Intro to Data Science with Python – Part 2

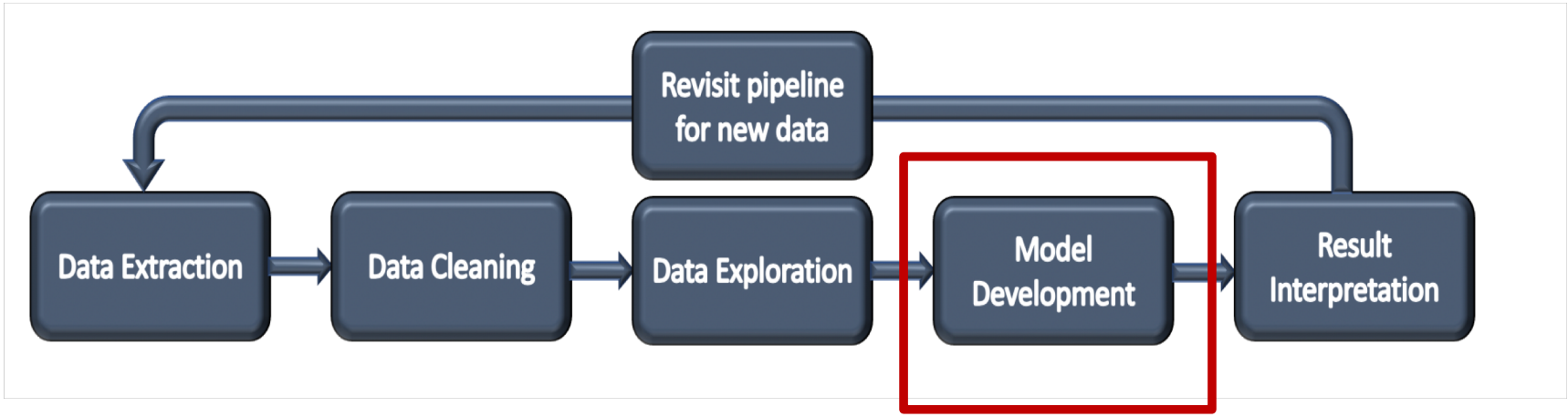
IBM **Developer**

Saishruthi Swaminathan| Developer Advocate | IBM San Francisco
Simon Plovyt | Developer Advocate | IBM San Francisco

Agenda

1. Types of machine learning algorithms
2. Metrics for evaluating the algorithms and error terms
3. Supervised ML algorithms at a glance + demo
4. Imbalance data handling
5. Tuning and Cross validation
6. Recap
7. Workshop

Data Science Pipeline



What is Machine Learning?

Who is this?????

Breaking down machine learning definition

"Machine learning is the science (and art) of programming computers so they can learn from data," writes Aurélien Géron in *Hands-on Machine Learning with Scikit-Learn and TensorFlow*.

ML is a subset of the larger field of artificial intelligence (AI) that "focuses on teaching computers how to learn without the need to be programmed for specific tasks," note Sujit Pal and Antonio Gulli in *Deep Learning with Keras*. "In fact, the key idea behind ML is that it is possible to create algorithms that learn from and make predictions on data."

Source: https://www.oreilly.com/ideas/machine-learning-a-quick-and-simple-definition

Terms

Dataset

Feature, field, variable, attribute or characteristics

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	Tab.
2	9	1	AA	Table
3	5	7	AA	Chair
4	6	9	AA	Chair
5	10	9	AA	Chair
6	20	40	AA	Dining
7	25	45	AA	Dining

Data point, record, sample, entity or instance

Here,

- 'Type' and 'Dealer' are **categorical variables** as they have finite number of distinct groups.
- 'X0' and 'x1' are **continuous variables**.

Machine learning categories

1. Supervised Learning

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	Tab.
2	9	1	AA	Table
3	5	7	AA	Chair
4	6	9	AA	Chair
5	10	9	AA	Chair
6	20	40	AA	Dining
7	25	45	AA	Dining

Learning from the previous data which has both predictors (attributes) and response (target).

Machine learning categories

2. Unsupervised Learning

	x0	x1	Dealer
0	5	1	AA
1	8	3	AA
2	9	1	AA
3	5	7	AA
4	6	9	AA
5	10	9	AA
6	20	40	AA
7	25	45	AA

The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together.

