Multiple Logistic Regression

Pavlos Protopapas, Ignacio Becker

Lecture 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.

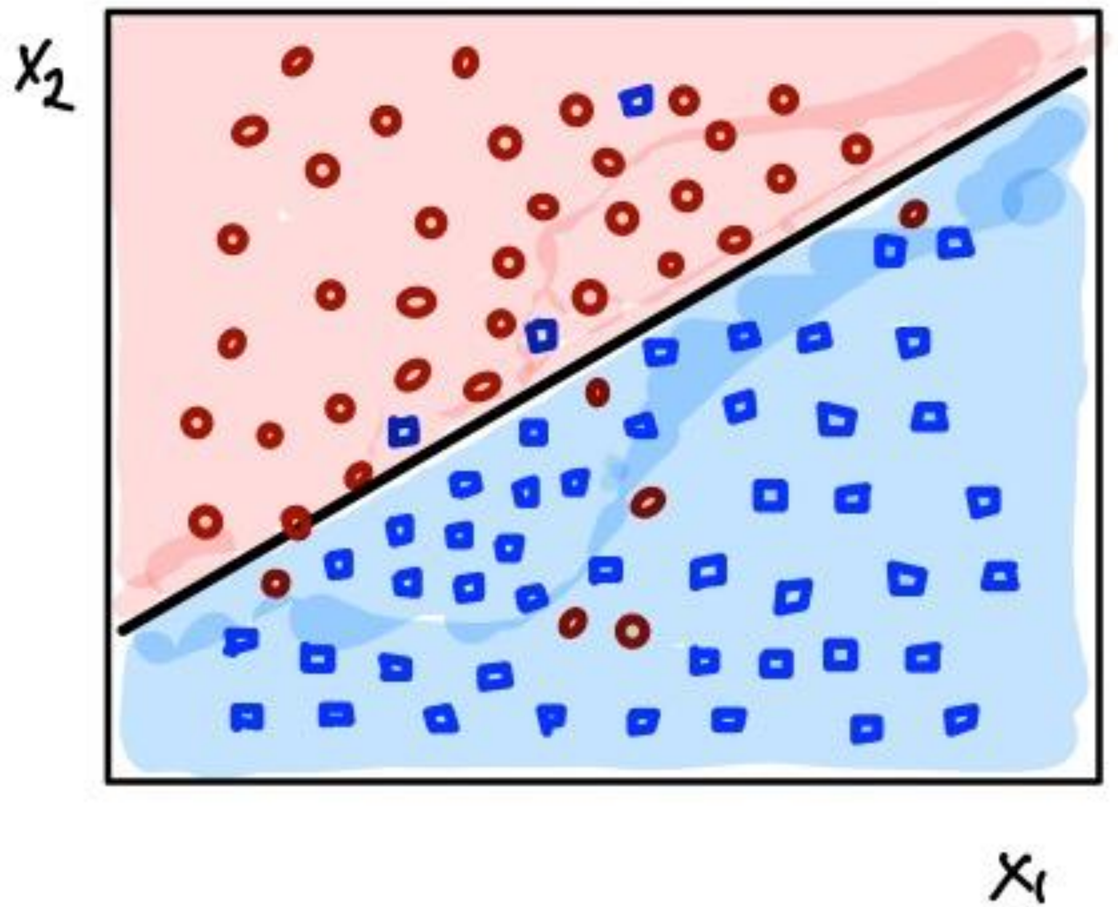
So how do we correct for such problems?

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 predictors, $x_1, x_2, \dots, x_p$.

Multiple logistic regression is a generalization to multiple predictors:

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

Fitting Multiple Logistic Regression

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

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

The actual fitting of a Multiple Logistic Regression is easy using software (of course there's a python package for that) as the iterative minimization of the -ve log-likelihood has already been hard coded.

In the `sklearn.linear_model` package, you just have to create your multidimensional design matrix X to be used as predictors in the `LogisticRegression` function.

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). But do we have to say, “Controlling for the other predictors in the model”?

We are trying to attribute the partial effects of each model controlling for the others (aka, controlling for possible *confounders*).

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.

