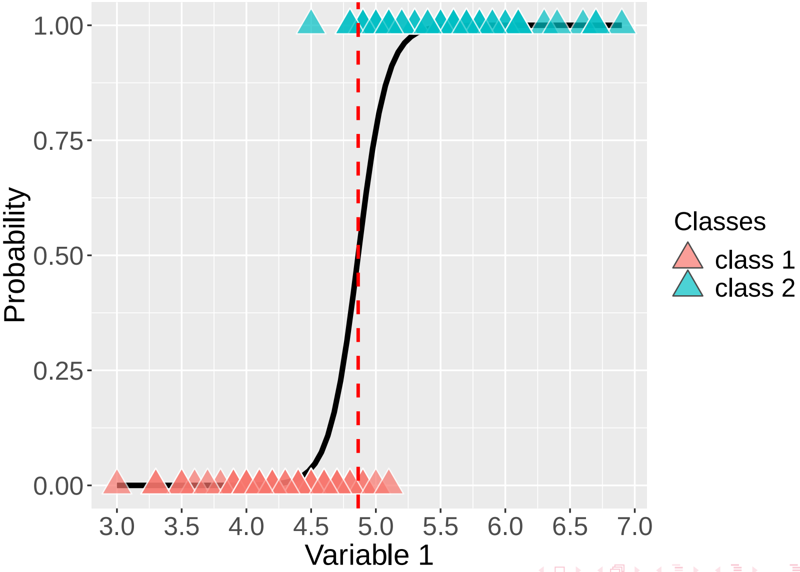
A Dive into Logistic Regression Modeling

Crista Moreno

June 7, 2019

Outline

Goal Build a Good Logistic Regression Model



Structure of Biomedical Data

Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Structure of Biomedical Data

Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Important Numbers

$m_{3,4}$ - 4th variable measurement for the 3rd patient

Structure of Biomedical Data

Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Important Numbers

$m_{3,4}$ - 4th variable measurement for the 3rd patient
N - Total Number of Data Points (number of rows, patients etc.)

Structure of Biomedical Data

Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Important Numbers

$m_{3,4}$ - 4th variable measurement for the 3rd patient
N - Total Number of Data Points (number of rows, patients etc.)
M - Total Number of Variables (measurements, parameters etc.)

Structure of Biomedical Data

Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Important Numbers

$m_{3,4}$ - 4th variable measurement for the 3rd patient
N - Total Number of Data Points (number of rows, patients etc.)
M - Total Number of Variables (measurements, parameters etc.)
C - Total Number of Classes (labels i.e. normal vs diseased)

Which Variables to Include in the Model?

- $N = |\{X^1, X^2, \ldots, X^N\}|$ (data)
- $C = |\{c_1, c_2\}| = 2$ e.g. $\{\text{normal}, \text{diseased}\}$
- $M = |\{v_1, v_2, \ldots, v_M\}|$ e.g. $\{\text{sex}, \text{weight}, \text{blood type}, \text{etc.}\}$

Which Variables to Include in the Model?

- $N = |\{X^1, X^2, \ldots, X^N\}|$ (data)
- $C = |\{c_1, c_2\}| = 2$ e.g. $\{\text{normal}, \text{diseased}\}$
- $M = |\{v_1, v_2, \ldots, v_M\}|$ e.g. $\{\text{sex}, \text{weight}, \text{blood type}, \text{etc.}\}$

Question 1

What is the probability that patient belongs to class c, given that the data X is equal to x?

$$P(Y = c|X = x)$$

Which Variables to Include in the Model?

- $N = |\{X^1, X^2, \ldots, X^N\}|$ (data)
- $C = |\{c_1, c_2\}| = 2$ e.g. $\{\text{normal}, \text{diseased}\}$
- $M = |\{v_1, v_2, \ldots, v_M\}|$ e.g. $\{\text{sex}, \text{weight}, \text{blood type}, \text{etc.}\}$

Question 1

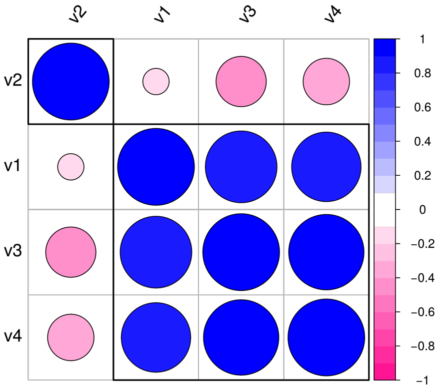
What is the probability that patient belongs to class c, given that the data X is equal to x?

$$P(Y = c | X = x)$$

Question 2

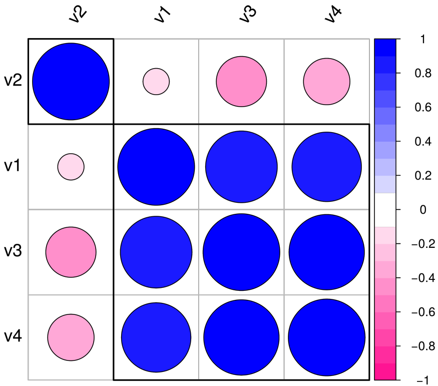
How to determine which of the M variables to use for the model?

