INFO 6105
Data Science Engineering Methods and Tools

Lecture 5
Logistic Regression

Ebrahim Nasrabadi
nasrabadi@northeastern.edu

College of Engineering
Northeastern University

Fall 2019

Statistic Applications

Statistics can be used to answer the following questions:

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?

Statistic Applications

Statistics can be used to answer the following questions:

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?
- Is there a relationship between advertising budget and sales? If yes, how strong is the relationship?

Statistic Applications

Statistics can be used to answer the following questions:

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?
- Is there a relationship between advertising budget and sales? If yes, how strong is the relationship?
- How accurately can we predict demand for a product or service?

Statistic Applications

Statistics can be used to answer the following questions:

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?
- Is there a relationship between advertising budget and sales? If yes, how strong is the relationship?
- How accurately can we predict demand for a product or service?
- What is the default risk (i.e., the chance that companies or individuals will be unable to make the required payments on their debt obligations)?

Statistic Applications

Statistics can be used to answer the following questions:

- What is the price sensitivity of a product or service (e.g., the degree to which price affects the sales)?
- Is there a relationship between advertising budget and sales? If yes, how strong is the relationship?
- How accurately can we predict demand for a product or service?
- What is the default risk (i.e., the chance that companies or individuals will be unable to make the required payments on their debt obligations)?
- What is the chance that a house goes under contract in its first two weeks on the market?

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as “sales”.
- Classification: A classification problem is when the response is a category, such as “click” or“non-click”

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as “sales”.
- Classification: A classification problem is when the response is a category, such as “click” or“non-click”

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as “sales”.
- Classification: A classification problem is when the response is a category, such as “click” or“non-click”

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as "sales".
- Classification: A classification problem is when the response is a category, such as "click" or"non-click"

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not
- Health : Predicting if a patient has diabetes based on certain diagnostic measurements

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as “sales”.
- Classification: A classification problem is when the response is a category, such as “click” or“non-click”

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not
- Health : Predicting if a patient has diabetes based on certain diagnostic measurements
- Marketing : Predicting if a given user will buy an insurance product or not

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as "sales".
- Classification: A classification problem is when the response is a category, such as "click" or"non-click"

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not
- Health : Predicting if a patient has diabetes based on certain diagnostic measurements
- Marketing : Predicting if a given user will buy an insurance product or not
- Banking: Predicting if a customer will default on a loan

Regression vs Classification

- Regression: A regression problem is when the response is a real value, such as "sales".
- Classification: A classification problem is when the response is a category, such as "click" or "non-click"

Examples of binary classification problems:

- Spam Detection : Predicting if an email is Spam or not
- Credit Card Fraud : Predicting if a given credit card transaction is fraud or not
- Health : Predicting if a patient has diabetes based on certain diagnostic measurements
- Marketing : Predicting if a given user will buy an insurance product or not
- Banking: Predicting if a customer will default on a loan
- Advertising: Predicting if a user will click on an ad

Credit Card Default Data

Consider a data set containing information on 10,000 customers:

- default A variable with levels No and Yes indicating whether the customer defaulted on their debt
- student A variable with levels No and Yes indicating whether the customer is a student
- balance The average balance that the customer has remaining on their credit card after making their monthly payment
- income Income of customer

Credit Card Default Data

default	student	balance	income
No	No	730	44362
No	Yes	817	12106
No	No	1074	31767
No	No	529	35704
No	No	786	38463
No	Yes	920	7492
No	No	826	24905
No	Yes	809	17600
No	No	1161	37469
No	No	0	29275
No	Yes	0	21871
No	Yes	1221	13269

Credit Card Default Example

- We are interested to predict predict which customers will default on their credit card debt on the basis of the other variables.
- We refer to the
  - default variable as *target* (also called *response* or *output*) variable
  - student, balance, income variables as *predictors* (also called *features* or *inputs*).

