

Intermediate Data Science

Linear and Logistic Regression

Joanna Bieri DATA201

Mini Review - Linear and Logistic Regression

REVIEW

- **Linear Regression** - for predicting *continuous* outcomes (e.g., prices, scores). It is a true regression model where we are trying to fit a straight line to some data!
- **Logistic Regression** - for predicting *categorical* outcomes (e.g., yes/no, 0/1). This is a classification model! A way to take the idea of linear regression and turn it into classification.

Mini Review - Linear Regression

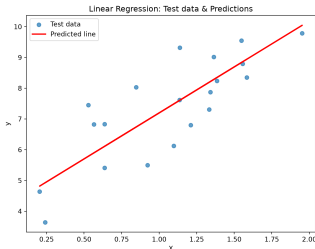
REVIEW

$\hat{y} = w_0 + w_1x$ The “machine learning” is figuring out what these constants should be!

```
linreg = LinearRegression()  
linreg.fit(X_train, y_train)
```

Intercept (w_0): 4.206

Coefficient (w_1): 2.990



Mini Review - MSE and R^2

REVIEW

- **Mean Squared Error (MSE)** measures the **average squared difference** between predicted values and actual values. Smaller MSE means predictions are closer to actual values.
- **Coefficient of Determination (R^2)** measures how much of the **variance in the target** is explained by the model.

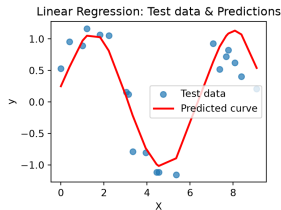
```
y_pred = linreg.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

MSE on test set: 0.918

R^2 on test set: 0.652

Mini Review - What if our data was nonlinear?

REVIEW



Left: a straight line. Right: a 6-degree polynomial. Polynomial regression includes **powers of predictors**: $\hat{y} = w_0 + w_1x + w_2x^2 + \dots$

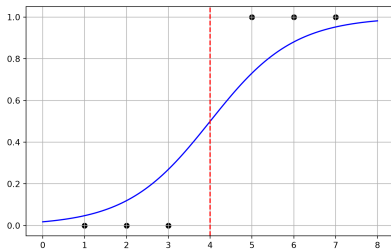
```
poly = PolynomialFeatures(degree=degree,  
                          include_bias=False)
```

