Logistic Regression - Day 1

Agenda

Logistic Regression - Day 1

In this session, you will be learning:

01 **Classification problems**

02 **Solving classification problems using linear regression**

03 **Interpreting coefficients of a logistic model**

04 **Finding out the optimal coefficients**

05 **Python demo**

Classification Problems

Classification Problems

Predict if a machine will fail in the next 14 days.

VibrationX_14day	VibrationY_14day	VibrationZ_14day	Failed
…	…	…	Yes
…	…	…	No

Classification Problems

The HR department of a company wants to understand which employees are at risk of resigning.

# promotions	Current salary	Market Salary	Resigned
...	...	...	Yes
...	...	...	No

Classification Problems

Can we predict which patients are at risk of re-admission?

Patient ID	Age	Gender	....
001	23	M	...
Patient ID	Age	Gender	....

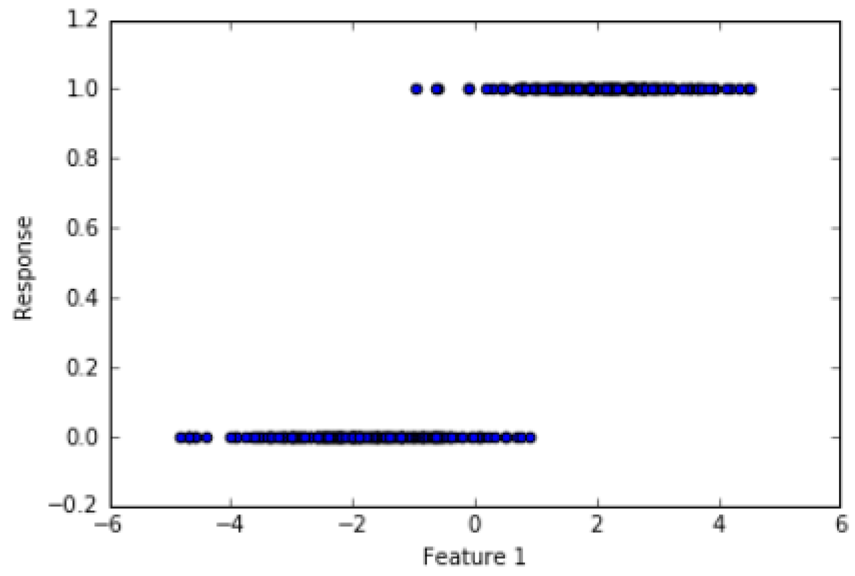
Classification Problems: Class Exercise

Take five minutes and discuss two scenarios in which the prediction problem is a classification problem. Also, discuss what kind of data you will need to collect.

- **Customer Churn Prediction**: Demographics, Past association with a product, and Number of times complaints registered
- **Predicting Fraud**: Demographics, Financial History, and Circumstances of transaction

Fitting Lines

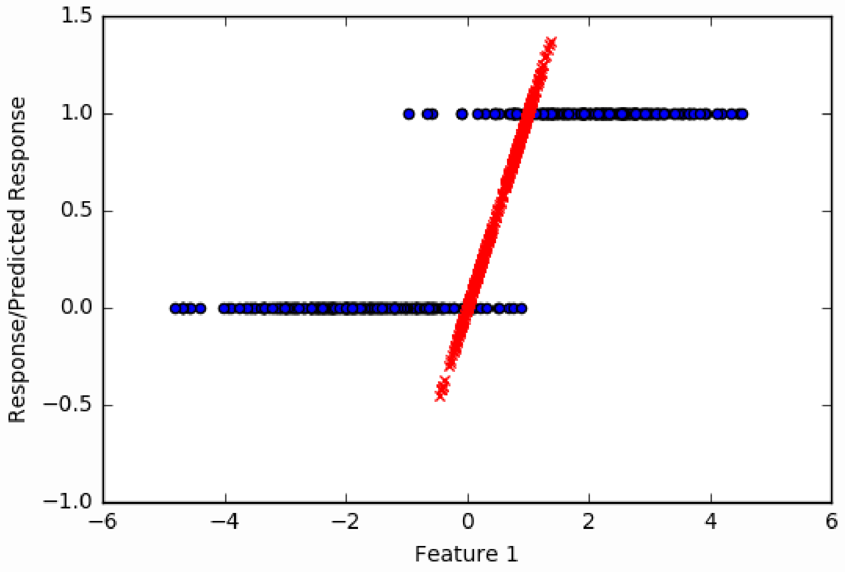
Supervised Learning: Fitting lines



Feature 1 (Age_Normalized)	Feature 2 (Income_Normalized)	Response (Good/Bad)	Feature 1 (Age_Normalized)
…..	……	Good = 1	…..
…..	……	Bad = 0	…..