Credit Card Default Example

- We are interested to predict predict which customers will default on their credit card debt on the basis of the other variables.
- We refer to the
  - default variable as *target* (also called *response* or *output*) variable
  - student, balance, income variables as *predictors* (also called *features* or *inputs*).
- Goal:
  - understand the relationship between response and predictors
  - make predictions: what is the chance if a customer will default on a loan?

Credit Card Default Example

- Consider only one predictor: balance
- Goal: We want to
  - understand the relationship between default and balance for Credit Card Default Data
  - predict the probability of default based on the balance.

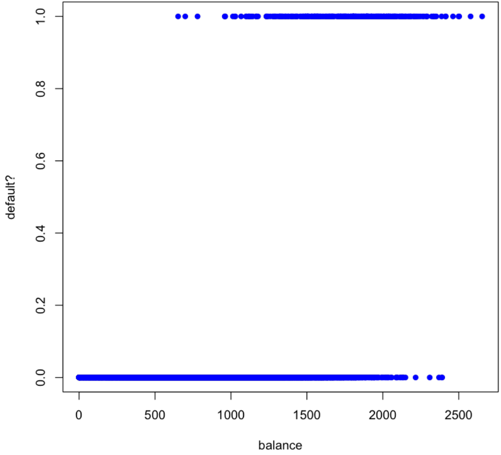
Credit Card Default Example

- Consider only one predictor: balance
- Goal: We want to
  - understand the relationship between default and balance for Credit Card Default Data
  - predict the probability of default based on the balance.
- The response variable is default and the predictor variable is balance.
- We denote the response variable by Y and the predictor variable by X

default	balance
No	1690.234
No	1505.783
No	1536.595
No	1578.064
No	1722.356
No	1557.345
Yes	2205.800
No	1802.903
Yes	1774.694
No	1747.259

We create a dummy variable to represent the qualitative variable default:

$$Y = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{Otherwise} \end{cases}$$



Qualitative Response Variable

Question: Can we use linear regression?

$$\textbf{default} \simeq \beta_0 + \beta_1 \times \textbf{balance}$$

or mathematically

$$Y \simeq \beta_0 + \beta_1 \times X.$$

Qualitative Response Variable

Question: Can we use linear regression?

$$\textbf{default} \simeq \beta_0 + \beta_1 \times \textbf{balance}$$

or mathematically

$$Y \simeq \beta_0 + \beta_1 \times X.$$

Notes:

- The response must be either 0 (default) or 1 (non-default). We seek a function/hypothesis $h : \mathbb{R} \longrightarrow \{0, 1\}$ to predict a customer will default on a loan given X.

Qualitative Response Variable

Question: Can we use linear regression?

$$\textbf{default} \simeq \beta_0 + \beta_1 \times \textbf{balance}$$

or mathematically

$$Y \simeq \beta_0 + \beta_1 \times X.$$

Notes:

- The response must be either 0 (default) or 1 (non-default). We seek a function/hypothesis $h : \mathbb{R} \longrightarrow \{0, 1\}$ to predict a customer will default on a loan given X.
- We are often interested in probabilities. In this case, we see a function $h : \mathbb{R} \longrightarrow [0, 1]$ to predict the probability that a customer will default on a loan given X to the positive class, that is,

$$h(X) = P(Y = 1|X)$$

Linear Regression vs Logistic regression

If we use linear regression, the output can be negative or greater than 1 whereas probability can not.

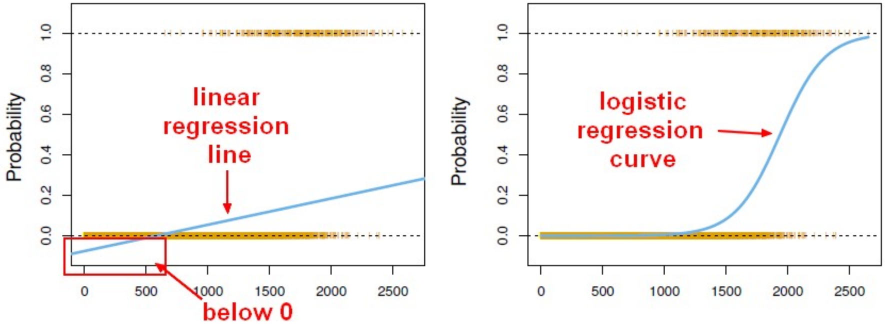


Figure: Source: http://gerardnico.com/wiki/data_mining/simple_logistic_regression

