Multiple Logistic Regression

Pavlos Protopapas

Lecture Outline

Outline:

- How do we interpret logistic regression coefficients when the features are categorical?
- How do we perform inference in logistic regression?
- How do we build logistic regression models with multiple features?
- Visualizing a 2-dimensional linear decision boundary
- How do we fit a multi-logistic-regression model?
- How do we interpret multiple logistic regression coefficients?
- Visualizing a 2-dimensional polynomial decision boundary

OLD OUTLINE:

- What is classification
- Classification: Why not Linear Regression?
- Classification using the Logistic Model
- Binary Response & Logistic Regression
- Estimating the Simple Logistic Model
- Inference in Logistic Regression
- Multiple Logistic Regression
- Polynomial Logistic Model

Categorical Predictors

Just like in linear regression, when the predictor, X, is binary, the interpretation of the model simplifies.

In this case, what are the interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$?

For the heart data, let X be the indicator that the individual is a male or female. What is the interpretation of the coefficient estimates in this case?

The observed percentage of HD for women is 26% while it is 55% for men.

Calculate the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ if the indicator for HD was predicted from the gender indicator.

Statistical Inference in Logistic Regression

The **uncertainty of the estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$ can be quantified and used to calculate both confidence intervals and hypothesis tests.

Of course, you can use **bootstrap** to perform these inferences.

Note:

The estimate for the standard errors of these estimates without bootstrap, is based on a quantity called Fisher's Information (beyond the scope of this class), which is related to the curvature of the log-likelihood function.

Due to the nature of the underlying Bernoulli distribution, if you estimate the underlying proportion p_i, you get the variance for free! Because of this, the inferences will be based on the normal approximation (and not t-distribution based).

Multiple Logistic Regression

Multiple Logistic Regression

It is simple to illustrate examples in logistic regression when there is just one predictors variable.

But the approach ‘easily’ generalizes to the situation where there are multiple predictors.

A lot of the same details as linear regression apply to logistic regression. Interactions can be considered, multicollinearity is a concern and so is overfitting.

So how do we correct for such problems?

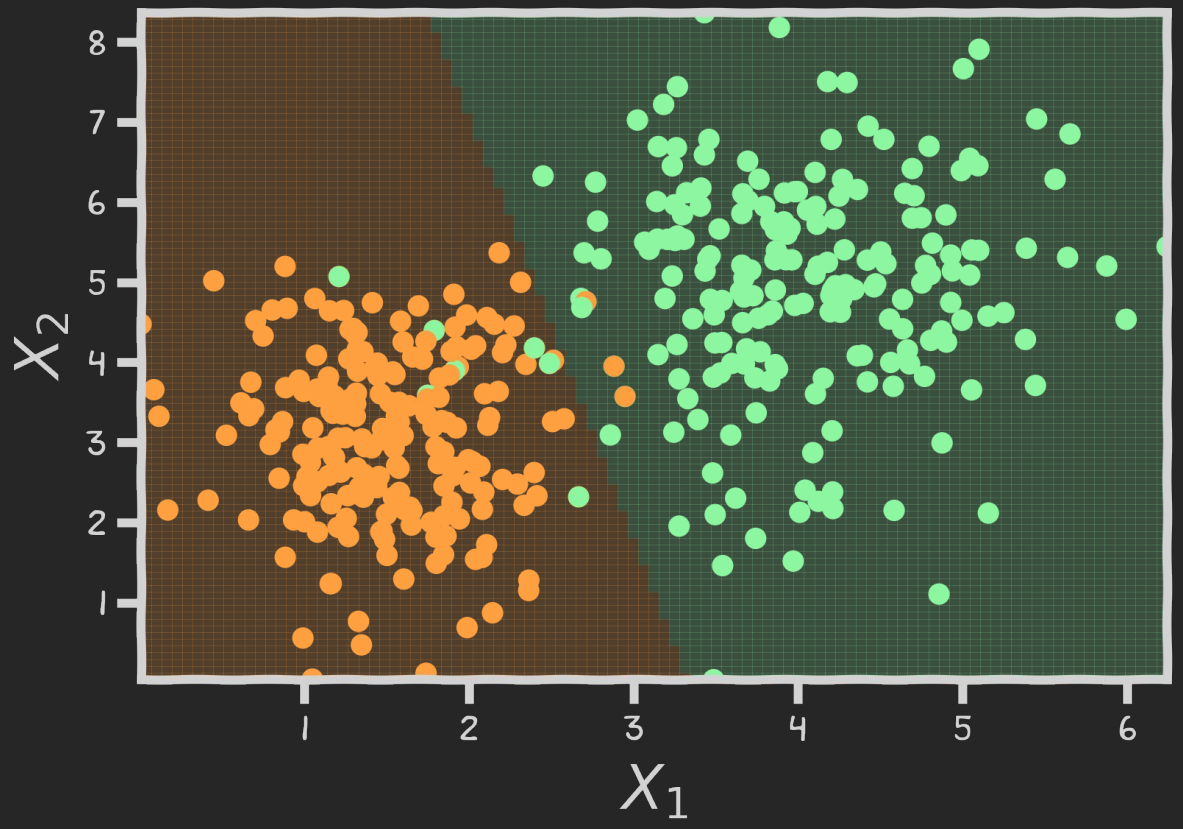
Regularization and checking though train and cross-validation!