Machine learning categories

3. Semi - supervised Learning

	x0	x1	Dealer	Type
0	5	1	AA	Table
1	8	3	AA	[redacted]
2	9	1	AA	[redacted]
3	5	7	AA	[redacted]
4	6	9	AA	Chair
5	10	9	AA	[redacted]
6	20	40	AA	Dining
7	25	45	AA	[redacted]

Training the system on partially labeled input data. Input data has a lot of unlabeled data and little of labeled data.

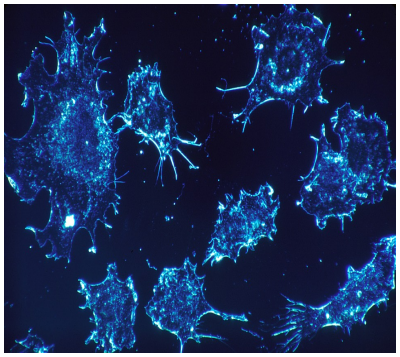
Face recognition in social media

Explanation: https://medium.com/@jrodthoughts/understanding-semi-supervised-learning-a6437c070c87

Machine learning categories

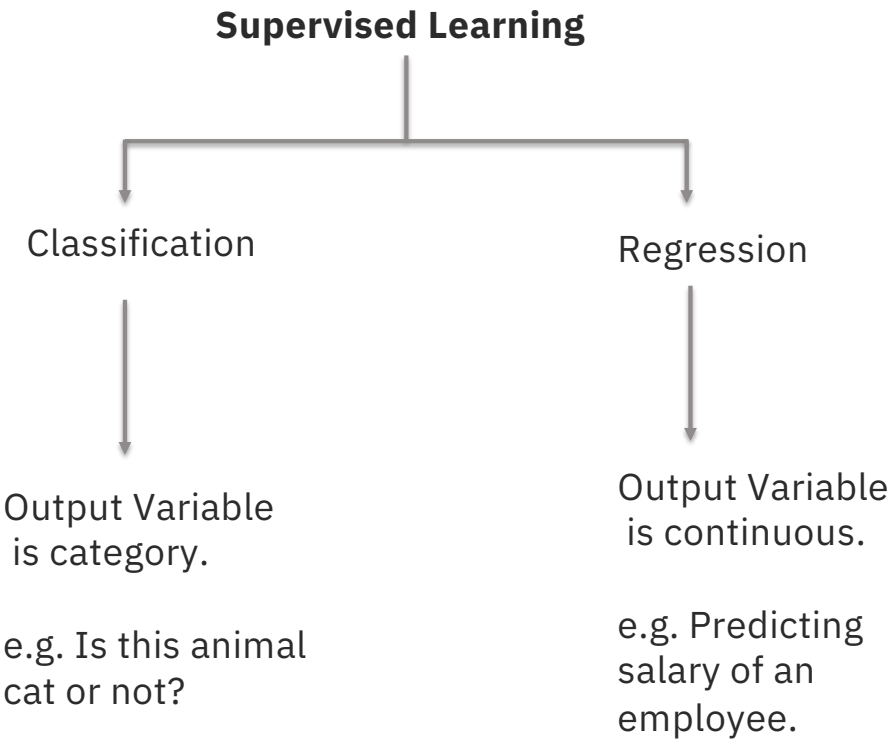
4. Transfer Learning

Reusing a model that was trained for solving one problem and applying it to a different but related problem.



Cat: CCO License | https://www.pexels.com/photo/animal-pet-cute-kitten-45201/
Cancer cells: Pixabay License | https://pixabay.com/photos/cancer-cells-cells-scan-541954/

Algorithms for the day!!



Source: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Algorithms for the day!!!

Classification

1. Logistic regression
2. Decision tree
3. K-Nearest Neighbor algorithm
4. Random Forest
5. Boosting and Adaboost

Regression

1. Linear regression

Metrics for classification

		Predicted Class	
		Yes	No
Actual class	Yes	True Positive (TP)	False Negative(FN)
	No	False Positive (FP)	True Negative (TN)

1. Accuracy

$$\frac{TP + TN}{TP + FN + FP + TN}$$

2. Precision

$$\frac{TP}{TP + FP}$$

What proportion of positive identifications was actually correct?

3. Recall

$$\frac{TP}{TP + FN}$$

What proportion of actual positives was identified correctly?