Logistic Regression

Modeling Choice: We choose

$$P(\text{the loan will be default given balance}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \textbf{balance})}}$$
$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times X)}}$$

where $f(z) = \frac{1}{1+e^{-z}}$ is called a *logistic function* and β_0, β_1 are called *model parameters.*

Modeling Choice: We choose

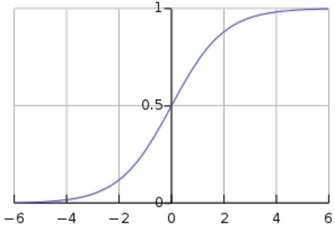
$$P(\text{the loan will be default given balance}) = \frac{1}{1+e^{-(\beta_0+\beta_1\times\textbf{balance})}}$$

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1\times X)}}$$

where $f(z) = \frac{1}{1+e^{-z}}$ is called a *logistic function* and β_0, β_1 are called *model parameters*.

Properties:

- $0 \leq f(z) \leq 1$
- $f(0) = 0.5$
- $f(z) \longrightarrow 1$ as $z \longrightarrow \infty$
- $f(z) \longrightarrow 0$ as $z \longrightarrow -\infty$



How to determine β_0 and β_1?

- We can represent data as

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

  where $x_i =$ balance and $y_i =$ default for i^{th} observation.
- Find the model parameters β_0 and β_1 such that the predicted probability $h_\beta(\vec{x}_i) = P(y_i|\vec{x}_i; \beta)$
  - is close to one if $y_i = 1$
  - is close to zero if $y_i = 0$

Maximum Likelihood

Note that

$$
\begin{aligned}
p(y_i|\vec{x}_i;\beta) &= h_\beta(\vec{x}_i)^{y_i} \cdot (1 - h_\beta(\vec{x}_i))^{1-y_i} \\
&= \begin{cases} h_\beta(\vec{x}_i) & \text{if } y_i = 1 \\ 1 - h_\beta(\vec{x}_i) & \text{if } y_i = 0 \end{cases}
\end{aligned}
$$

Maximum Likelihood

Note that

$$p(y_i|\vec{x}_i;\beta) = h_\beta(\vec{x}_i)^{y_i} \cdot (1 - h_\beta(\vec{x}_i))^{1-y_i}$$
$$= \begin{cases} h_\beta(\vec{x}_i) & \text{if } y_i = 1 \\ 1 - h_\beta(\vec{x}_i) & \text{if } y_i = 0 \end{cases}$$

Assuming the data is generated independently, the probability of the observing $y_1, \ldots, y_n$ is given by

$$L(\beta) = P(Y|\vec{x};\beta) = \prod_{i=1}^{m} P(y_i|\vec{x}_i;\beta)$$
$$= \prod_{i:y_i=1} h_\beta(X_i) \cdot \prod_{i:y_i=0} (1 - h_\beta(\vec{x}_i))$$

Maximum Likelihood

Note that

$$p(y_i|\vec{x}_i;\beta) = h_\beta(\vec{x}_i)^{y_i} \cdot (1 - h_\beta(\vec{x}_i))^{1-y_i}$$
$$= \begin{cases} h_\beta(\vec{x}_i) & \text{if } y_i = 1 \\ 1 - h_\beta(\vec{x}_i) & \text{if } y_i = 0 \end{cases}$$

Assuming the data is generated independently, the probability of the observing $y_1, \ldots, y_n$ is given by

$$L(\beta) = P(Y|\vec{x};\beta) = \prod_{i=1}^{m} P(y_i|\vec{x}_i;\beta)$$
$$= \prod_{i:y_i=1} h_\beta(X_i) \cdot \prod_{i:y_i=0} (1 - h_\beta(\vec{x}_i))$$

Choose β so as to maximize $L(\beta)$ or equivalently minimize $-\log L(\beta)$.

