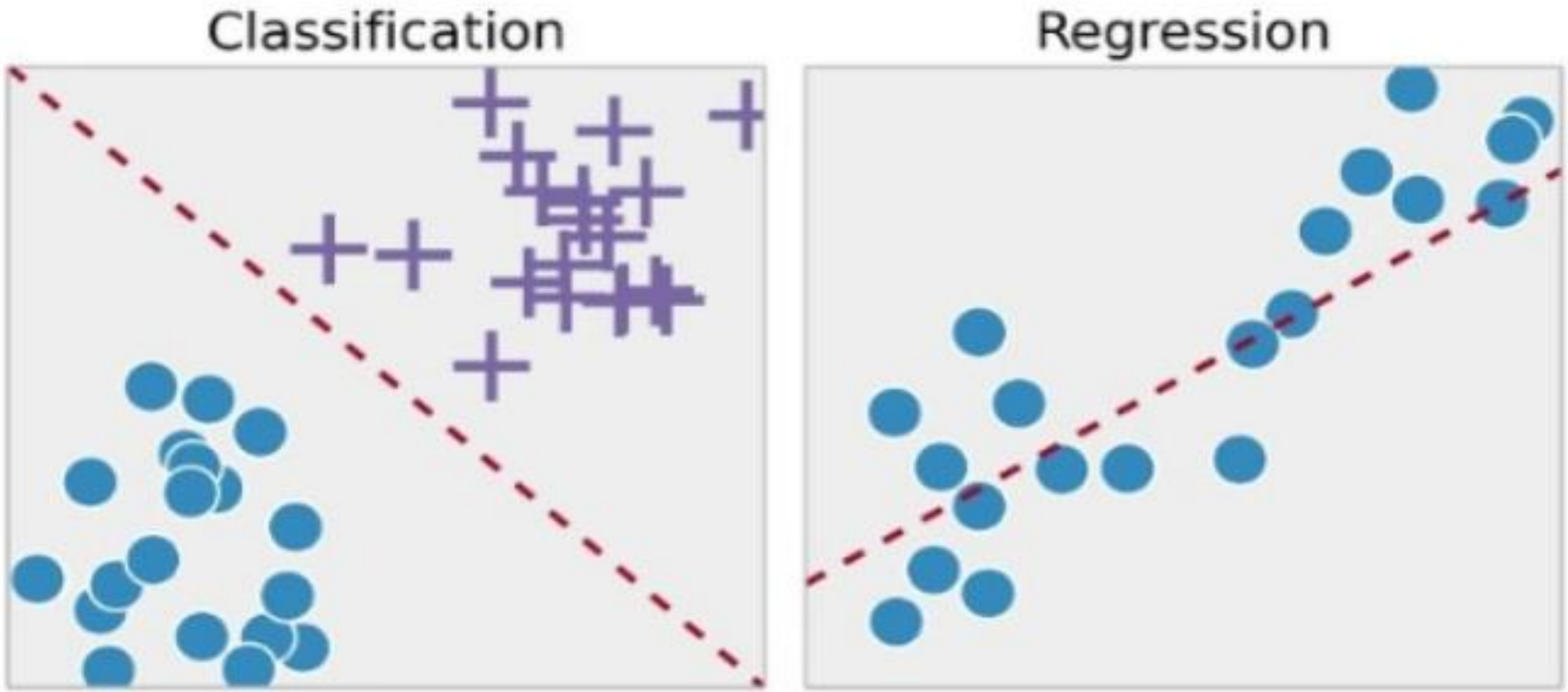
Logistic Regression

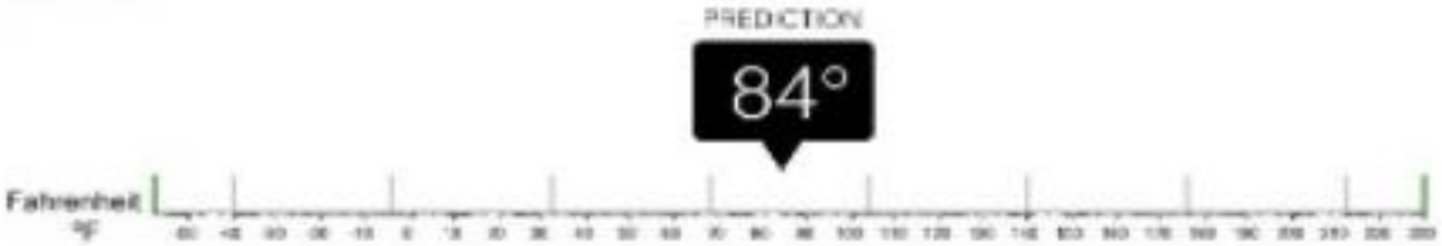
Regression
v/s
Classification


Classification
Regression

- Classification
  - **Output type:** discrete
  - **Trying to find:** a boundary
  - **Evaluation:** accuracy
- Regression
  - **Output type:** continuous
  - **Trying to find:** best fit line
  - **Evaluation:** sum of sqaured errors