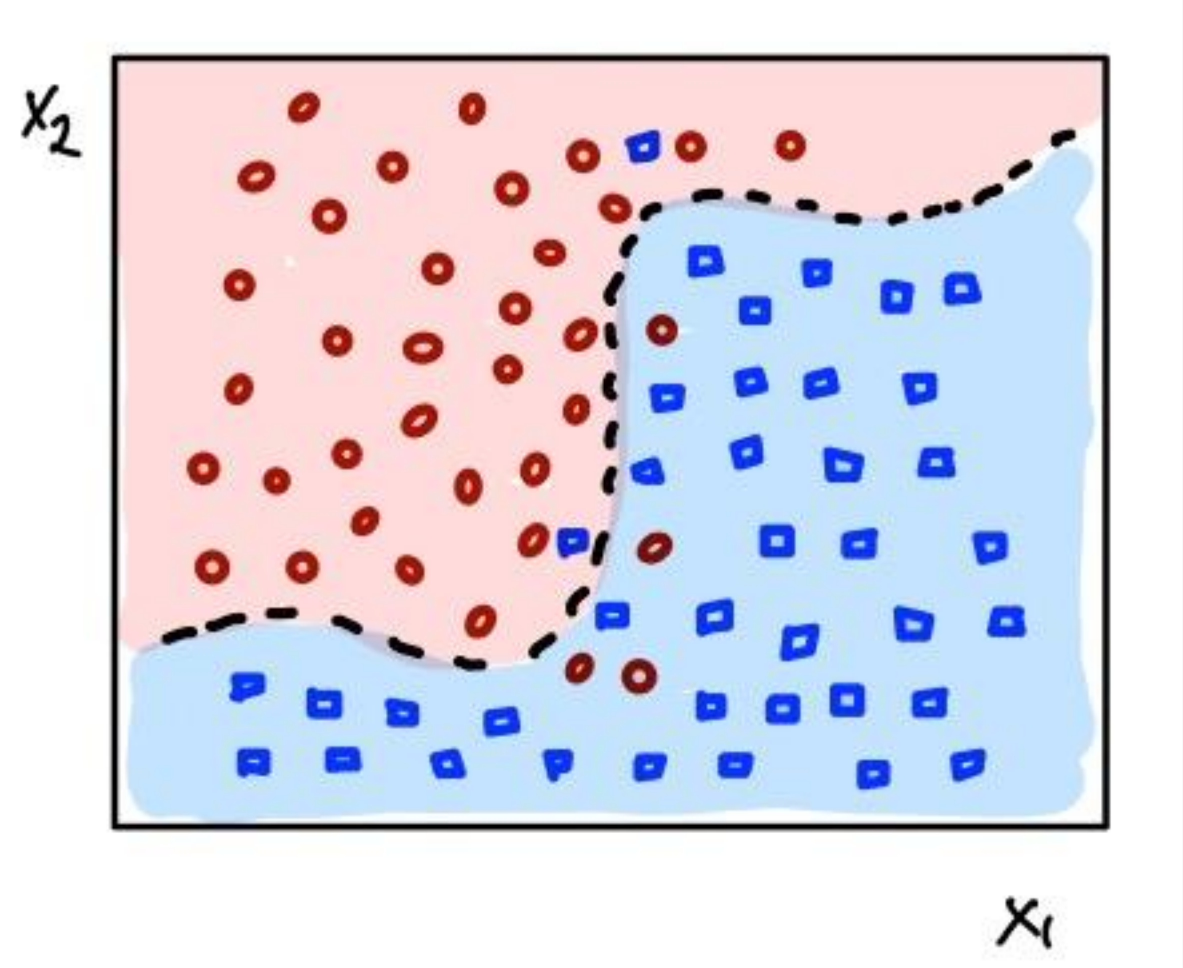
We transform the data by adding new predictors:

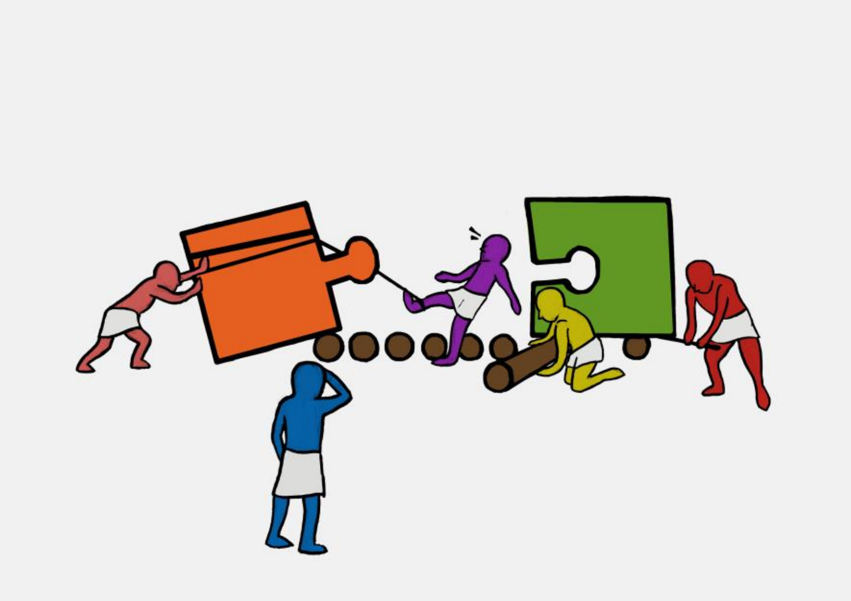
$$\tilde{x} = [1, \tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_M]$$