Fit a model of the form, Response=b0+b1Feature1

Supervised Learning: Fitting lines



We fit a straight line to the data, where the response is binary in nature $y \in \{0,1\}$.

Notice the predictions in red color

1. Predictions < 0 or Predictions > 1
2. 0 < Predictions < 1

We are trying to fit:

$\hat{y} = \beta 0 + \beta 1 X1 + \beta 2 X2 + \varepsilon = f(X1, X2)$

The problem with the fitting linear model is:

$$f(X1, X2) \in (-\infty, +\infty)$$

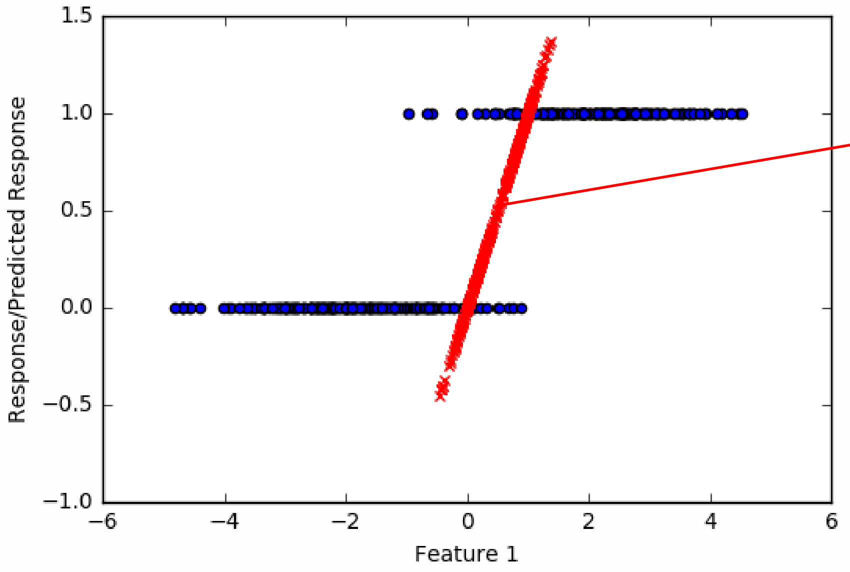
$$y \in \{0,1\}$$

Supervised Learning: Fitting lines

Age	Target
20	1
20	0
21	1
21	0

Age	Proportion (1)
20	0.50
21	0.50

Supervised Learning: Fitting lines



Instead of estimating $y \in \{0,1\}$ we can try to estimate the Prob(y=1) = $\hat{p}$

$\hat{p} \in (0,1)$

These estimates make sense now.

We are trying to fit:

$\hat{p}$ = β0+β1X1+β2X2+ε = $f(X1,X2)$

The problem still is:

$$f(X1,X2) \in (-\infty,+\infty)$$

$$\hat{p} \in (0,1)$$

Supervised Learning: Fitting lines

Age	Target
20	1
20	1
20	1
20	0
21	1
21	0

Age	Pr (1)	Pr/1-Pr
20	0.75	3
21	0.50	0.5

Supervised Learning: Fitting lines

Age	Target
20	1
20	1
20	1
20	0
21	1
21	0

Age	Pr (1)	Pr/1-Pr
20	0.75	3
21	0.50	0.5

Maximum and Minimum value Pr/(1-Pr)?

