An overview of statistical learning.

Statistical learning refers to a VAST set of tools for understanding data.

These tools can be classified as:

Supervised.

Unsupervised.

Broadly Speaking:

Supervised statistical learning involves:

Building a statistical model for predicting / estimating an output Based on one or more inputs.

In supervised learning, you act as a teacher.

you provide Both the Dataset and the the correct answers!

The GOAL is:

For the model to learn relationship (function) between the inputs and the outputs so it can predict the Statistical answer for new unseen data.

* Data is labeled (knows input X, output Y).
* Goal: is a statistical prediction.
* Its common tasks:

1. Regression: predicting a continuous number (like predicting a house price based on it's size & the area).
2. Classification: predicting a category. (predicting if an email is a spam or not.)

Unsupervised statistical learning

You provide the model only inputs (features) without any labels.

The goal is for the model

to explore the data on it's own!

to find hidden structures, patterns, relationships.

* Data: is unlabeled (input X only).
* Goal: is Discovery or structure extraction.
* Common tasks:

1. Clustering: Grouping similar data points together.

(Segmenting customers into groups based on purchasing behavior)

2. **Dimentionality Reduction :** Simlifying tons of variables data by reducing the number of variables.

איפה אני אתקל בזה ב mql5?

I will encounter these concepts when I start Integrating machine learning libraries in MQL5 / Or using advanced mathematical functions.

SUPERVISED learning in MQL5 :

Teaching your EA to predict a specific outcome based on historical data :

* Python integration / ONNX : This is the most common modern method.

You train a model (like a neural network / random forest) In python using labeled data ("In the past when ATR was between 2.5 to 3.5 price went up.

then you save this model & load it into your MQL5 -A usin the ONNX function.

Theory + Example : Linear Regression in Trading

The core formula : $Y = \beta_1 x + \beta_0$

Y (Response) = The price (XAUUSD 1m candle close price)

X (predictor) = Time (The candle index : 1, 2, 3 ...)

β_0 (Intercept) = The starting price level of the regression line (Ground Zero).

β_1 (slope) = The momentum. (שיפוע) of the candles length.

$\beta_1 >\approx 0$ Uptrend

$\beta_1 <\approx 0$ Downtrend

$\beta_1 \approx 0$ Consolidation

The goal of the linear regression

To find the "Best Fit" line, we must minimize the RSS (Residual Sum of Squares)

$$RSS = \sum (y_i - \bar{y}_i)^2$$

We want the total overall distance between the candles and the regression line to be as small as possible.

Example

We want to create a regression line to be able to figure out if it's a down/up/cons...

For the last 5 candles n=5.

The Dataset:

* Time (x) = 1, 2, 3, 4, 5

* Price (y) = 10, 11, 12, 14, 15 (Simplified price)

Solution

1. Calculate the means (AVG of the price & time each)

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3$$

$$\bar{y} = \frac{10+11+12+14+15}{5} = \frac{62}{5} = 12.4$$

2. Calculate the slope (β_1) using the function & table below:

$$\text{Slope} = \frac{(x-\bar{x})\cdot(y-\bar{y})}{(x-\bar{x})^2}$$

time(x)	price(y)	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	10	-2	-2.4	4.8	4
2	11	-1	-1.4	1.4	1
3	12	0	-0.4	0	0
4	13	1	1.6	1.6	1
5	14	2	2.6	5.2	4
				13.0	10.0

$$\text{Slope} = \frac{(x-\bar{x})\cdot(y-\bar{y})}{(x-\bar{x})^2} = \frac{13}{10} = 1.3$$

The linear regression is = $y = 1.3x + 10$

kNN regression

- kNN is None parametric.
- KNN is used in supervised learning
- The algorithm both used in Regression & Classification.

It assumes that similar data points exist in close proximity to each other

How the algorithm works ?

① Choose $\boxed{k}$ -> Select the number of neighbors.
(hyperparameter)

② Calculate Distance -> Measure the distance between the new query point (Xq) <u>AND</u> all the other dataset points

③ Find Neighbors -> After the Distance sorting we pick the K nearest datapoints.

AFTER ALL THE DATASET IS READY

For Regression

We calculate the AVERAGE of the Y values of the K neighbors

For Classification

We take a MAJORITY VOTE The most frequent class among the neighbors

THE MOST IMPORTENT PART

CHOOSING THE RIGHT K

Bias variance tradeoff

The choice of k is the most important because it will decide if the algorithm will work / work idealy / not work / kinda work...

2 SCENARIOS DEPENDS ON THE K

Overfitting Small K (e.g K = 1)

* The model follows the training data too closely (High Variance)
* Capture noise & outliers.
* Low training error BUT Likely high test error

Underfitting Large K

* The model is too simple & ignores local details (High Bias)
* The prediction become Too close to the global average (naive model).

MODEL EVALUATION

To find the optimal k:

We typically use the RSS (Residual sum of squares) or MSE (Mean squared error) on a validation/test set.

The goal is to:

MINIMIZE the errors on unseen data not just the training data.