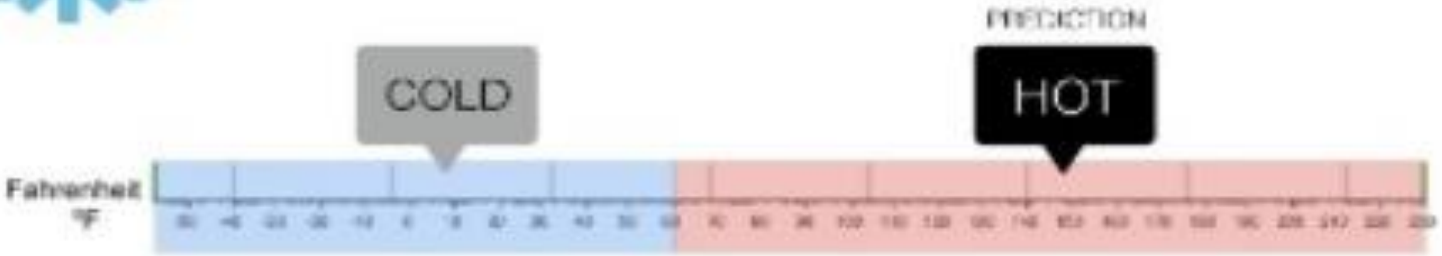
Regression

What is the temperature going to be tomorrow?