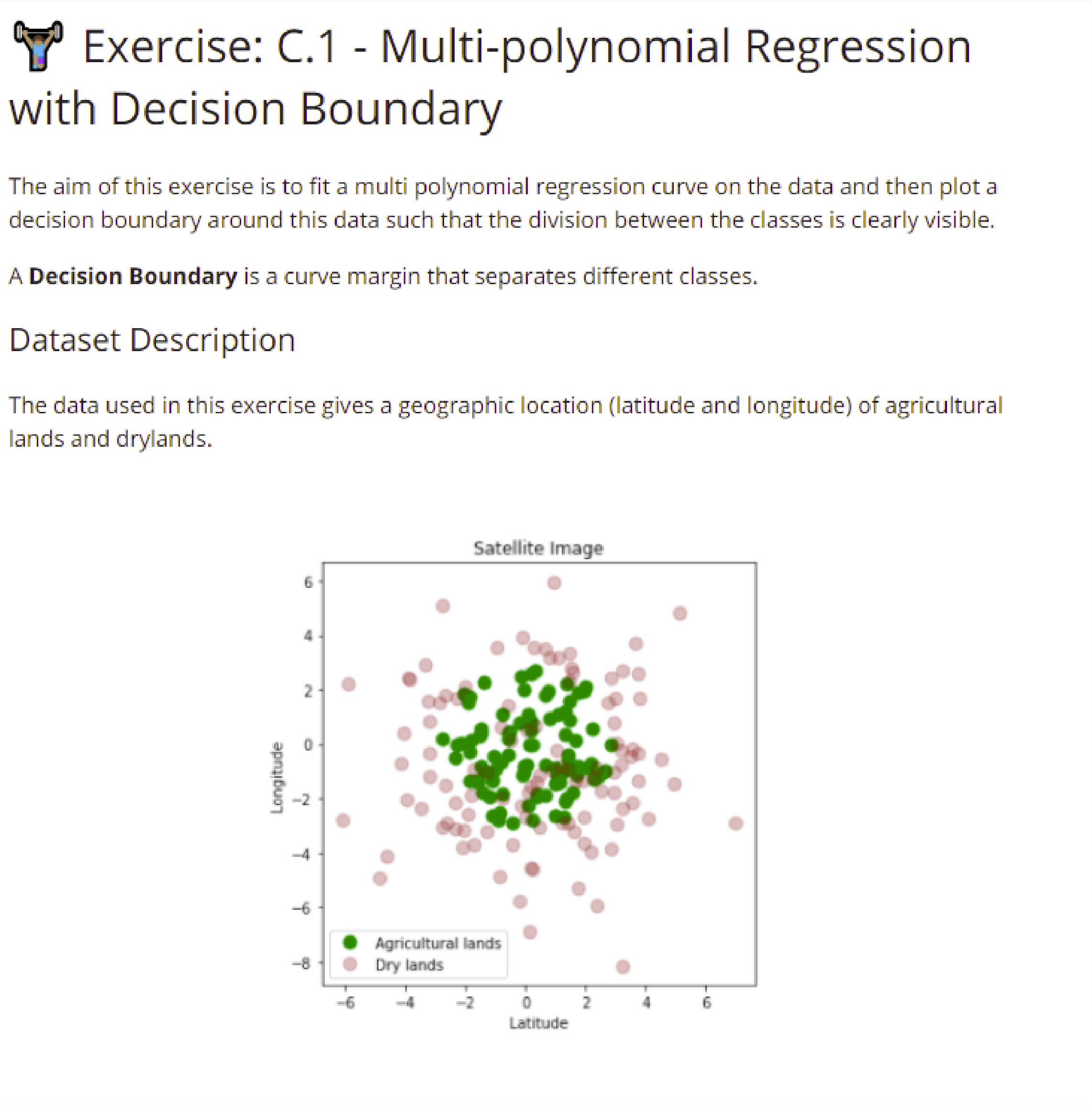
where $\tilde{x}_k = x^k$

The polynomial Logistic Regression can be expressed as:

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







Exercise: C.1 - Multi-polynomial Regression with Decision Boundary

The aim of this exercise is to fit a multi polynomial regression curve on the data and then plot a decision boundary around this data such that the division between the classes is clearly visible.

A **Decision Boundary** is a curve margin that separates different classes.

Dataset Description

The data used in this exercise gives a geographic location (latitude and longitude) of agricultural lands and drylands.