Minimum value of Pr(1)? Pr(1) = 0

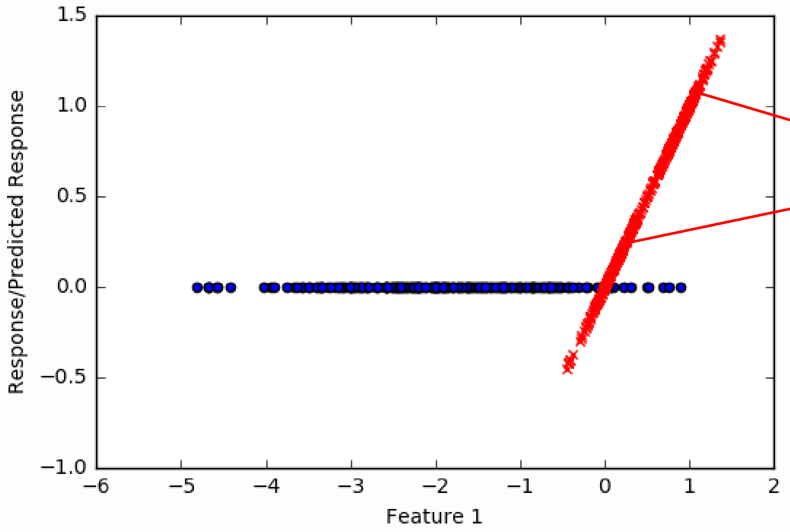
Maximum value of Pr(1)? Pr(1) = 1

Min Pr(1) = 0, Pr/(1-Pr) = ? Pr/(1-Pr) = 0

Max Pr(1) = 1, Pr/(1-Pr) = ? Pr/(1-Pr) = Infinity

0<=Pr/(1-Pr)<=Infinity

Supervised Learning: Fitting lines



Instead of estimating $y \in \{0,1\}$ we can try to estimate the $\hat{p}/(1-\hat{p})$

$$\hat{p}/(1-\hat{p}) \in (0,+\infty)$$

These estimates make sense now.

We are trying to fit:

$\hat{p}/(1-\hat{p}) = \beta 0+\beta 1X1+\beta 2X2+\varepsilon = f(X1,X2)$

The problem still is:

$$f(X1,X2) \in (-\infty,+\infty)$$

$$\hat{p}/(1-\hat{p}) \in (0,+\infty)$$

Supervised Learning: Fitting lines

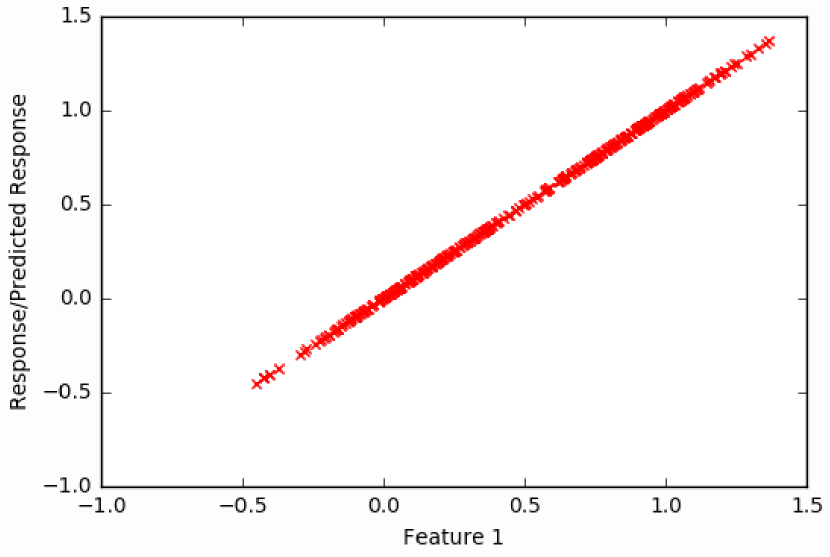
Age	Target
20	1
20	1
20	1
20	0
21	1
21	0

Age	Pr (1)	Pr/1-Pr	log(Pr/(1-Pr))
20	0.75	3	1.09
21	0.50	0.5	-0.70

0 <= Pr/(1-Pr) <= Infinity

-Infinity <= log(Pr/(1-Pr)) <= Infinity

Supervised Learning: Fitting lines



Instead of estimating $y \in \{0,1\}$ we can try to estimate the $\log(\hat{p}/(1-\hat{p}))$

$$\log \hat{p}/(1-\hat{p}) \in (-\infty,+\infty)$$

All these estimates make sense now.

We are trying to fit:

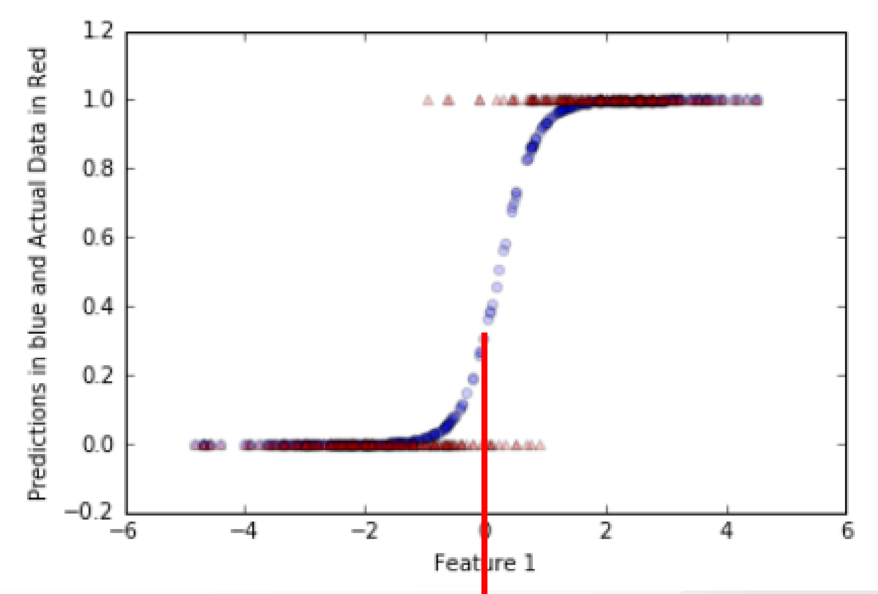
$\log \hat{p}/(1-\hat{p}) = \beta 0+\beta 1 X1+\beta 2 X2+\varepsilon = f(X1,X2)$

$f(X1,X2) \in (-\infty,+\infty) \log \hat{p}/(1-\hat{p}) \in (-\infty,+\infty)$

Supervised Learning: Fitting lines

$$\log\frac{\hat{p}}{1-\hat{p}} = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$$

$$\hat{p} = \frac{e^{\beta 0+\beta 1X1+\beta 2X2+\varepsilon}}{1 + e^{\beta 0+\beta 1X1+\beta 2X2+\varepsilon}}$$



Logistic Function

Classification Problems: Class Exercise

Imagine that there are 125 customers of age 25 years. Of them, 25 have subscribed to a premium subscription of an OTT platform. Find out the odds ratio.

$$\boxed{p = \frac{25}{125}, 1 - p = \frac{100}{125}, so\ odds\ ratio = \frac{p}{1-p} = \frac{25}{125} * \frac{125}{100} = \frac{1}{4}}$$

Interpreting Coefficients of a Logistic Model

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = 2.1+0.08Age$$

Age	Churned
28	Yes
32	Yes
40	No

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = 2.1+0.08Age$$

Changing Age by 1 unit will change the log odds of someone churning by 0.08 units.

Interpreting Logistic Regression Coefficients

Predicting churn based on a person’s age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = 2.1 + \mathbf{0.08}\,Age$$

Changing Age by 1 unit will change the log odds of someone churning by 0.08 units.

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$\log(p/(1-p)) = 2.1 + \mathbf{0.08 * 20}$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$\log(p/(1-p)) = 2.1 + 1.6$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$\log(p/(1-p)) = 3.7$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person’s age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = e^{3.7}$$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = 2.71^{3.7}$$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$$\log(p/(1-p)) = 39.99$$

Interpreting Logistic Regression Coefficients

Predicting churn based on a person's age

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$$

$P = 0.97$

Class Exercise

- Imagine you were trying to model the propensity of customers to churn, and you were able to build the following logistic regression model:

$$\log \frac{p}{1-p} = 1.95 - 2.5Age + 1.68Income$$

- Can we conclude that as age increases, the propensity for a customer to churn decreases, keeping all else constant?
- What would be the churn probability for a person aged 25 with an income of 15000?

Class Exercise

- Imagine you were trying to model the propensity of customers to churn, and you were able to build the following logistic regression model:

$$\log\frac{p}{1-p} = 0.0001 - 0.005Age + 0.008Income$$

- Can we conclude that as age increases, the propensity for a customer to churn decreases, keeping all else constant? (Yes, negative sign on the coefficient of Age.)
- What would be the probability of churn for a person with age 25 and income 150?($p = \frac{E}{1+E}$, where $\mathrm{E} = \mathrm{e}^{0.0001-0.005*25+0.008*150}$, $\mathrm{p} = 0.75$)