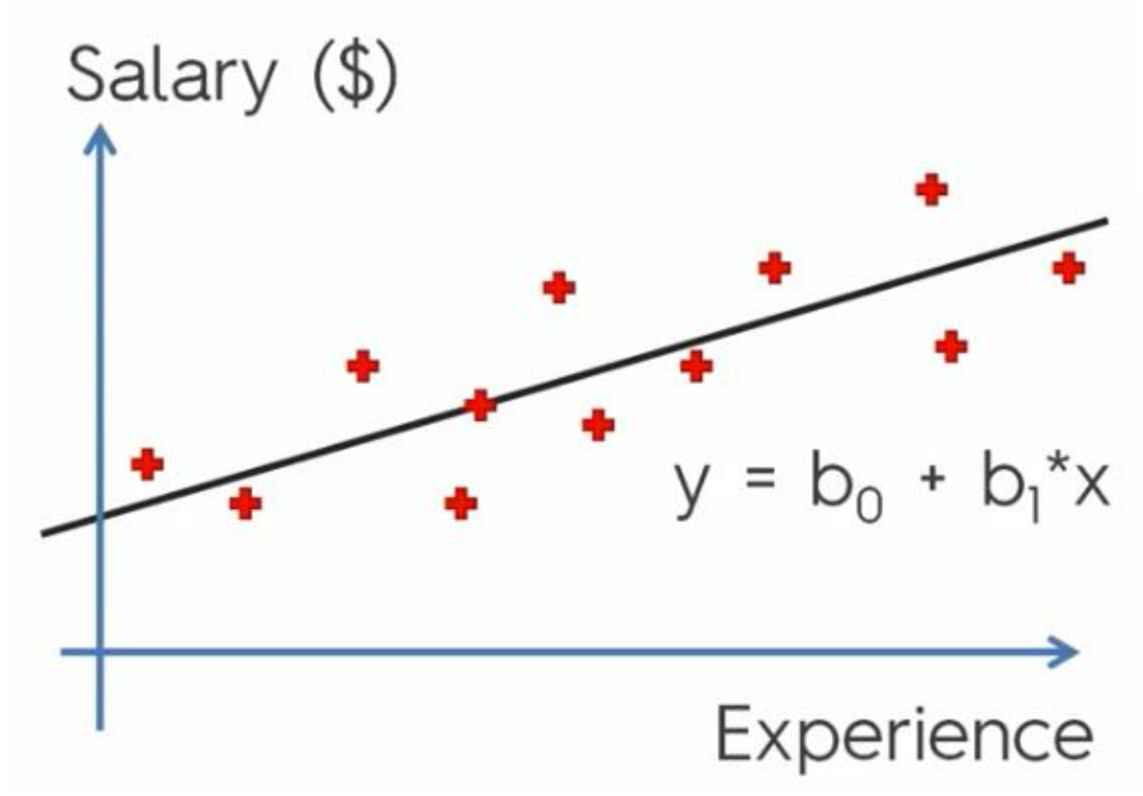


Classification

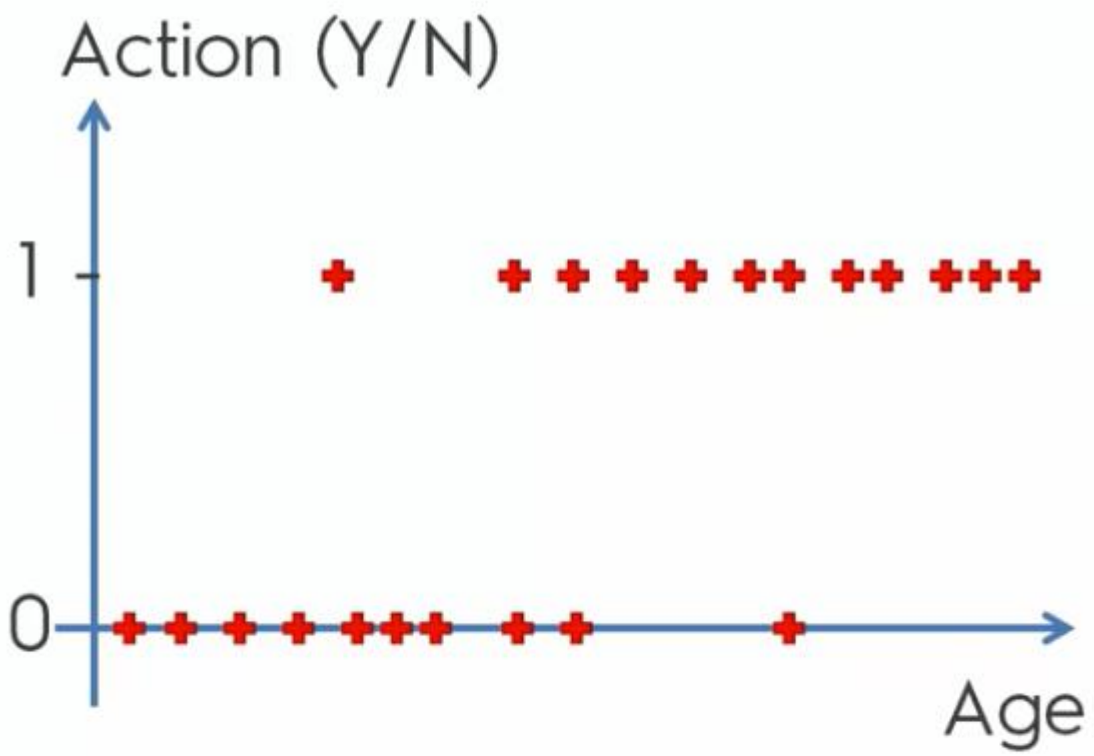
Will it be Cold or Hot tomorrow?