Maximum Likelihood

Choose β so as to maximize $L(\beta)$ or equivalently minimize $-\frac{1}{n}\log L(\beta)$.

$$\mathcal{L}(\vec{\beta}) := -\frac{1}{n}\log L(\beta) = -\frac{1}{n}\sum_i \left[y_i \log(h_\beta(x_i)) + (1-y_i)\log(1-h_\beta(x_i))\right]$$

where $h_\beta(x_i) = \frac{1}{1+e^{-\beta^T \vec{x}_i}}$.
This does not have an explicit solution, we need to minimize numerically!

Gradient Descent

Start with some initial β and repeatedly take a step in the direction of steepest decrease of $-\log L(\beta)$:

$$\beta_j = \beta_j - \eta \frac{\partial}{\partial \beta_j} \mathcal{L}(\vec{\beta})$$

η is called *learning rate*.

Gradient Descent

Start with some initial β and repeatedly take a step in the direction of steepest decrease of $-\log L(\beta)$:

$$\beta_j = \beta_j - \eta \frac{\partial}{\partial \beta_j} \mathcal{L}(\vec{\beta})$$

η is called *learning rate*.

Update Rule: For training examples $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\beta_j = \beta_j + \frac{\eta}{n} \sum_{i=0}^{n} (y_i - h_\beta(x_i)) x_{ij} \qquad\qquad j = 0, 1$$