4. F1-score

$$\frac{2 * \text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

Classification metric example

Actual	Predicted
0	0
0	1
1	1
1	0
0	0
1	0

	Predicted Class		
Actual class		Yes	No
	Yes	1 (TP)	2(FN)
	No	1 (FP)	2 (TN)

Accuracy = (TP + TN) / (TP + FN + FP + TN) = 3/6 = 0.5

Precision = (TP) / (TP + FP) = 1 /2 = 0.5

Recall = (TP) / (TP + FN) = 1/3 = 0.33

F1 - Score = 2 *((precision * recall)/(precision + recall))
= 2 * ((0.5 * 0.33)/(0.5 + 0.33))
= 0.40 (approx.)

Regression metrics

1. Root Mean Squared Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

2. Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

3. R-Squared (R2)

$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Error terms

1. Irreducible Error

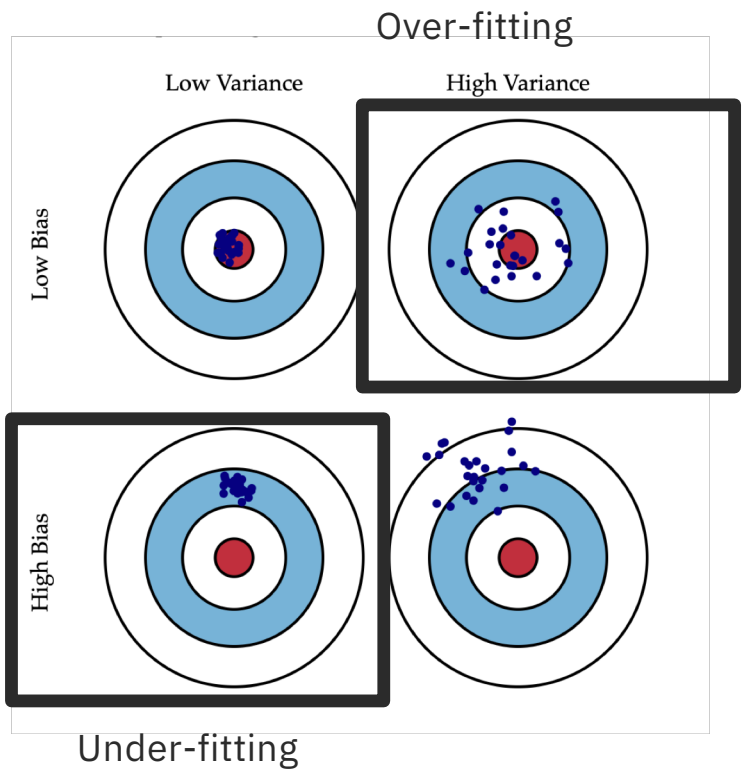
An error that cannot be reduced regardless of what algorithm is being applied.

2. Bias

Bias measures how far off in general these models' predictions are from the correct value.

3. Variance

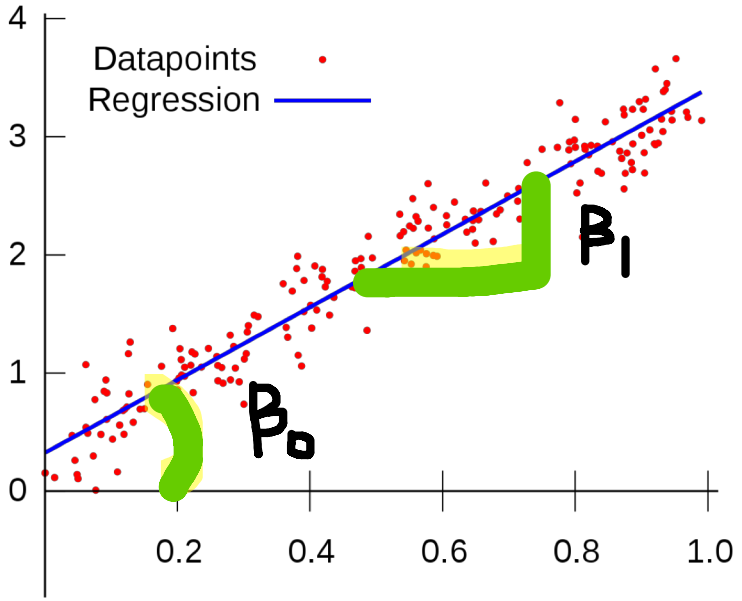
Algorithm sensitivity to specific kind of dataset.



https://elitedatascience.com/bias-variance-tradeoff and http://scott.fortmann-roe.com/docs/BiasVariance.html

Algorithm overview – Simple Linear Regression

Line of best fit



https://commons.wikimedia.org/wiki/File:LinearRegression.svg

- Assumes predictors (attributes) and target are having linear relationship.

$$Y = \beta_0 + \beta_1 X$$

Here,
X --- > Independent variable
Y --- > Dependent variable
β0 -----> intercept and β1 ----- > Slope

- All the terms are either constant or a parameter multiplied by an independent variable.
- In simple linear regression, we have one predictor and one target.
- Best fit line is the one in which the total error from predictions of all data points are as small as possible.

