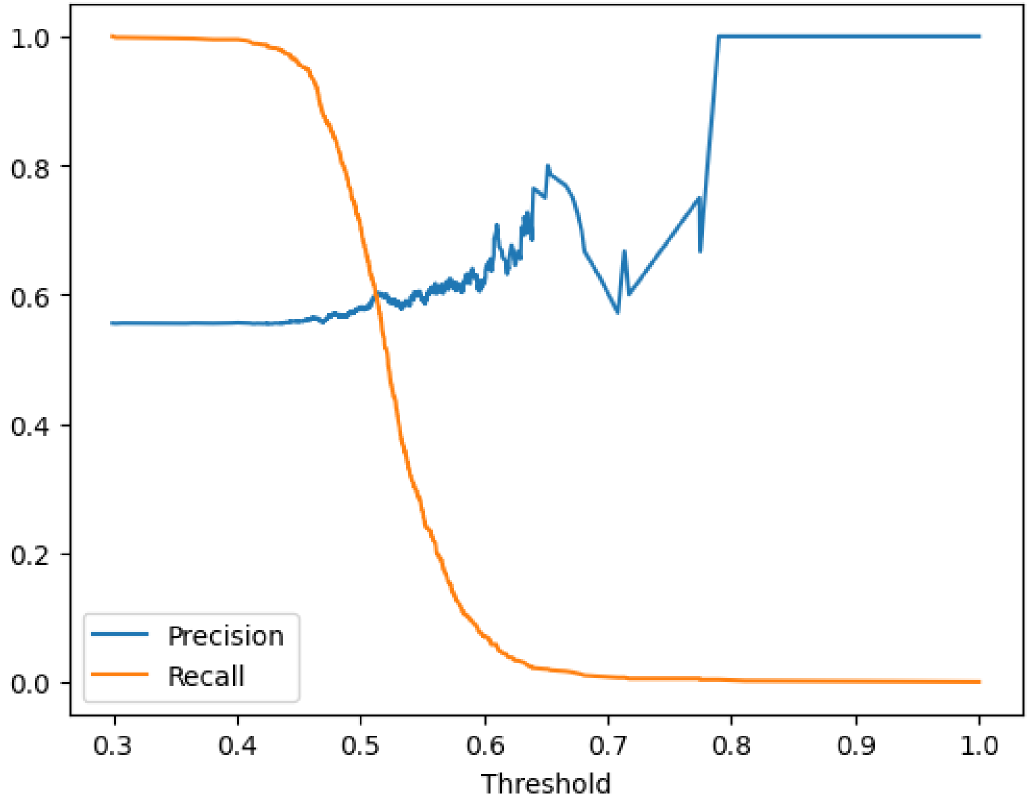
Problem 3 This problem uses the Weekly.csv dataset (uploaded on Bruinlearn) containing 1089 weekly stock returns for 21 years. (a) Use the full dataset to fit a logistic regression of today's stock movement (up or down) on the five lags of returns and the trading volume. (b) Calculate the confusion matrix, accuracy, precision, recall, and F1 score for the in-sample predictions. Does the model uniformly beat random guessing in terms of these performance metrics? (c) On the same graph, plot precision and recall against the threshold (varying over $[0, 1]$) used to generate predicted labels from predicted probabilities. Explain the pattern you see. (d) Now fit the logistic regression using only data up to (and including) the year 2008, with Lag2 as the only predictor. (e) Use the remaining observations as a test sample to repeat (b) and (c). (f) Which of the two fitted models would you use for real-time stock return prediction?