Finding Out the Optimal Coefficients

Estimating Coefficients of Logistic Model

Age	Good_Bad (Good =1)	Prediction
20	1	
21	1	
24	0	
25	0	
29	0	
30	1	
38	1	

$$\log\frac{\hat{p}}{1-\hat{p}} = \beta 0 + \beta 1 Age$$

$$\hat{p} = \frac{e^{\beta 0+\beta 1 Age}}{1 + e^{\beta 0+\beta 1 Age}}$$

$$\beta 0 = 0.7$$
$$\beta_1 = 1.7$$

$$\boxed{\hat{p} = \frac{e^{0.7+1.7Age}}{1 + e^{0.7+1.7Age}}}$$

Estimating Coefficients of Logistic Model

Age	Good_Bad (Good =1)	Prediction
20	1	
21	1	
24	0	
25	0	
29	0	
30	1	
38	1	

$$\beta 0 = 0.7$$
$$\beta_1 = 1.7$$

$$\log \frac{\hat{p}}{1-\hat{p}} = \beta 0 + \beta 1 Age$$

$$\hat{p} = \frac{e^{\beta 0+\beta 1 Age}}{1 + e^{\beta 0+\beta 1 Age}}$$

Age	Good_Bad (Good =1)	Prediction
20	1	
21	1	
24	0	
25	0	
29	0	
30	1	
38	1	

$$\beta 0 = 0.3$$
$$\beta_1 = 2.2$$

Estimating Coefficients of Logistic Model

Age	Good_Bad (Good =1)	Prediction
20	1	
21	1	
24	0	
25	0	
29	0	
30	1	
38	1	

$$\beta 0 = 0.7$$
$$\beta_1 = 1.7$$

$$\log \frac{\hat{p}}{1-\hat{p}} = \beta 0 + \beta 1 Age$$

$$\hat{p} = \frac{e^{\beta 0 + \beta 1 Age}}{1 + e^{\beta 0 + \beta 1 Age}}$$

Age	Good_Bad (Good =1)	Prediction
20	1	0.70
21	1	0.60
24	0	0.50
25	0	0.45
29	0	0.70
30	1	0.62
38	1	0.40

$$\beta 0 = 0.3$$
$$\beta_1 = 2.2$$

Estimating Coefficients of Logistic Model

Age	Good_Bad (Good =1)	Prediction
20	1	
21	1	
24	0	
25	0	
29	0	
30	1	
38	1	

$$\beta 0 = 0.7$$
$$\beta_1 = 1.7$$

Clearly $\beta 0 = 0.7$ and $\beta 1 = 1.7$ is a better choice

$$\log\frac{\hat{p}}{1-\hat{p}} = \beta 0 + \beta 1 Age$$

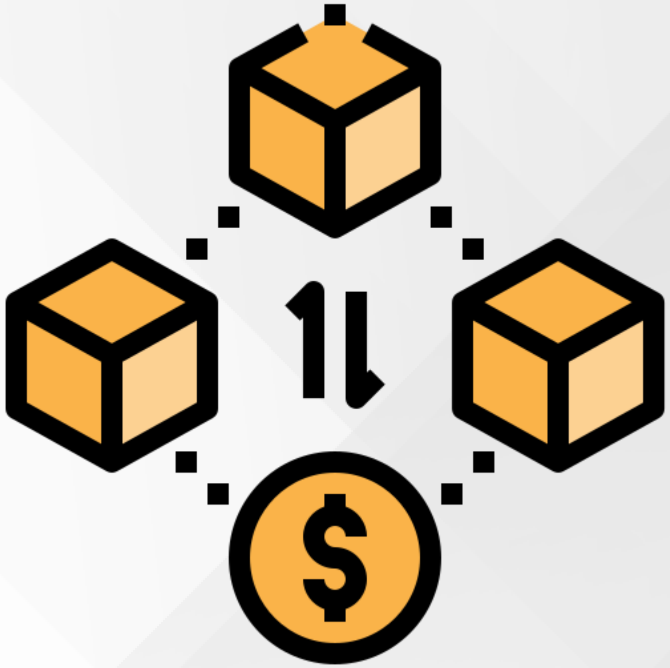
$$\hat{p} = \frac{e^{\beta 0+\beta 1 Age}}{1 + e^{\beta 0+\beta 1 Age}}$$

Age	Good_Bad (Good =1)	Prediction
20	1	0.70
21	1	0.60
24	0	0.50
25	0	0.45
29	0	0.70
30	1	0.62
38	1	0.40

$$\beta 0 = 0.3$$
$$\beta_1 = 2.2$$

Estimating Coefficients of Logistic Model

- One would like to choose the model coefficients so that the model gives a high score to events and a low score to non-events.
- But how will we measure a model's ability to assign a high score to events and a low to non-events? Cost Function. See the excel sheet **logistic_cost.xlsx**



Thank You!