Credit Card Default Example

```
Call:
glm(formula = default ~ balance, family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2697  -0.1465  -0.0589  -0.0221   3.7589

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Making Predictions

Once the model parameters have been determined, it is a simple matter to predict the probability that a new customer will default on a loan:

$$
\begin{aligned}
P(\text{the loan will be default given balance}) &= \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times\textbf{balance})}} \\
&= \frac{1}{1+e^{10.65-0.005499\times\textbf{balance}}}
\end{aligned}
$$

Making Predictions

Once the model parameters have been determined, it is a simple matter to predict the probability that a new customer will default on a loan:

$$P(\text{the loan will be default given balance}) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times \textbf{balance})}}$$
$$= \frac{1}{1 + e^{10.65 - 0.005499 \times \textbf{balance}}}$$

We next address two questions:

- Model Fit: How does the model fit the data?
- Build a classifier: How to decide whether a new customer will default on a loan?

Confusion Matrix

We need to determine a threshold:
if the estimated probability $\geq$ threshold then $\hat{y}_i = 1$, o.w. $\hat{y}_i = 0$.

Confusion Matrix

We need to determine a threshold:
if the estimated probability $\geq$ threshold then $\hat{y}_i = 1$, o.w. $\hat{y}_i = 0$.

i	y_i	$h_\beta(\vec{x}_i)$	Prediction Outcome (Threshold=0.6)	
1	1	0.9	1	TP
2	1	0.8	1	TP
3	0	0.7	1	FP
4	1	0.6	1	TP
5	0	0.5	0	TN
6	1	0.4	0	FN
7	0	0.3	0	TN
8	0	0.2	0	TN
9	0	0.1	0	TN
10	0	0.05	0	TN

Confusion Matrix

We need to determine a threshold:
if the estimated probability $\geq$ threshold then $\hat{y}_i = 1$, o.w. $\hat{y}_i = 0$.

i	y_i	$h_\beta(\vec{x}_i)$	Prediction Outcome (Threshold=0.6)	
1	1	0.9	1	TP
2	1	0.8	1	TP
3	0	0.7	1	FP
4	1	0.6	1	TP
5	0	0.5	0	TN
6	1	0.4	0	FN
7	0	0.3	0	TN
8	0	0.2	0	TN
9	0	0.1	0	TN
10	0	0.05	0	TN

		Actual Class		
		Positive	Negative	Total
Prediction Outcome	Positive	3	1	4
	Negative	1	5	6
	Total	4	6	10

Confusion Matrix

We need to determine a threshold:
if the estimated probability $\geq$ threshold then $\hat{y}_i = 1$, o.w. $\hat{y}_i = 0$.

i	y_i	$h_\beta(\vec{x}_i)$	Prediction Outcome (Threshold=0.6)	
1	1	0.9	1	TP
2	1	0.8	1	TP
3	0	0.7	1	FP
4	1	0.6	1	TP
5	0	0.5	0	TN
6	1	0.4	0	FN
7	0	0.3	0	TN
8	0	0.2	0	TN
9	0	0.1	0	TN
10	0	0.05	0	TN

		Actual Class		
		Positive	Negative	Total
Prediction Outcome	Positive	3	1	4
	Negative	1	5	6
	Total	4	6	10

- Accuracy=fraction of examples that are classified correctly=8/10=0.8
- True Positive Rate=fraction of positive examples that are classified correctly=$\frac{3}{4}$ $= 0.75$
- False Positive Rate=fraction of negative examples that are classified incorrectly=$\frac{1}{6}$ =0.16

Confusion Matrix

		Actual Class		
		Positive	Negative	Total
Prediction Outcome	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
	Total	TP+FN	FP+TN	n