Answer Question 2: Correlation Matrix



$$\mathrm{Cor}(x, y) = \frac{\langle x^*, y^* \rangle}{||x^*|| \cdot ||y^*||}, \qquad x^* = x - \bar{x}, y^* = y - \bar{y}$$

Answer Question 2: Correlation Matrix



$$\mathrm{Cor}(x,y) = \frac{\langle x^*, y^* \rangle}{||x^*|| \cdot ||y^*||}, \qquad x^* = x - \bar{x}, y^* = y - \bar{y}$$

v1 and v2 look interesting

Answer Question 1: Logistic Regression Model

$$p(Y = c|X = x)$$

Answer Question 1: Logistic Regression Model

$$p(Y = c | X = x)$$

$$= \mathrm{p(X)} = \frac{e^{\boldsymbol{\beta} \cdot \boldsymbol{X}}}{1 + e^{\boldsymbol{\beta} \cdot \boldsymbol{X}}}$$

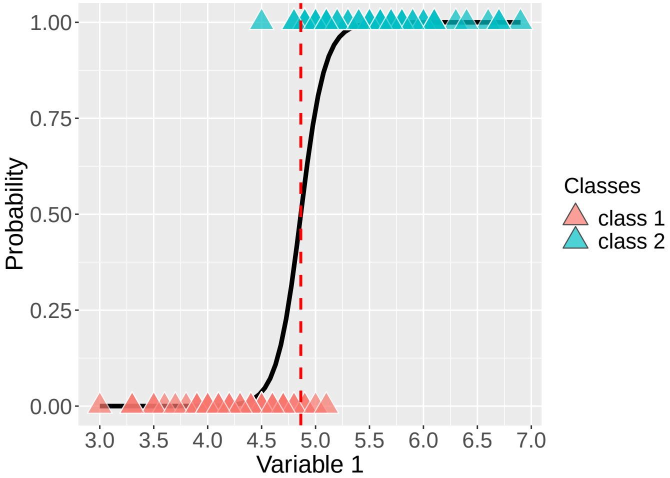
Answer Question 1: Logistic Regression Model

$$p(Y = c|X = x)$$

$$= \text{p(X)} = \frac{e^{\boldsymbol{\beta} \cdot \boldsymbol{X}}}{1 + e^{\boldsymbol{\beta} \cdot \boldsymbol{X}}}$$

$$\text{p(-}\infty) = 0, \qquad p(+\infty) = 1$$

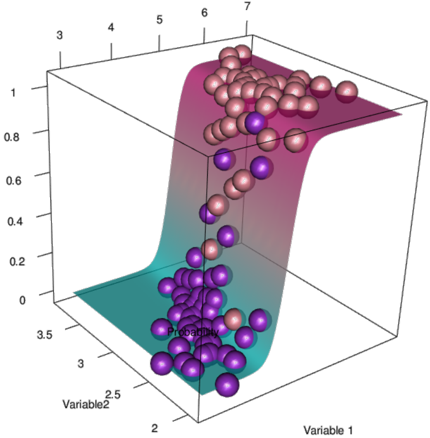
Logistic Regression Model with a Single Variable



$$p(X) = \frac{e^{\beta_o + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_1}}$$

Red dashed line is the decision boundary.

Logistic Regression Model with Multiple Variables



$$p(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_\gamma X_\gamma}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_\gamma X_\gamma}} = \frac{e^{\boldsymbol{\beta}\cdot\boldsymbol{X}}}{1+e^{\boldsymbol{\beta}\cdot\boldsymbol{X}}}$$

Is the Model Good?

Great! We have a model!
Are we done?

Is the Model Good?

Great! We have a model!

Are we done?

Not so fast.

