Logistic Regression

University of the Witwatersrand

2022

Review Question

- I roll two regular dice and add the values on the upward faces up. Find the probability that I get a seven.

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Idea of multivariate regression

Lesson Plan

- Review Question
- Idea of multivariate regression
- Mathematical formalisation

Review Question

- I roll two regular dice and add the values on the upward faces up. Find the probability that I get a seven.

Review Question

- I roll two regular dice and add the values on the upward faces up. Find the probability that I get a seven.
- Thirty-six equally likely options.

Review Question

- I roll two regular dice and add the values on the upward faces up. Find the probability that I get a seven.
- Thirty-six equally likely options.
- Six of them $(1, 6)$, $(2, 5)$,$(3, 4)$, $(4, 3)$, $(5, 2)$ and $(6, 1)$ give a score of seven

Review Question

- I roll two regular dice and add the values on the upward faces up. Find the probability that I get a seven.
- Thirty-six equally likely options.
- Six of them $(1, 6)$, $(2, 5)$,$(3, 4)$, $(4, 3)$, $(5, 2)$ and $(6, 1)$ give a score of seven
- $\frac{6}{36} = \frac{1}{6}$

Idea of Logistic Regression

- Sometimes we want to classify things.

Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.

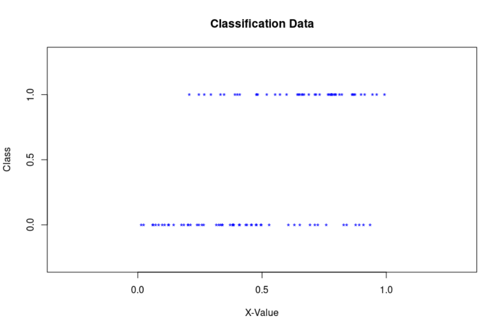
Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.
- The idea of logistic regression is that we'll set up an indicator random variable. 1 if in the group 0 if not (two class classification problem).

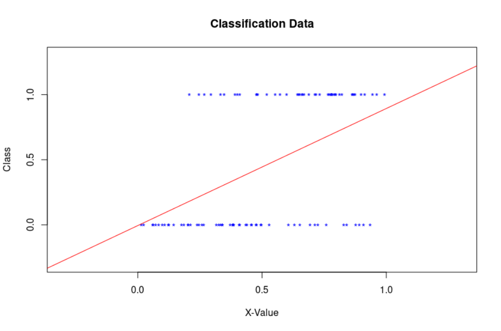
Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.
- The idea of logistic regression is that we'll set up an indicator random variable. 1 if in the group 0 if not (two class classification problem).
- Then we try to do regression on it.

A problem

- 

A problem



Fitting a curve

- OK so we'd like to fit a curve that's bounded between zero and one.

Fitting a curve

- OK so we'd like to fit a curve that's bounded between zero and one.
- This can be thought of as representing probability of being in the class.

Fitting a curve

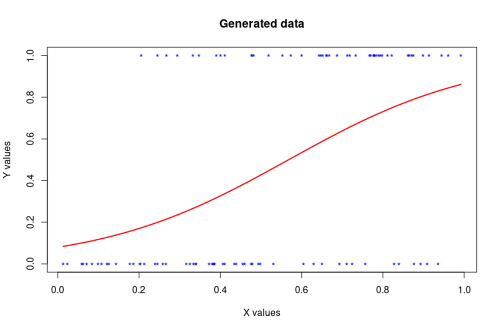
- OK so we'd like to fit a curve that's bounded between zero and one.
- This can be thought of as representing probability of being in the class.
- Could use a lot of functions. In machine learning we sometimes do. For logistic regression we use a sigmoid function

A problem

- Looks like:

A problem

- Looks like:



-

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$
- This is really a function of $Z = e^{\beta^T X}$. It's monotonic and ranges from 0 to 1.

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$
- This is really a function of $Z = e^{\beta^T X}$. It's monotonic and ranges from 0 to 1.
-

$$
\begin{aligned}
p(X) &= \frac{e^{\beta^T X}}{1 + e^{\beta^T X}} \\
p(X) &= \frac{Z}{1 + Z} \\
p(X)(1 + Z) &= Z \\
p(X) + Zp(X) &= Z
\end{aligned}
$$

Sigmoid Function

- $$\begin{aligned} p(X) + Zp(X) &= Z \\ p(X) &= Z(1 - p(X)) \\ Z &= \frac{p(X)}{1 - p(X)} \\ \frac{p(X)}{1 - p(X)} &= e^{\beta^T X} \\ \ln \frac{p(X)}{1 - p(X)} &= \beta^T X \end{aligned}$$

Sigmoid Function

- $$\begin{aligned} p(X) + Zp(X) &= Z \\ p(X) &= Z(1 - p(X)) \\ Z &= \frac{p(X)}{1 - p(X)} \\ \frac{p(X)}{1 - p(X)} &= e^{\beta^T X} \\ \ln \frac{p(X)}{1 - p(X)} &= \beta^T X \end{aligned}$$
- That is to say that the "log odds" is modeled as a linear function of the X variables.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.
- We will cover the derivation of this but the upshot is that we don't have a nice closed form solution for β and in practice rely of software to compute things (mostly these use Newton's method because gradient decent turns out to be slower for this problem).

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.
- We will cover the derivation of this but the upshot is that we don't have a nice closed form solution for β and in practice rely of software to compute things (mostly these use Newton's method because gradient decent turns out to be slower for this problem).
- Due to similarities with Linear Regression we can do inference on β. We won't here but your favourite software package will have tests.

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$
- $$L(\theta) = \prod_{i} p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$

- $$L(\theta) = \prod_i p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$

- $$\begin{aligned} l(\beta) &= \sum_i y_i \ln p(x_i) + (1 - y_i) \ln[1 - p(x_i)] \\ &= \sum_i y_i \ln \frac{p(x_i)}{1 - p(x_i)} + \ln[1 - p(x_i)] \\ &= \sum_i y_i(X^T\beta) - \ln[1 + e^{X^T\beta}] \end{aligned}$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$
- $$L(\theta) = \prod_i p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$
- $$\begin{aligned} l(\beta) &= \sum_i y_i \ln p(x_i) + (1 - y_i) \ln[1 - p(x_i)] \\ &= \sum_i y_i \ln \frac{p(x_i)}{1 - p(x_i)} + \ln[1 - p(x_i)] \\ &= \sum_i y_i(X^T\beta) - \ln[1 + e^{X^T\beta}] \end{aligned}$$
- Numerics