Confusion Matrix

		Actual Class		
		Positive	Negative	Total
Prediction Outcome	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
	Total	TP+FN	FP+TN	n

- Accuracy: fraction of examples that are classified correctly$=\frac{TP+TN}{n}$
- Balanced Accuracy: $\frac{1}{2}(\frac{TP}{N_+}+\frac{TN}{N_-})$
- True Positive Rate: fraction of positive examples that are classified correctly$=\frac{TP}{TP+FN}$
- False Positive Rate: fraction of negative examples that are classified incorrectly$=\frac{FP}{FP+TN}$
- True Negative Rate: fraction of negative examples that are classified correctly$=\frac{TN}{FP+TN}$
- False Negative Rate: fraction of positive examples that are classified incorrectly$=\frac{FN}{TP+FN}$

Receiver Operator Characteristic

We want

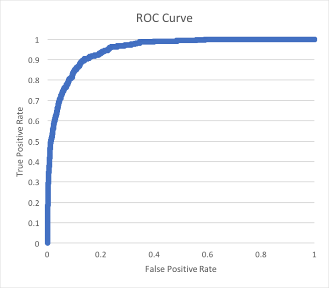
- TP rate to be as large as possible
- FP rate to be as small as possible

Receiver Operator Characteristic

We want

- TP rate to be as large as possible
- FP rate to be as small as possible

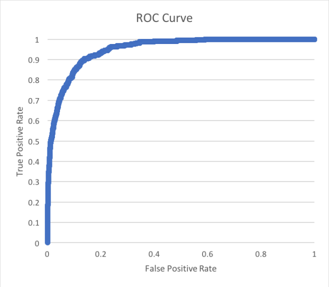
Threshold	TP rate	FP rate
0.0	$4/4 = 1.00$	$1/6 = 1.00$
0.1	$4/4 = 1.00$	$5/6 = 0.83$
0.2	$4/4 = 1.00$	$4/6 = 0.67$
0.3	$4/4 = 1.00$	$3/6 = 0.50$
0.4	$4/4 = 1.00$	$2/6 = 0.33$
0.5	$3/4 = 0.75$	$2/6 = 0.33$
0.6	$3/4 = 0.75$	$1/6 = 0.17$
0.7	$2/4 = 0.50$	$1/6 = 0.17$
0.8	$2/4 = 0.50$	$0/6 = 0.00$
0.9	$1/4 = 0.25$	$0/6 = 0.00$
1.0	$0/4 = 0.00$	$0/6 = 0.00$



We want

- TP rate to be as large as possible
- FP rate to be as small as possible

Threshold	TP rate	FP rate
0.0	$4/4 = 1.00$	$1/6 = 1.00$
0.1	$4/4 = 1.00$	$5/6 = 0.83$
0.2	$4/4 = 1.00$	$4/6 = 0.67$
0.3	$4/4 = 1.00$	$3/6 = 0.50$
0.4	$4/4 = 1.00$	$2/6 = 0.33$
0.5	$3/4 = 0.75$	$2/6 = 0.33$
0.6	$3/4 = 0.75$	$1/6 = 0.17$
0.7	$2/4 = 0.50$	$1/6 = 0.17$
0.8	$2/4 = 0.50$	$0/6 = 0.00$
0.9	$1/4 = 0.25$	$0/6 = 0.00$
1.0	$0/4 = 0.00$	$0/6 = 0.00$



Receiver Operator Characteristic (ROC) is a plot of TP rate against FP rate, it shows the tradeoff between FP and TP rates for various thresholds.

Area Under Curve

The area under the ROC curve (AUC or "*Area Under Curve*") is another measure of classification accuracy: the closer the AUC to one the more accurate the classification.

Area Under Curve

The area under the ROC curve (AUC or "*Area Under Curve*") is another measure of classification accuracy: the closer the AUC to one the more accurate the classification.

AUC: the probability that the classifier will rank a randomly positive example higher than a randomly chosen negative example:

$$\begin{aligned} \text{AUC} &\simeq P\big(\text{score}(X^+) \geq \text{score}(X^-)\big) \\ &\simeq \frac{U}{|N_+| \cdot |N_-|} \end{aligned}$$

where

- N_+ is the set of positive examples
- N_- is the set of negative examples
- $U = \sum_{X_i \in N_+} \sum_{X_j \in N_-} [\text{score}(X_i) > \text{score}(X_j)]$

Credit Card Default Example

Predict the probability of default based on the student, balance, and income.

Logistic regression:

```
Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Credit Card Default Example

Predict the probability of default based on the student, balance, and income.

Logistic regression:

```
Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- There a negative relationship between student and default
- There a positive relationship between balance and default

Making Prediction:

$$P(\text{the loan will be default})$$

$$= \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times \textbf{studentYes} + \hat{\beta}_2 \times \textbf{balance} + \hat{\beta}_3 \times \textbf{income})}}$$

$$= \frac{1}{1 + e^{10.87 + 0.6468 \textbf{studentYes} - 0.005737 \textbf{balance} - 0.000003033 \times \textbf{income}}}$$

Credit Card Default Example

Making Prediction:

$$P(\text{the loan will be default})$$

$$= \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times \textbf{studentYes} + \hat{\beta}_2 \times \textbf{balance} + \hat{\beta}_3 \times \textbf{income})}}$$

$$= \frac{1}{1 + e^{10.87 + 0.6468 \textbf{studentYes} - 0.005737 \textbf{balance} - 0.000003033 \times \textbf{income}}}$$

studentYes	balance	income	Probability of Default
1	$ 1,200	$ 40,000	0.011
0	$ 1,200	$ 40,000	0.021
1	$ 1,500	$ 40,000	0.059
1	$ 1,800	$ 40,000	0.256
0	$ 2,000	$ 40,000	0.674

Making Prediction:

$$
\begin{aligned}
&P(\text{the loan will be default}) \\
&\quad = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1\times\mathbf{studentYes}+\hat{\beta}_2\times\mathbf{balance}+\hat{\beta}_3\times\mathbf{income})}} \\
&\quad = \frac{1}{1+e^{10.87+0.6468\mathbf{studentYes}-0.005737\mathbf{balance}-0.000003033\times\mathbf{income}}}
\end{aligned}
$$

Making Prediction:

$$P(\text{the loan will be default})$$

$$= \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \times \textbf{studentYes} + \hat{\beta}_2 \times \textbf{balance} + \hat{\beta}_3 \times \textbf{income})}}$$

$$= \frac{1}{1 + e^{10.87 + 0.6468\textbf{studentYes} - 0.005737\textbf{balance} - 0.000003033 \times \textbf{income}}}$$

studentYes	balance	income	Probability of Default
1	$ 1,200	$ 40,000	0.011
0	$ 1,200	$ 40,000	0.021
1	$ 1,500	$ 40,000	0.059
1	$ 1,800	$ 40,000	0.256
0	$ 2,000	$ 40,000	0.674

Credit Card Default Example

Confusion Matrix (Threshold =0.6)

		Actual Class		
		Positive	Negative	Total
Prediction Outcome	Positive	81	23	104
	Negative	252	9,633	9896
	Total	333	9667	10,000

Credit Card Default Example

Confusion Matrix (Threshold =0.6)

		Actual Class: Positive	Actual Class: Negative	Total
Prediction Outcome	Positive	81	23	104
	Negative	252	9,633	9896
	Total	333	9667	10,000

- Accuracy: fraction of examples that are classified correctly$=\frac{81+9{,}644}{10{,}000} = 0.9725$
- Balanced Accuracy: $\frac{1}{2}(\frac{81}{333} + \frac{9644}{9667}) = 0.6204$
- True Positive Rate: fraction of positive examples that are classified correctly$=\frac{81}{333} = 0.2432$
- False Positive Rate: fraction of negative examples that are classified incorrectly$=\frac{23}{9{,}667} = 0.0023$
- True Negative Rate: fraction of negative examples that are classified correctly$=\frac{9{,}644}{9{,}667} = 0.9976$
- False Negative Rate: fraction of positive examples that are classified incorrectly$=\frac{252}{333} = 0.7567$

ROC Curve

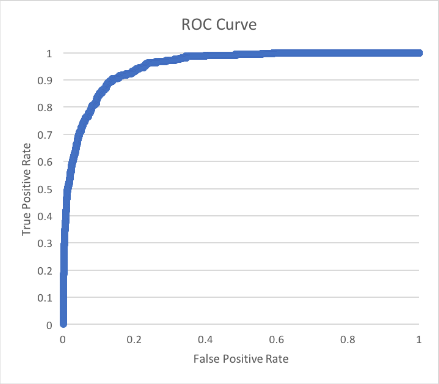


Figure: AUC=0.945

Unbalanced Classes

In many applications such as click prediction, we have disproportional data for various classes.

Typical manifestation of this problem is classifier classifying everything to one class.

Question How to tackle this problem?

Unbalanced Classes

In many applications such as click prediction, we have disproportional data for various classes.

Typical manifestation of this problem is classifier classifying everything to one class.

Question How to tackle this problem?

- Down-sample Downsample the over-represented classes such that all classes have similar amount of data

  Downside doesn't use all the data

- Up-sampling

  Downside overfilling and computational overhead

Consider the case where the response variable has multiple classes $C_1, C_2, \ldots, C_K$ with $K > 2$.

Multiclass Classification

Consider the case where the response variable has multiple classes $C_1, C_2, \ldots, C_K$ with $K > 2$.

We decompose this problem into a set of binary problems.

One-vs-All (OVA) Classification: We build K different binary classifiers:

- For $k = 1, ..., K$, do
  - let the positive examples be all the points in class k and the negative examples be all points not in class k.
  - let h_k be the k^{th} classifier:

$$h_k(X) = P(y \in C_k | X) = P(\text{the example is positive}|X)$$

  - Predict the most likely class: the predicted class for a new observation X is $\arg\max_k h_k(X)$

All-vs-All (AVA) Classification: We build $K(K-1)/2$ different binary classifiers:

- For $k, \ell = 1, ..., K$ with $k \neq \ell$, do
    - let the positive examples be all the points in class k and the negative examples be all points in class ℓ.
    - let $h_{k\ell}$ be the classifier:

$$h_{k\ell}(X) = P(y \in C_k | X) = 1 - P(y \in C_\ell | X)$$

    - Predict the most likely class: the predicted class for a new observation X is $\arg\max_k \sum_\ell h_{k\ell}(X)$

Multiclass Classification

One-vs-All or All-vs-All?: The choice between OVA and AOA is largely computational.

- OVA requires $O(K)$ classifiers
- AVA requires $O(K^2)$ classifiers, but on smaller data sets