- Adding too many variables $\Rightarrow$ Overfitting & $\uparrow$ Accuracy

Dangers with Overfitting the Model

- Adding too many variables $\Rightarrow$ Overfitting & $\uparrow$ Accuracy
- $\uparrow$ Accuracy $\Rightarrow$ $\uparrow$ Sensitivity and $\uparrow$ Specificity

Dangers with Overfitting the Model

- Adding too many variables $\Rightarrow$ Overfitting & $\uparrow$ Accuracy
- $\uparrow$ Accuracy $\Rightarrow$ $\uparrow$ Sensitivity and $\uparrow$ Specificity
- $\uparrow$ Sensitivity and $\uparrow$ Specificity $\Rightarrow$ Misleading Model

Dangers with Overfitting the Model

- Adding too many variables $\Rightarrow$ Overfitting & $\uparrow$ Accuracy
- $\uparrow$ Accuracy $\Rightarrow$ $\uparrow$ Sensitivity and $\uparrow$ Specificity
- $\uparrow$ Sensitivity and $\uparrow$ Specificity $\Rightarrow$ Misleading Model

The simplest model that fits the data is also the most plausible.(Occam's Razor)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Sensitivity} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}$$

Validation

How do we validate our model?

Validating a Model: Back to Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Validating a Model: Back to Data

Patient	Variable 1	Variable 2	$\cdots$	Variable M	Class
1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,M}$	normal
2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,M}$	sick
3	$m_{3,1}$	$m_{3,2}$	$\cdots$	$m_{3,M}$	normal
4	$m_{4,1}$	$m_{4,2}$	$\cdots$	$m_{4,M}$	normal
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
N	$m_{N,1}$	$m_{N,2}$	$\cdots$	$m_{N,M}$	sick

Data $\mathcal{D}$ for Model

Data	Observation	Input (Variables)	Output (Class)
$\mathcal{D}_1$	1	X^1	Y^1
$\mathcal{D}_2$	2	X^2	Y^2
$\mathcal{D}_3$	3	X^3	Y^3
$\mathcal{D}_4$	4	X^4	Y^4
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{D}_N$	N	X^N	Y^N

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \ldots, \mathcal{D}_N\}$$

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

Classic Approach: Partition Data for Validating the Model

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

- Partition the data $\mathcal{D}$ into a Training set and Test set

$$\underbrace{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7}_{\text{Training Set}}, \underbrace{\mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}}_{\text{Test Set}}$$

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

- Partition the data $\mathcal{D}$ into a Training set and Test set

$$\underbrace{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7}_{\text{Training Set}}, \underbrace{\mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}}_{\text{Test Set}}$$

- Use Training set to build a model

Classic Approach: Partition Data for Validating the Model

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

- Partition the data $\mathcal{D}$ into a Training set and Test set

$$\underbrace{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7}_{\text{Training Set}}, \underbrace{\mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}}_{\text{Test Set}}$$

- Use Training set to build a model
- Test model on Test set to get the model's performance.

Classic Approach: Partition Data for Validating the Model

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

- Partition the data $\mathcal{D}$ into a Training set and Test set

$$\underbrace{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7}_{\text{Training Set}}, \underbrace{\mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}}_{\text{Test Set}}$$

- Use Training set to build a model
- Test model on Test set to get the model's performance.

Now we have the model's performance, are we done?

- Given the data for modeling

$$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}\}$$

- Partition the data $\mathcal{D}$ into a Training set and Test set

$$\underbrace{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7}_{\text{Training Set}}, \underbrace{\mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}}_{\text{Test Set}}$$

- Use Training set to build a model
- Test model on Test set to get the model's performance.

Now we have the model's performance, are we done?
Classic approach works well for large data sets.

k-fold Cross Validation Works for Smaller Data Sets

k-fold Cross Validation Works for Smaller Data Sets

- Select k e.g. $k = 5$ folds

k-fold Cross Validation Works for Smaller Data Sets

- Select k e.g. $k = 5$ folds
- Partition data into a Training Set and Validation Set in the following fashion:

k-fold Cross Validation Works for Smaller Data Sets

- Select k e.g. $k = 5$ folds
- Partition data into a Training Set and Validation Set in the following fashion:

fold 1	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 2	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 3	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 4	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 5	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$

- Select k e.g. $k = 5$ folds
- Partition data into a Training Set and Validation Set in the following fashion:

fold 1	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 2	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 3	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 4	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 5	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$

- For each fold train model on the Training Set and get performance on Validation Set

k-fold Cross Validation Works for Smaller Data Sets

- Select k e.g. $k = 5$ folds

- Partition data into a Training Set and Validation Set in the following fashion:

fold 1	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 2	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 3	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 4	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$
fold 5	$\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5, \mathcal{D}_6, \mathcal{D}_7, \mathcal{D}_8, \mathcal{D}_9, \mathcal{D}_{10}$

- For each fold train model on the Training Set and get performance on Validation Set
- The variance and average of the performance helps indicate how well this model can make predictions on future data.

Software

- **R** - A **free software** (GNU Affero GPL) environment for statistical computing and graphics.
- **RStudio** is a free and open-source integrated development environment (IDE) for R.