```
X_train_poly = poly.fit_transform(X_train)
```

```
X_test_poly = poly.transform(X_test)
```

Mini Review - Logistic Regression

REVIEW



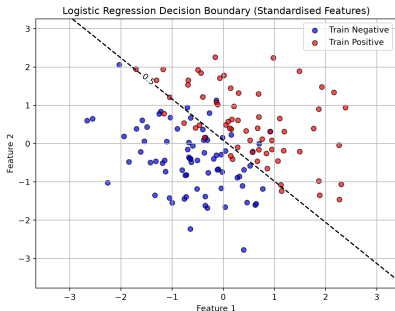
$$p = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

The “sigmoid” function maps the linear part back into the interval $(0, 1)$.

If the probability is less than 0.5 then it belongs to class 0 and otherwise to class 1.

Mini Review - Understanding Classification Metrics

REVIEW



Accuracy on test set: 0.850

Confusion Matrix:

$\begin{bmatrix} 28 & 6 \end{bmatrix}$

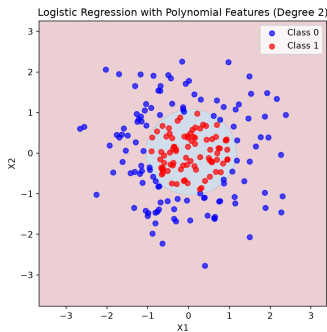
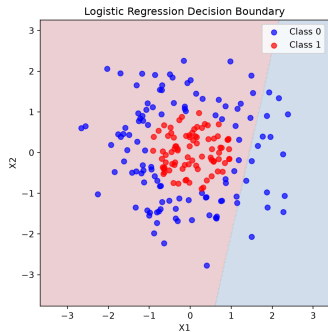
$\begin{bmatrix} 3 & 23 \end{bmatrix}$

TN	FP
FN	TP

- **Precision** measures the accuracy of positive predictions. High precision means few false positives.
- **Recall** measures how well the model finds all actual positives. High recall means few false negatives.

Mini Review - What if our decision boundary was not linear?

REVIEW



Accuracy: 0.517

polynomial features: 0.917

We can use polynomial features here too! We still have over vs underfitting!

Which Model?

ASK

You want to predict

(1) the price of a house

(2) if an email is spam.

Which model for each? Why?

Hint. Is the thing you are predicting a number or a category?

Which Model?

ANSWER

*You want to predict (1) the price of a house and (2) if an email is spam.
Which model for each? Why?*

- House price is a number, so Linear Regression. Measure it with MSE and R^2 .
- Spam or not spam is a category, so Logistic Regression. Measure it with accuracy, the confusion matrix, precision and recall.
- The name is confusing: Logistic Regression is a classification model.

Linear Regression

ASK

We made the data with $y = 4 + 3x$ plus noise.

The model found 4.21 and 2.99.

Why not exactly 4 and 3?

Hint. The model never sees the formula. What does it see?

Linear Regression

ANSWER

We made the data with $y = 4 + 3x$ plus noise. The model found 4.21 and 2.99. Why not exactly 4 and 3?

- The model only sees 80 noisy training points, not the formula.
- It finds the line that fits THOSE points best. A different random split gives slightly different numbers.
- More data, or less noise, gets you closer to 4 and 3.

Linear Regression

ASK

A model has $R^2 = 0.005$
on the test set.

What does that tell you?

Hint. What would a model that always predicts the average score?

Linear Regression

ANSWER

A model has $R^2 = 0.005$ on the test set. What does that tell you?

- It is doing about as well as always predicting the average of y . $R^2 = 0$ is exactly that.
- The model is missing the pattern in the data. In the notes this was a straight line on a sine wave.
- Plot the data and the predictions, then try a different model.

Over and Underfitting

ASK

A polynomial is a curve.
Why is polynomial regression
still a LINEAR model?

Hint. Look at the w s in $\hat{y} = w_0 + w_1x + w_2x^2$. What are we solving for?

Over and Underfitting

ANSWER

A polynomial is a curve. Why is polynomial regression still a LINEAR model?

- The computer is solving for the w 's, and each w is only multiplied by a number and added up.
- `PolynomialFeatures` makes x^2 a new column. After that it is the same `LinearRegression()`.
- Linear means linear in the coefficients, not a straight line in x .

Over and Underfitting

ASK

R^2 on the training data: 0.99

R^2 on the test data: 0.40

What is going on?

What do you try?

Hint. Think about memorizing the practice exam.

Over and Underfitting

ANSWER

R^2 on the training data is 0.99 and on the test data it is 0.40. What is going on? What do you try?

- Overfitting. The model fits the training points very closely and does not carry over to new data.
- Try a simpler model (a lower degree), or get more data.
- Underfitting is the other problem: both scores are low because the model is too simple.

Logistic Regression

PREDICT

confusion matrix

	predicted 0	predicted 1
real 0	8	2
real 1	1	9

The model said "class 1" for 11 things. How many of those are right? Which metric is that?

Hint. Read down the "predicted 1" column.

Logistic Regression

ANSWER

The model said "class 1" for 11 things. How many of those are right? Which metric is that?

9 of 11 = 0.82 precision

Precision is $TP / (TP + FP) = 9 / (9 + 2)$. Recall reads across the "real 1" row: $9 / (9 + 1) = 0.90$. Accuracy is everything on the diagonal: $17 / 20 = 0.85$.

Logistic Regression

ASK

A model predicts if a patient
has a serious disease.

Do you move the cutoff
below 0.5 or above 0.5? Why?

Hint. Which mistake is worse here, a false positive or a false negative?

Logistic Regression

ANSWER

A model predicts if a patient has a serious disease. Do you move the cutoff below 0.5 or above 0.5? Why?

- Below 0.5. A false negative (missing a sick patient) is the worse mistake.
- A lower cutoff calls more people positive, so recall goes up and precision goes down.
- The cutoff is a choice, and it depends on what each kind of mistake costs.

Data Science Interview Practice

INTERVIEW

Explain the linear regression model.

What does it assume?

Data Science Interview Practice

INTERVIEW

Explain logistic regression
and give an example
of when you would use it.

Questions?

Questions?