We will get into the details of this, along with other extensions of logistic regression, in the next lecture.

Classifier with two predictors

How can we estimate a classifier, based on logistic regression, for the following plot?

The challenge here is that we have



Multiple logistic regression is a generalization to multiple predictors:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Decision Boundaries

Multiple Logistic Regression

Earlier we saw the general form of *simple* logistic regression, meaning when there is just one predictor used in the model:

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

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Fitting Multiple Logistic Regression

The estimation procedure is identical to that as before for simple logistic regression:

- a likelihood approach is taken, and the –ve log-likelihood is minimized across all parameters $\beta_0, \beta_1, \ldots, \beta_p$ using an iterative method like Gradient Descent.

Interpretation of Multiple Logistic Regression

Interpreting the coefficients in a multiple logistic regression is similar to that of linear regression.

Key: since there are other predictors in the model, the coefficient $\hat{\beta}_j$ is the association between the j^{th} predictor and the response (on log odds scale).

Interpreting Multiple Logistic Regression: an Example

Let's get back to the Heart Data. We are attempting to predict whether someone has HD based on MaxHR and whether the person is female or male. The simultaneous effect of these two predictors can be brought into one model.

Recall from earlier we had the following estimated models:

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = 6.30 - 0.043 \cdot X_{MaxHR}$$

$$\log\left(\frac{\widehat{P(Y=1)}}{1-\widehat{P(Y=1)}}\right) = -1.06 + 1.27 \cdot X_{gender}$$

Some questions

1. Estimate the odds ratio of HD comparing men to women using this model.
2. Is there any evidence of multicollinearity in this model?
3. Is there any confounding in this problem?

Interactions in Multiple Logistic Regression

Just like in linear regression, interaction terms can be considered in logistic regression. An **interaction terms** is incorporated into the model the same way, and the interpretation is very similar (on the log-odds scale of the response of course).

Write down the model for the Heart data for the 2 predictors plus the interactions term.

Some questions

1. Write down the complete model. Break this down into the model to predict log-odds of heart disease (HD) based on MaxHR for women and the same model for men. How is this different from the previous model (without interaction)?