Linear Regression example

$$Y = \beta_0 + \beta_1 X$$

X	Y
2	1
3	2
4	7
5	9

$$\beta_0 = \frac{1}{n}(\sum y_i - \beta_1 \sum x_i)$$

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$\beta_1 \Rightarrow \quad n\Sigma x_i y_i = (2\times1) + (3\times2) + (4\times7) + (5\times9)$

$= 2 + 6 + 28 + 45$

$\Sigma x_i y_i = 81 \quad \Rightarrow \quad n\Sigma x_i y_i = 4\times81 = \boxed{324}$

$\Sigma x_i \Sigma y_i = (2+3+4+5) * (1+2+7+9) = \boxed{266}$

$n\Sigma x_i^2 = 4(4+9+16+25) = \boxed{216}$

$(\Sigma x_i)^2 = (14)^2 = \boxed{196}$

$\beta_1 = 2.9$

$\beta_0 = \frac{1}{n}(\Sigma y_i - \beta_1 \Sigma x_i) = \boxed{-4.675}$

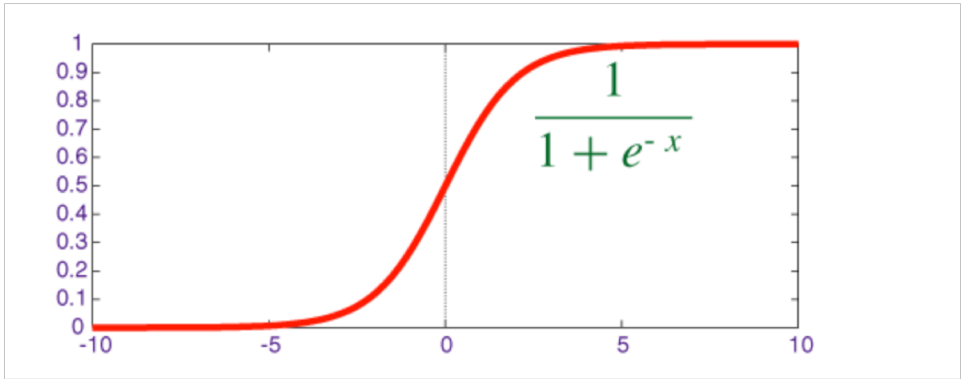
Y = -4.675 + (2.9)X

Logistic Regression

- This is a classification algorithm.

$$Y = \text{Sigmoid}\ (\beta_0 + \beta_1 X)$$

- Logistic regression is like linear regression in the goal of finding the co-efficient values of each attribute.
- The output is transformed using a non-linear function called the logistic (sigmoid) function.