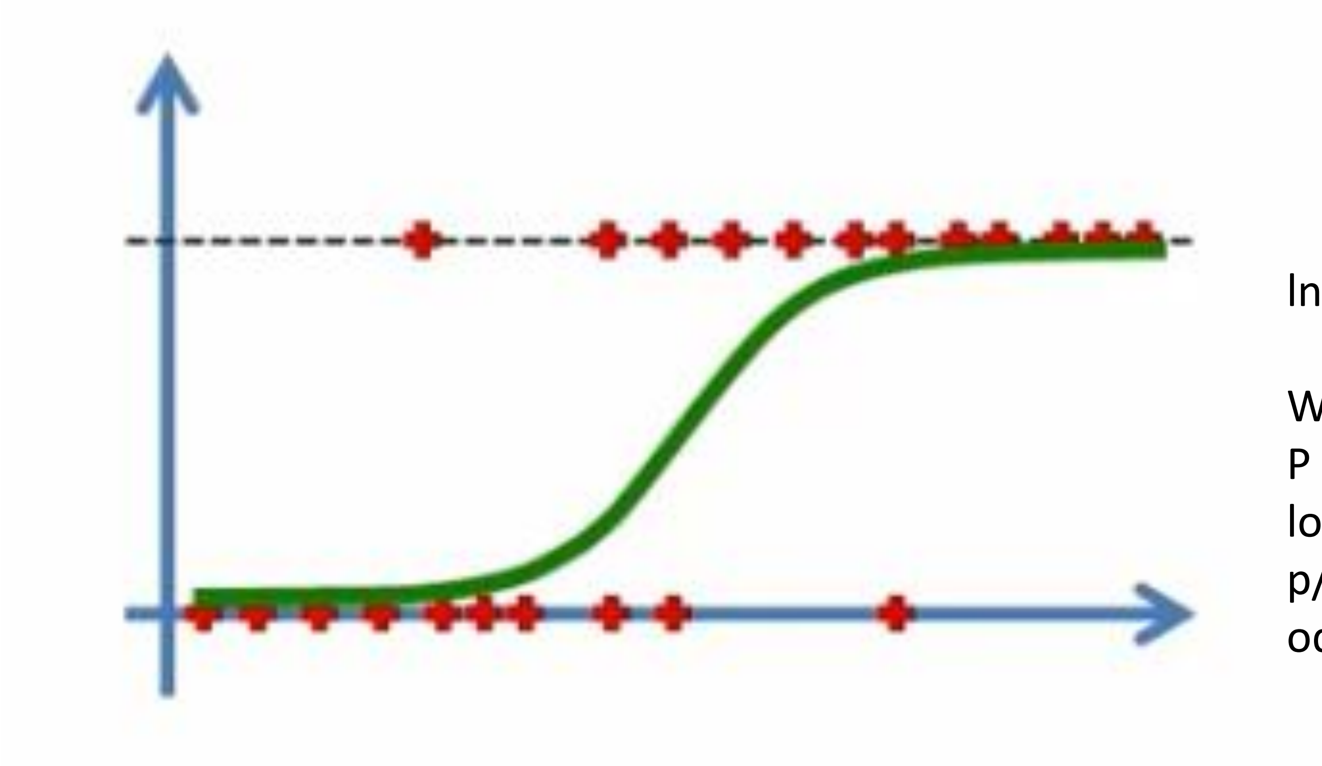
Logistic Regression - Intuition



Logistic Regression - Intuition



Logistic Regression - Intuition



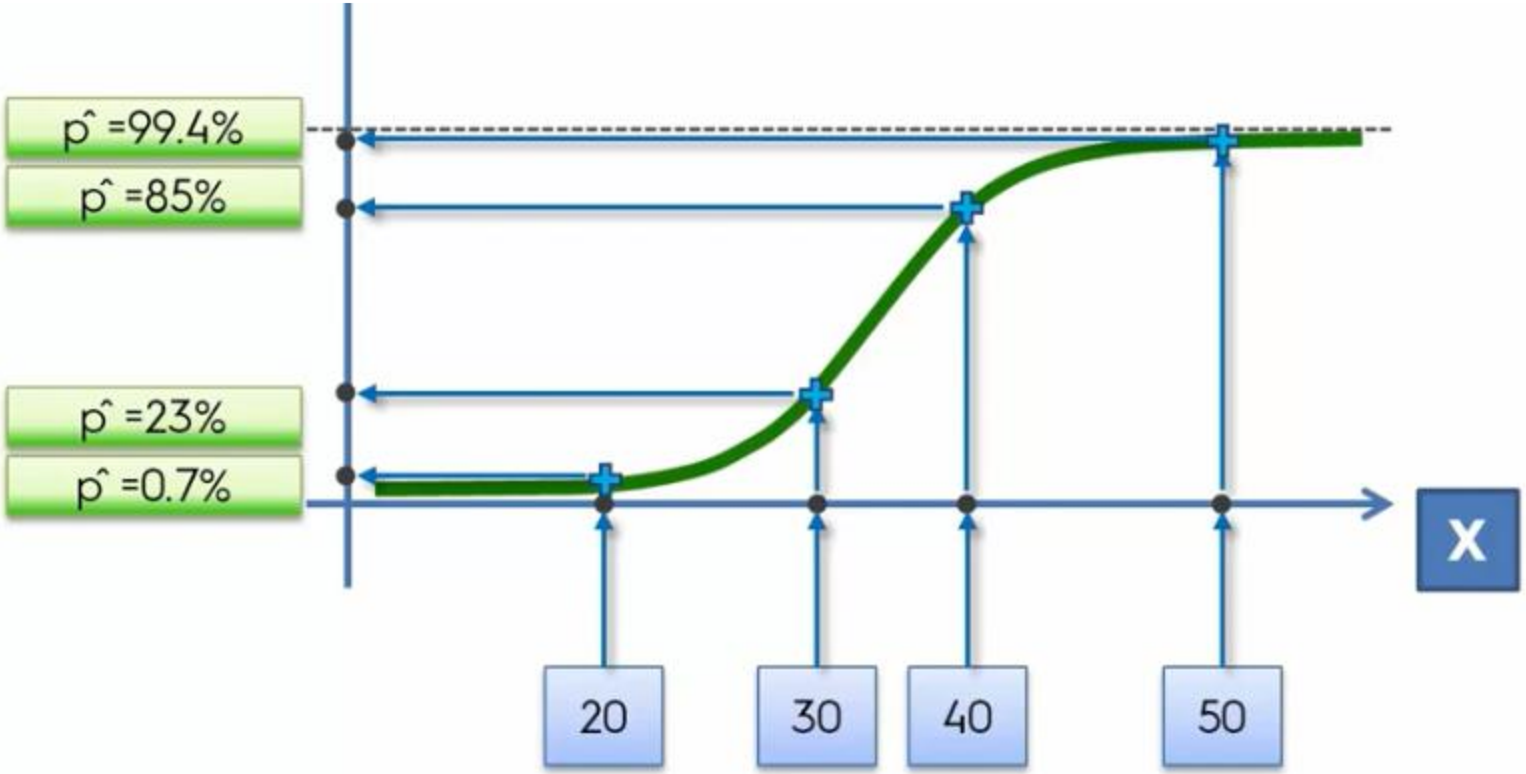
$y = b0 + b1*x$

Sigmoid Transformation

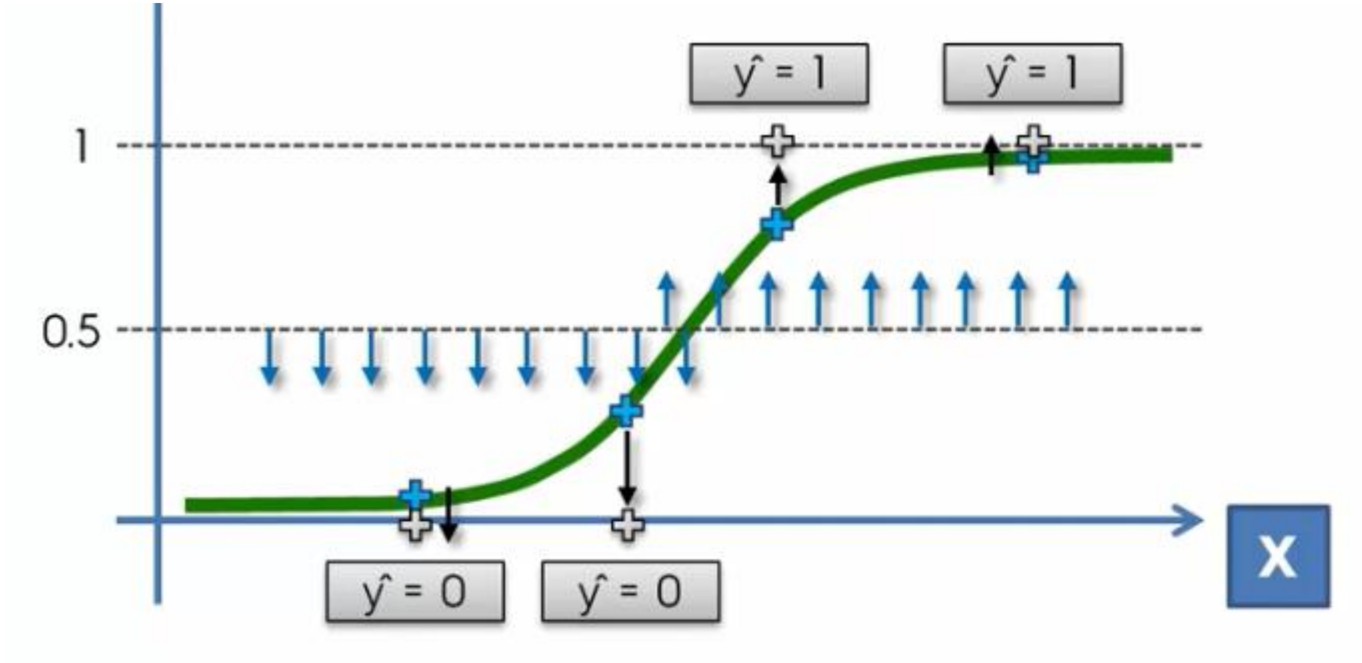
$\ln(p/(1-p)) = b0 + b1*x$,

Where
$P = 1/(1+e^{-y})$ is the logit/sigmoid function.
$p/(1-p)$ is called the odds ratio

Logistic Regression - Intuition



Logistic Regression - Intuition



Logistic Regression - Intuition

- $P(y = 1|x) = y^{-hat}$
- $P(y = 0|x) = (1-y^{-hat})$
- This can be written compactly as $P(y|x) = y^{-hat}$^y * $(1-y^{-hat})$^(1-y)
- This is called likelihood.
- If we take the log of likelihood, we get

$$Log(y^{-hat}, y) = y*log(y^{-hat}) + (1-y)*log(1-y^{-hat})$$

- This is the cost function for logistic regression. The objective is to maximize this log likelihood for all the inputs variables

More on Logistic Regression

- Belongs to the family of Generalized Linear Models

More on Logistic Regression

- What are Generalized Linear Models?
- GLMs are an extension of the linear model framework, which includes dependent variables which are non-normal. In general, they possess three characteristics:
  - These models comprise a linear combination of input features.
  - The mean of the response variable is related to the linear combination of input features via a link function.
  - The response variable is considered to have an underlying probability distribution belonging to the family of exponential distributions such as binomial distribution, Poisson distribution, or Gaussian distribution.