2. Interpret the results of this model. What does the coefficient for the interaction term represent?

3. Estimate the odds ratio of HD comparing men to women using this model [trick question].

4. Is there any evidence of multicollinearity in this model?

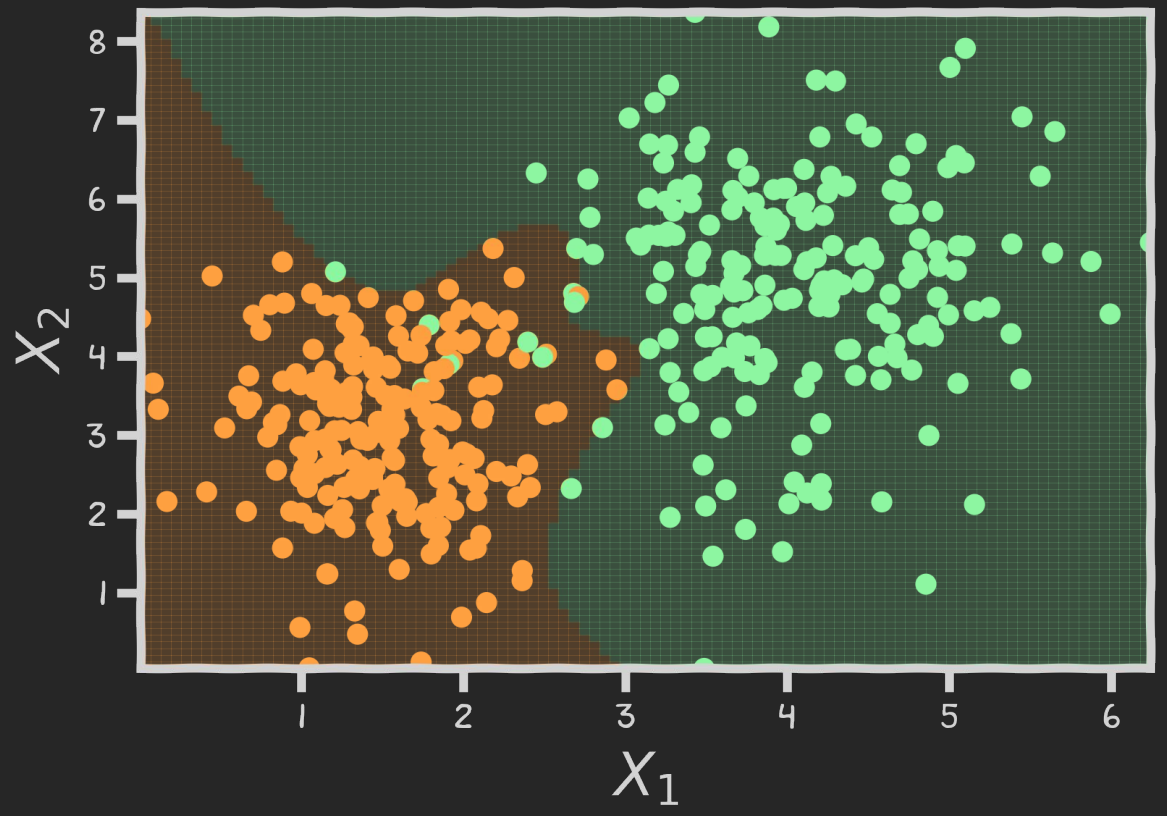
5. Is there any confounding in this problem?

Polynomial Logistic Regression

Polynomial Logistic Regression

We saw a 2-D plot last time which had two predictors, X_1, X_2. A similar one is shown here but the decision boundary is not linear.

We can extend multiple Logistic Regression as we did with polynomial regression:

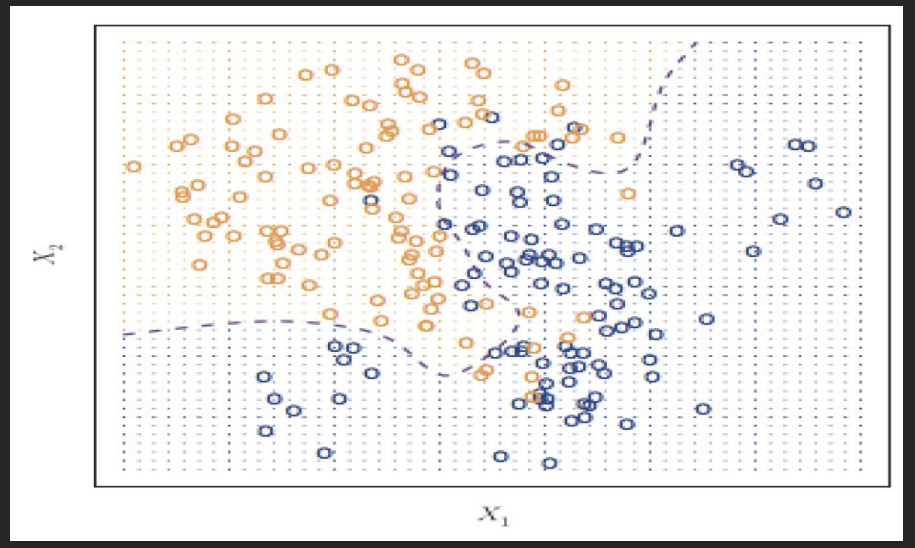


$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \tilde{X}\beta$$

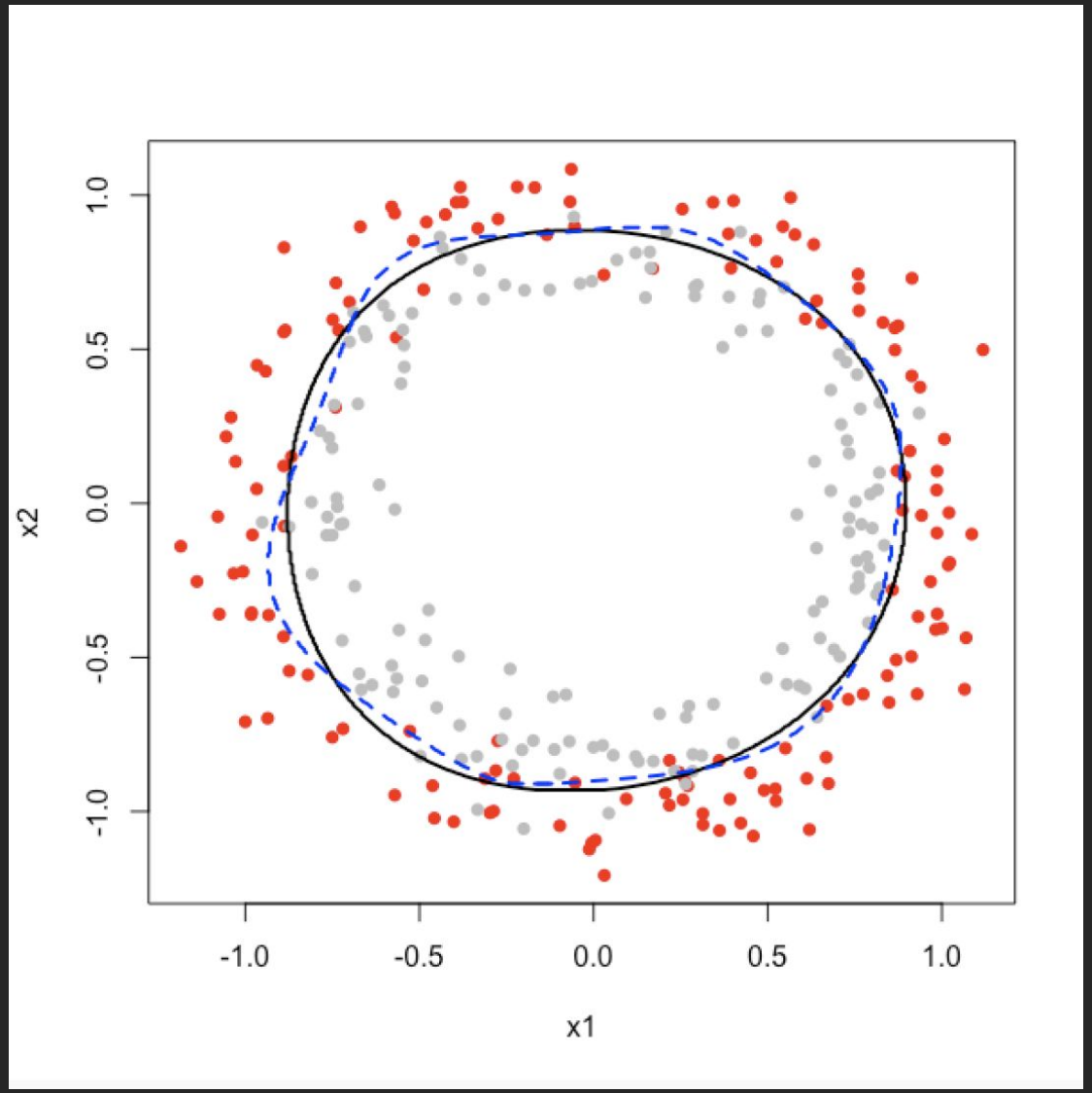
Classification boundaries

Recall that we could attempt to purely classify each observation based on whether the estimated $P(Y = 1)$ from the model was greater than 0.5.

When dealing with 'well-separated' data, logistic regression can work well in performing classification.



2D Classification in Logistic Regression: an Example



2D Classification in Logistic Regression: an Example

Would a logistic regression model perform well in classifying the observations in this example?

What would be a good logistic regression model to classify these points?

Based on these predictors, two separate logistic regression models were

2D Classification in Logistic Regression: an Example

In the previous plot, which classification boundary performs better? How can you tell? How would you make this determination in an actual data example?

We could determine the misclassification rates in left out validation or test set(s)