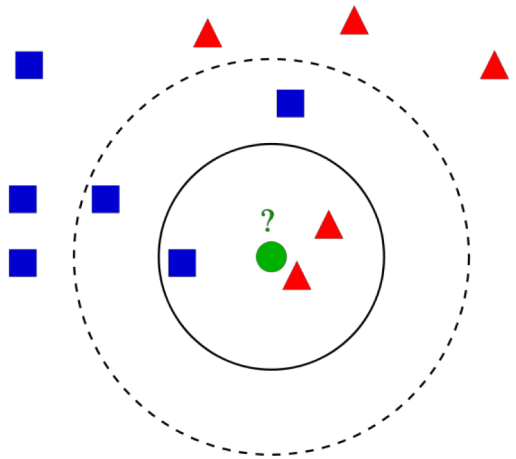


$P \geq 0.5$, class = 1

$P < 0.5$, class = 0

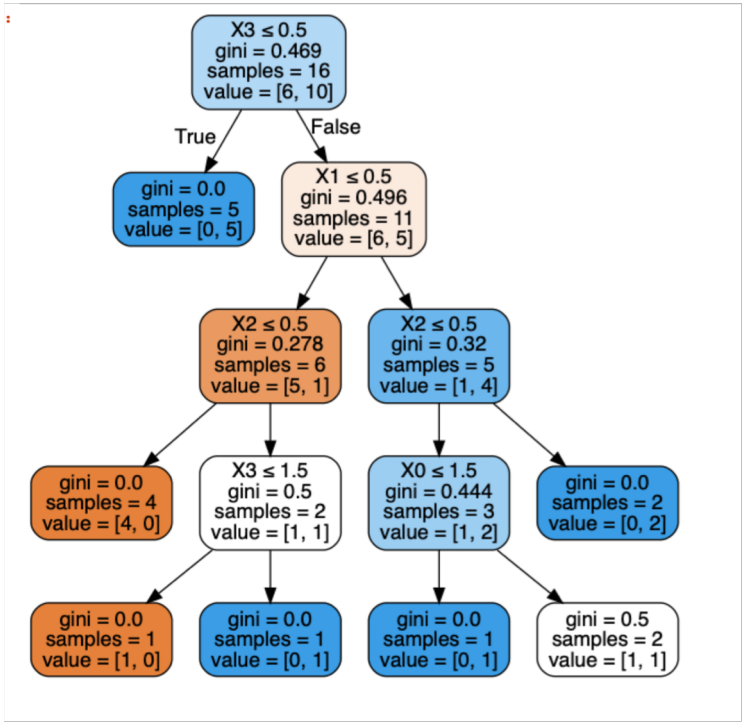
K-Nearest Neighbor

- Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) .
- For regression problems, this might be the mean output variable.
- For classification problems this might be the mode (or most common) class value.
- If all the attributes are in the same scale, the similarity between the data instances can be calculated using the Euclidean distance.
- Cosine similarity can also be used.



Decision Tree

Age	Has_Job	Own_house	Credit_rating	Class
young	no	no	fair	no
young	no	no	good	no
young	yes	no	good	yes
young	yes	yes	fair	yes
young	no	no	fair	no
middle	no	no	fair	no
middle	no	no	good	no
middle	yes	yes	good	yes
middle	no	yes	excellent	yes
middle	no	yes	excellent	yes
old	no	yes	excellent	yes
old	no	yes	good	yes
old	yes	no	good	yes
old	yes	no	excellent	yes
old	no	no	fair	no
old	no	no	excellent	yes
young	yes	no	good	no
young	no	no	good	no
middle	yes	no	excellent	yes
middle	no	yes	fair	no

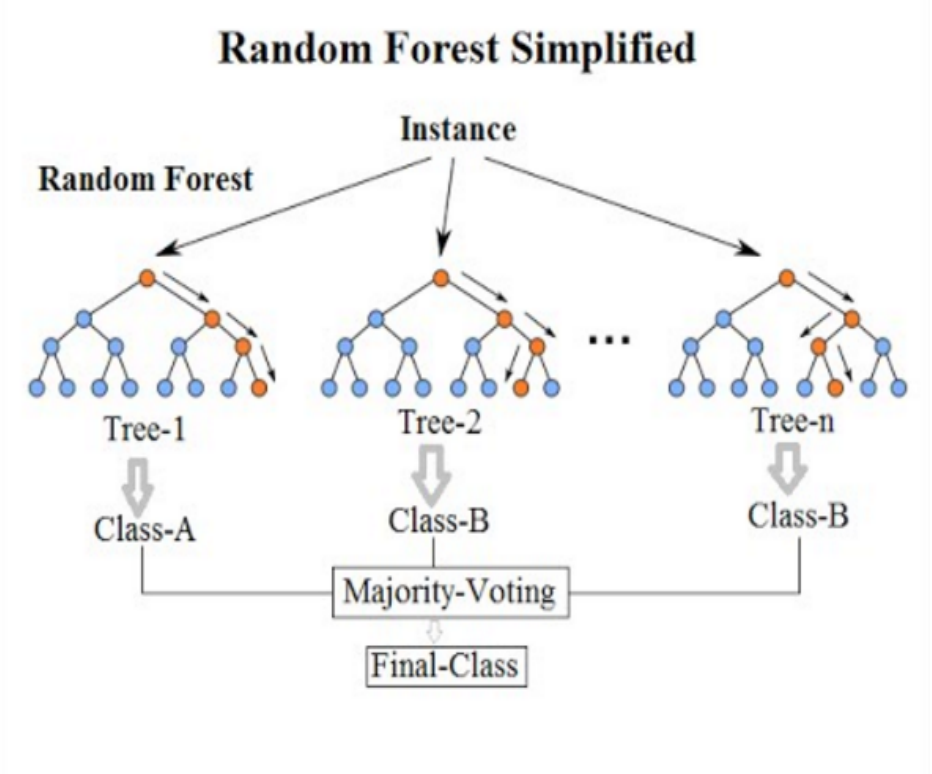


http://sirius.cs.put.poznan.pl/~inf89721/Seminarium/Web_Data_Mining__2nd_Edition__Exploring_Hyperlinks__Contents__and_Usage_Data.pdf

