

# A Machine Learning Refresher

Q & A on the main concepts and terminology

Huascar Sanchez

`no.email.now`

Some Research Lab

March 28, 2020

The views expressed do not necessarily reflect the position of my employer.

## Q. What is Machine Learning?

Machine Learning is fitting a function to examples and using that function to generalize and make predictions about new examples.

Machine Learning, by large, falls into two categories:

- Supervised learning
- Unsupervised learning

## Q. How do you represent data in ML?

In general, the given data is expressed in a form of a bunch of vectors  $\vec{v}_j \in \mathbb{R}^d$  that belong to some high dimensional vector space.

For instance, in image recognition, the vector of an each image is a set of pixels (i.e., a pixelated version of the image).

If you have a notion of distance  $\Delta(\vec{v}_i, \vec{v}_j)$ , then you can compare which vectors are close to each other in this high dimensional vector space; e.g., the norm  $\|\vec{v}_i - \vec{v}_j\|^2$

## Q. What is Supervised Learning?

In **Supervised Learning** (or **SL**), you are given a bunch of examples and their labels (e.g., A or B) and the goal is to classify (or assign), when you are given a new example, to which label we assign the new example.

You could think of these labels the name of the **class or cluster** to which certain portions of the data belong.

## Q. Can you give an example of Supervised Learning?

**Support vector machines or SVM**, which goal is to construct the optimal separating *hyperplane* between pieces of data; e.g.,

- Say a hyperplane between clusters of data represented by labels A and B.
- These clusters sit in some high dimensional space, and the idea is to construct a plane that maximizes the margins between the plane and the data.
- If a new datum sits closer to one area of the data, say A, then we assign this new datum to A.

(For historical reasons) This algorithm is called **support vector machines** because the vectors that lie on the margin of the plan are called the *support* vectors.

This is a method for constructing a device to discriminate. If we're having a supervised learning problem then this method gives me an optimal form of discrimination.

## Q. What is Unsupervised Learning?

In **Unsupervised Learning** (or **UL**), you are given a bunch of data and you are not told it falls naturally into clusters, but you are not told what the clusters are.

The goal is identify the clusters of data, how many clusters there are, and then be able to assign new things to these different clusters.

## Q. Can you give an example of Unsupervised Learning?

Principal Component Analysis (or **PCA**) is a classical UL algorithm.

In **PCA**, the way this works, we construct a *covariance matrix*, and my covariance matrix is just the following object:  $C = \sum_j \vec{v}_j \vec{v}_j^+$ , where  $\vec{v}_j^T$  is the transpose of  $\vec{v}_j$ .

- In other words, we construct  $C$  from the data by taking these vectors  $\vec{v}_j$  and multiply them by their transpose  $\vec{v}_j^+$ .

E.g., Financial forecasting: the vectors could be, for example, can represent the changes in stock prices over a 24 hr period, and the covariance matrix  $C$  would give the correlations (or covariances in the data) between the prices of the different stocks in different times within the 24 hrs period.

In **PCA**, you diagonalize  $C$  and say  $C = \sum_k P_k \vec{\omega}_k \vec{\omega}_k^+$ ,  $P_k$  is piece of the data with size  $k$ , and  $\vec{\omega}_k$  are the set of vectors you need to find.

If and only if a small set of  $P_k \gg 0$ , then  $C$  is effectively low-rank, and the corresponding  $\vec{\omega}_k$  are the principal components.

Harder questions, please!



## Q. What is cross-validation?

Cross-validation is a technique for assessing how well a model performs on new independent data.

The simplest example of cross-validation is when you split your data into two groups<sup>1</sup>: (1) training data, and (2) testing data.

One uses training data to build the model and testing data to test the model.

---

<sup>1</sup>e.g., a ~60%~40% split

## Q. How to define/select metrics?

**There isn't a one-size-fits-all metric.**

The metric(s) chosen to evaluate a ML model depends on various factors:

Is it a regression or classification task? E.g., MSE vs Accuracy.

What is the business objective? E.g., precision vs recall.

What is the distribution of the target variable?

Examples of other metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared, Confusion Matrix and related metrics (Precision, Recall, Accuracy).

## Q. Can you explain what precision and recall are?

**Recall** is a measure of completeness or quantity, whereas

**Precision** is a measure of exactness or quality:

$$\underbrace{Recall = \frac{TP}{TP + FN}}$$

What proportion of actual positives  
was identified correctly?

$$\underbrace{Precision = \frac{TP}{TP + FP}}$$

What proportion of positive identifications  
was actually correct?

In simple terms, **high** precision means your algorithm has returned substantially *more relevant results than irrelevant ones*, while **high** recall means your algorithm has returned *most of the relevant results*.

## Q. What is a confusion matrix?

A **Confusion Matrix** is a simple way of understanding how well an algorithm is doing at classifying data. It is just the idea of **false positives** and **false negatives**

		Predicted	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

## Q. Can you explain what are false positives and false negatives?

There are two errors that often rear their head when you are learning about hypothesis testing -- false positives and false negatives, technically referred to as type I error and type II error respectively.

**False positives** are incorrect classifications of the presence of a condition when it is actually absent. A **false positive** is **when you reject a true null hypothesis**.

**False negatives** are incorrect classifications of the absence of a condition when it is actually present. A **false negative** is **when you accept a false null hypothesis**.

**Q. Provide examples when false positives are more important than false negatives, false negatives are more important than false positives and when these two types of errors are equally important**

E.g., cancer screening. It is much worse to say that someone doesn't have cancer when they do (false negative) rather than saying that someone has it when in fact they don't (false positive).

E.g., (from a psychological perspective) pregnancy test. Given the null hypothesis: "I am not pregnant." A false negative for someone who really does not want a child, is not ready for one and when assuring themselves with a negative result.

E.g., forecasting (health weather). During the covid-19 outbreak. Tracking rate of false positives and false negatives in March 2020 in the United States of America.

## Q. Can you list some of metrics derived from Confusion Matrix?

Confusion matrix related metrics

Accuracy rate:  $\frac{TP+TN}{Total}$  where Total is  $TP + TN + FP + FN$ ,

Error rate:  $1 - \frac{TP+TN}{Total}$ ,

False positive rate:  $\frac{FP}{TN+FP}$ ,

Recall:  $\frac{TP}{TP+FN}$ , Precision:  $\frac{TP}{TP+FP}$ ,

Specificity:  $1 - \frac{FP}{TN+FP}$ ,

Prevalence:  $\frac{TP+FN}{Total}$

**Q. How do you measure how good your classification algorithm is?  
You have unbalanced numbers, with our class being much larger or smaller than others**

You use the Matthews correlation coefficient. The number it yields is between plus or minus one. Plus one means perfect prediction, zero means no better than random.

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



## Q. What is linear regression?

Linear regression attempts to model the relationship between two continuous variables – a scalar response (or dependent variable) and one or more explanatory variables (or independent variables) – by fitting a linear equation to observed data.

The governing equation for linear regression is also quite simple. It takes the form:  $y = Bx + A$  where  $y$  is the dependent variable,  $x$  is the independent variable, and  $A$  and  $B$  are coefficients dictating the equation

e.g. a model that assumes a linear relationship between the input variables  $x$  and the single output variable  $y$ .

**Q. What is the general form of multiple regression?**

The general form of the equation for linear regression.