[42]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.discrete.discrete_model import Logit
from sklearn.metrics import confusion_matrix, precision_score, accuracy_score, f1_score, recall_score, precision_recall_curve


datapath = '/Users/franguri/Library/CloudStorage/Dropbox/_TEACHING_/Econ 425 ML UCLA Winter 2024/WEEK 2 Logistic regression/Weekly.csv'
```

```
data = pd.read_csv(datapath, header=0)
data['Direction_num'] = data['Direction'] == 'Up'
#data['Direction_num'] = pd.to_numeric(data['Direction_num'])
data.describe()

# logit regression with full data:
X_train = data[['Lag1', 'Lag2', 'Lag3', 'Lag4', 'Lag5', 'Volume']]
Y_train = data['Direction_num']
model = Logit(Y_train, X_train).fit()
print(model.summary())

Y_pred_raw = model.predict(X_train)
Y_pred_label = Y_pred_raw >= 0.5
confusion_matrix(Y_train, Y_pred_label)
print('Precision = ', precision_score(Y_train, Y_pred_label))
print('Recall = ', recall_score(Y_train, Y_pred_label))
print('F1-score = ', f1_score(Y_train, Y_pred_label))
print('Accuracy = ', accuracy_score(Y_train, Y_pred_label))
prec, rec, thresholds = precision_recall_curve(Y_train, Y_pred_raw)

plt.plot(np.append(thresholds,1), prec, label='Precision')
plt.plot(np.append(thresholds,1), rec, label='Recall')
plt.xlabel('Threshold')
plt.legend()
plt.show()


# logit regression with 1990-2008 data and Lag2 only:
data_08 = data[(data['Year'] >= 1990) & (data['Year'] <= 2008)]
Y_train_08 = data_08['Direction_num']
X_train_08 = data_08['Lag2']
model = Logit(Y_train_08, X_train_08).fit()
print(model.summary())

data_test = data[(data['Year'] > 2008)]
X_test = data_test['Lag2']
Y_test = data_test['Direction_num']

Y_test_pred_raw = model.predict(X_test)
Y_test_pred_label = Y_test_pred_raw >= 0.5
confusion_matrix(Y_test, Y_test_pred_label)
print('Precision = ', precision_score(Y_test, Y_test_pred_label))
print('Recall = ', recall_score(Y_test, Y_test_pred_label))
print('F1-score = ', f1_score(Y_test, Y_test_pred_label))
print('Accuracy = ', accuracy_score(Y_test, Y_test_pred_label))
```

```
Optimization terminated successfully.
         Current function value: 0.686896
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:           Direction_num   No. Observations:                 1089
Model:                           Logit   Df Residuals:                     1083
Method:                            MLE   Df Model:                            5
Date:                 Thu, 18 Jan 2024   Pseudo R-squ.:                9.505e-05
Time:                         16:43:40   Log-Likelihood:                -748.03
converged:                        True   LL-Null:                       -748.10
Covariance Type:             nonrobust   LLR p-value:                    0.9996
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Lag1          -0.0327      0.026     -1.250      0.211      -0.084       0.019
Lag2           0.0682      0.027      2.556      0.011       0.016       0.120
Lag3          -0.0081      0.026     -0.306      0.759      -0.060       0.044
Lag4          -0.0194      0.026     -0.740      0.459      -0.071       0.032
Lag5          -0.0069      0.026     -0.261      0.794      -0.058       0.045
Volume         0.0569      0.027      2.125      0.034       0.004       0.109
==============================================================================
Precision =  0.5778078484438431
Recall =  0.7057851239669422
F1-score =  0.6354166666666666
Accuracy =  0.5500459136822773
```



```
Optimization terminated successfully.
         Current function value: 0.690654
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:          Direction_num   No. Observations:                  985
Model:                          Logit   Df Residuals:                      984
Method:                           MLE   Df Model:                            0
Date:                Thu, 18 Jan 2024   Pseudo R-squ.:               -0.004340
Time:                        16:43:40   Log-Likelihood:                -680.29
converged:                       True   LL-Null:                       -677.35
Covariance Type:            nonrobust   LLR p-value:                       nan
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Lag2           0.0629      0.029      2.192      0.028       0.007       0.119
==============================================================================
Precision =  0.6166666666666667
Recall =  0.6065573770491803
F1-score =  0.6115702479338844
Accuracy =  0.5480769230769231
```