More on Logistic Regression

- Following are the assumptions made by Logistic Regression:
  - The response variable must follow a binomial distribution.
  - Logistic Regression assumes a linear relationship between the independent variables and the link function (logit).
  - The dependent variable should have mutually exclusive and exhaustive categories
  - There should be little or no multicollinearity among the independent variables.

Lets consider an example…

- Social_Network_Ads.csv

Lets understand the output....

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output….

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output….

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output....

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output....

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output….

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand the output....

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

AIC = -2LL+2k
LL = Log Likelihood of the model
(max of the cost function),
k = Number of parameters

Lets understand the output....

```
Call:
glm(formula = Purchased ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0753  -0.5235  -0.1161   0.3224   2.3977

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1923     0.2018  -5.908 3.47e-09 ***
Age               2.6324     0.3461   7.606 2.83e-14 ***
EstimatedSalary   1.3947     0.2326   5.996 2.03e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 390.89  on 299  degrees of freedom
Residual deviance: 199.78  on 297  degrees of freedom
AIC: 205.78

Number of Fisher Scoring iterations: 6
```

Lets understand Confusion Matrix in details…

```
> cm

   FALSE TRUE
0     57    7
1     10   26
```

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

Lets understand Confusion Matrix in details…

Predicted / Actual	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Actual \ Predicted	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Actual \ Predicted	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Predicted / Actual	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Predicted / Actual	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Predicted / Actual	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Lets understand Confusion Matrix in details…

Predicted / Actual	1 (+ve)	0 (-ve)
1	TP	FN
0	FP	TN

Why not just use accuracy?

Lets consider a spam classifier, where we have 10 spams in the test data

Non-spam / Spam	1 (+ve)	0 (-ve)
1	*5*	5
0	10	90

Lets consider a cancer patients classifier, where we have 10 cancer patients in the test data

No Cancer / Cancer	1 (+ve)	0 (-ve)
1	5	*5*
0	10	90

Why not just use accuracy?

What is the precision & recall in these cases?

Non-spam / Spam	1 (+ve)	0 (-ve)
1	*5*	5
0	10	90

No Cancer / Cancer	1 (+ve)	0 (-ve)
1	5	*5*
0	10	90

Which terms to remember? And How??

If nothing else, just remember Precision & Recall!!

Precision = TP/(Predicted Positive)
Recall = TP/(Real Positive)

Dealing with Imbalanced Data

- Choose appropriate threshold using ROC-AUC
- Balance the data and stick to 0.5 threshold

ROC Curves

- One of the problems with the confusion matrix is that all the values that are populated is based on an arbitrary choice of the threshold
- Lets reconsider the Cancer patients example. Do you think a 50% threshold is justified in this case?
- What if we could evaluate multiple thresholds and see their impact? This is how ROC curves are generated
- ROC curves also help in deciding the threshold when you have unbalanced data with you

ROC Curves

- Lets understand more about ROC curves
- http://www.navan.name/roc/
- https://youtu.be/OAl6eAyP-yo