Bagging and Random Forest

Bootstrap Aggregating = Bagging

1. Random samples are drawn from the dataset with replacement and models are built.
2. Predict an unlabeled instance using all classifiers and return the most frequently predicted class as the prediction.
3. Reduce variance and increases test accuracy. (**NOTE: All features are used**)
4. Tweaking bagging process on decision trees will give random forest algorithm.
5. Only 'm' features are selected from the 'k' available features. Each classifier will create a different tree with different top split attribute.



https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png

Boosting and Adaboost

- Produce a sequence of classifiers and each one is dependent on the other.
- Weights of the wrongly classified samples are given more weights. This will let these samples to be selected again.
- Final result is the combination of all these classifiers with their respective weights.

Imbalance dataset handling

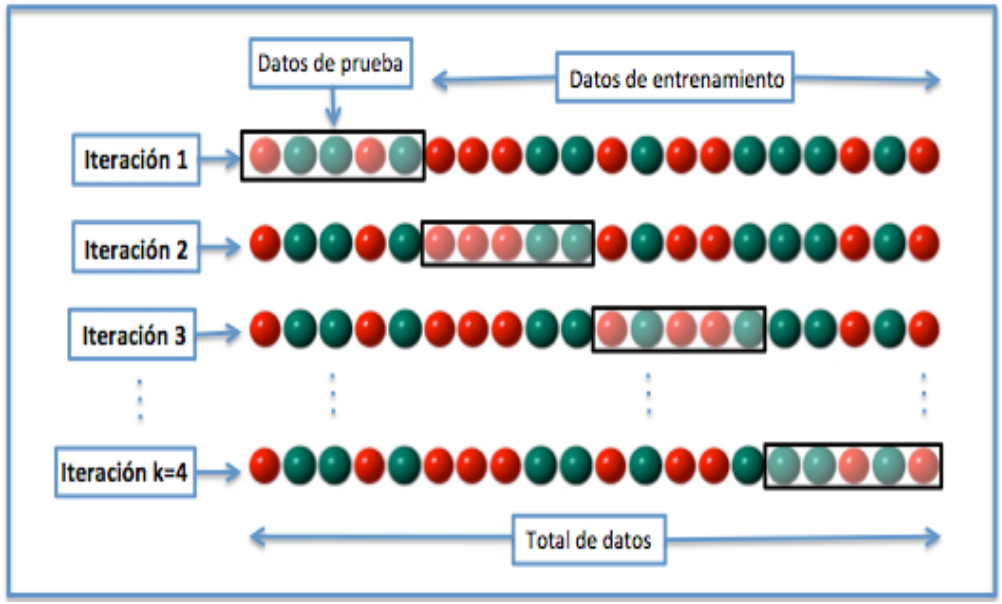
Common approaches are:

1. Random over-sampling.
2. Random under-sampling
3. Synthetic Minority Over-sampling Technique
4. Boosting techniques

Reference: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

Finding parameters and cross validation

- K-Fold cross validation is required to understand how the model will work in new data.
- Based on the ‘K’ value, the dataset will be split into number of groups. One of the data groups will be used as test data and others for training the model.
- To improve the accuracy of the model, proper selection of parameters is mandatory. Common function for tuning the parameter is ‘GridSearchCV` which comes with K-Fold cross validation.



https://machinelearningmastery.com/k-fold-cross-validation/. And
https://commons.wikimedia.org/wiki/File:K-fold_cross_validation.jpg

Summary

1. Types of machine learning algorithms
2. Evaluation metrics and error terms
3. Supervised machine learning algorithms in a glance
4. Handle imbalance in data, tuning and cross validation

Let’s practice

Resources and Credits

Github Resource : http://bit.ly/www-ds

saishruthis

SSaishruthi

splovyt