Other ways of dealing with Unbalanced Data

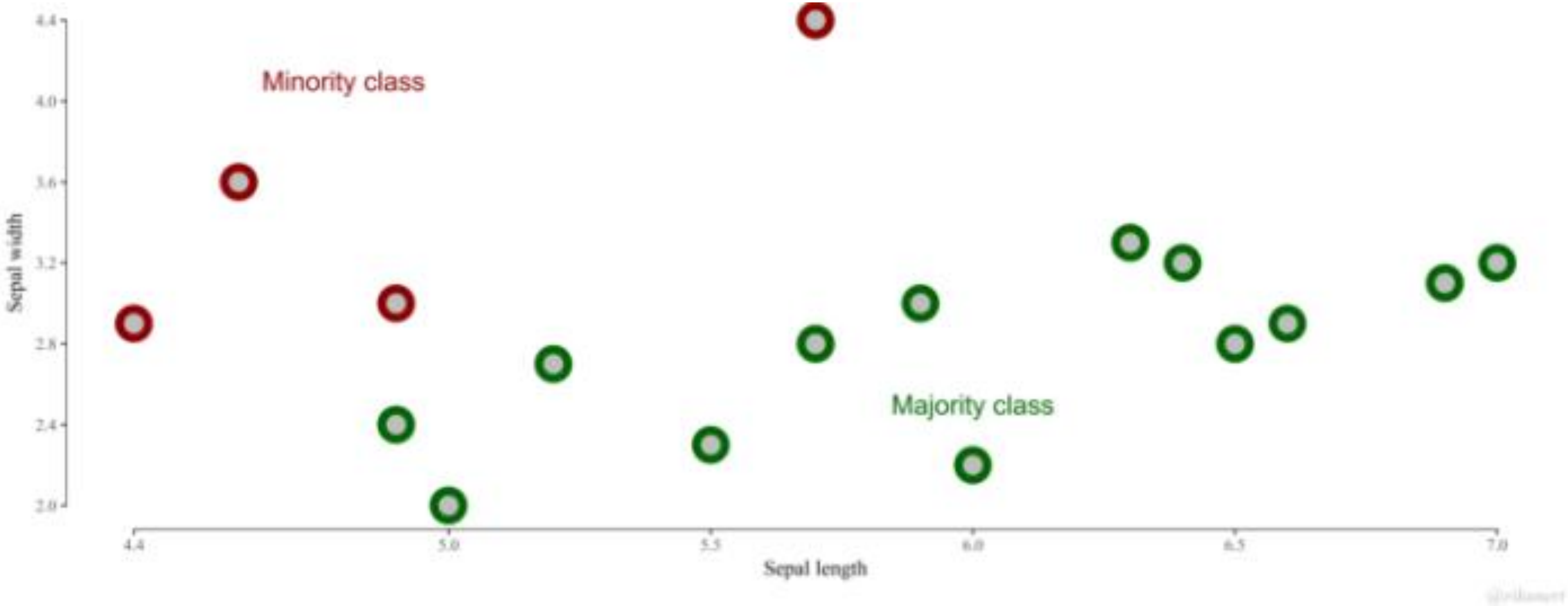
Sampling Techniques:

- Under sampling
- Over sampling
- Under & Over Sampling
- Synthetic Data Generation
  - SMOTE (Synthetic Minority Over-sampling Technique)
  - ROSE (Random Over Sampling Examples)

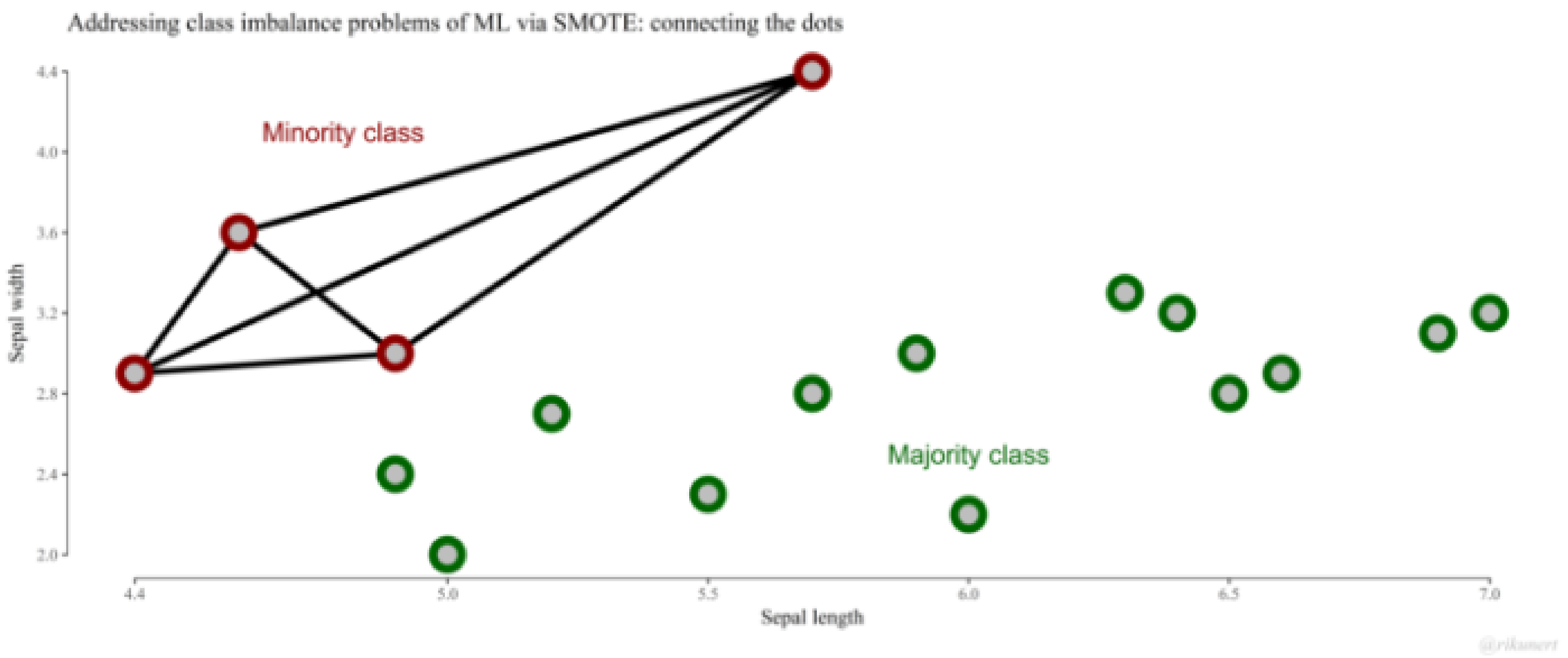
SMOTE & ROSE

- ROSE uses smoothed bootstrapping to draw artificial samples from the feature space neighborhood around the minority class
- SMOTE draws artificial samples by choosing points that lie on the line connecting the rare observation to one of its nearest neighbors in the feature space
- https://www.openstarts.units.it/bitstream/10077/4002/1/Menardi%20Torelli%20DEAMS%20WPS2.pdf

SMOTE

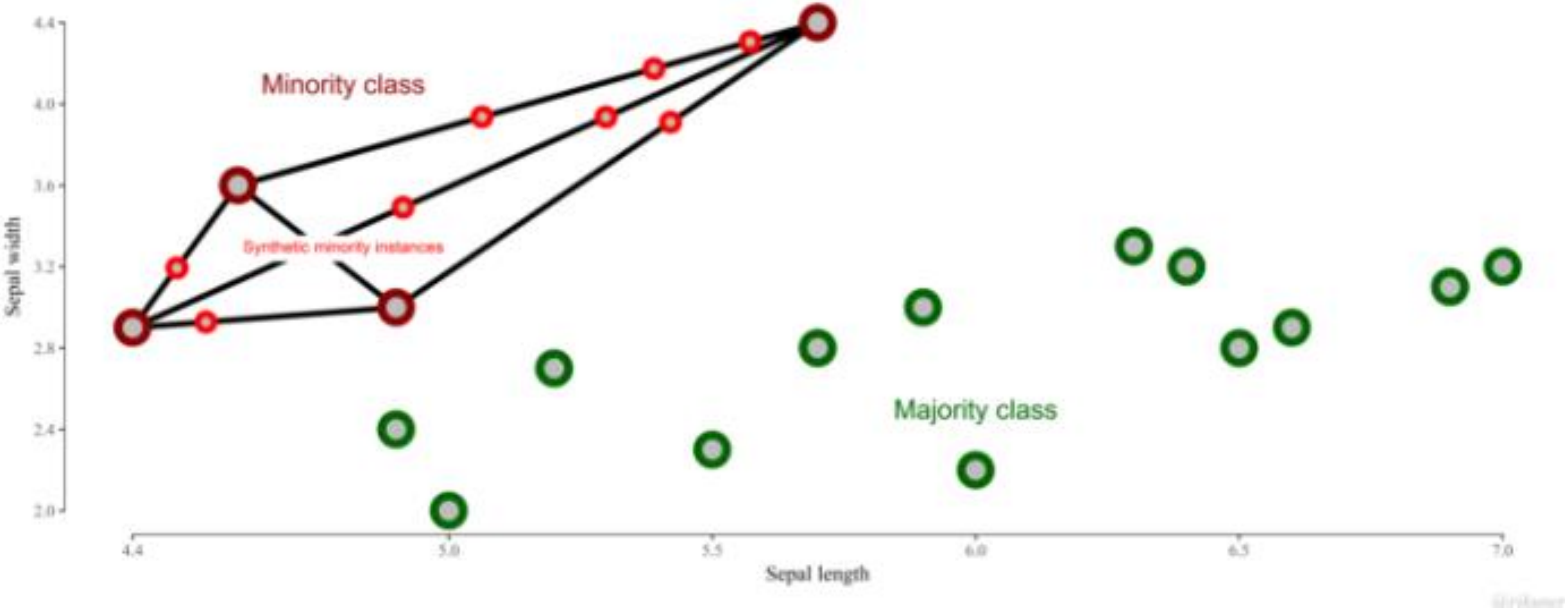


SMOTE



Addressing class imbalance problems of ML via SMOTE: connecting the dots
Minority class
Majority class
Sepal width
Sepal length
4.4
4.0
3.6
3.2
2.8
2.4
2.0
4.4
5.0
5.5
6.0
6.5
7.0

SMOTE



Minority class
Synthetic minority instances
Majority class
Sepal width
Sepal length
4.4
4.0
3.6
3.2
2.8
2.4
2.0
4.4
5.0
5.5
6.0
6.5
7.0