**PCA hoạt động như thế nào**

Đầu tiên, thuật toán PCA sẽ chuẩn hóa khung dữ liệu đầu vào, tính toán ma trận hiệp phương sai của các feature

**Eigen vector**

Bây giờ, chúng ta hãy thử tưởng tượng rằng mọi giá trị từ ma trận hiệp phương sai là một vector. Vector đó chỉ ra một hướng trong không gian n chiều (n là số lượng feature trong khung dữ liệu ban đầu). Vector là gì? Chúng ta có thể tưởng tượng rằng như một "mũi tên" chỉ theo một hướng nào đó trong không gian n-dimensional đó.

Những vectơ đó có thể được "trung bình" bằng cách tạo ra một vector khác "chỉ" ít nhiều theo cùng một hướng như tất cả các vectơ trung bình đó. Chúng ta gọi nó là eigenvector. Nó cũng có một giá trị (hãy tưởng tượng nó như là một "chiều dài" của "mũi tên") tương quan với số lượng vectơ trung bình của eigenvector

Sau đó, chúng tôi sắp xếp các eigenvectors bằng eigenvalues của họ.  Điều đó có nghĩa là chúng ta có thể chọn các eigenvectors cộng lại đến ngưỡng mong muốn của phương sai được giải thích.

**fit()**

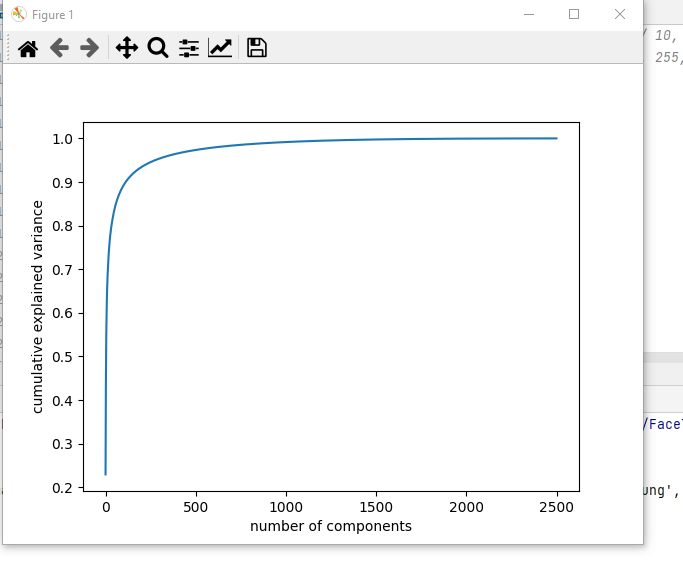
The fit method is calculating the mean and variance of each of the features present in our data. The transform method is transforming all the features using the respective mean and variance.

**transform()**

Using the transform method we can use the same mean and variance as it is calculated from our training data to transform our test data.

# **Now the question is why we did this?**

If we will use the fit method on our test data too, we will compute a new mean and variance that is a new scale for each feature and will let our model learn about our test data too. Thus, what we want to keep as a surprise is no longer unknown to our model and we will not get a good estimate of how our model is performing on the test (unseen) data which is the ultimate goal of building a model using machine learning algorithm.



Clasification report

Báo cáo phân loại được sử dụng để đo lường chất lượng dự đoán từ thuật toán phân loại. Có bao nhiêu dự đoán là đúng và bao nhiêu là sai. Cụ thể hơn, True Positives, False Positives, True Negatives và False Negatives được sử dụng để dự đoán các số liệu của một báo cáo phân loại

There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative**: the case was negative and predicted negative
2. **TP / True Positive**: the case was positive and predicted positive
3. **FN / False Negative**: the case was positive but predicted negative
4. **FP / False Positive**: the case was negative but predicted positive

**Precision — *What percent of your predictions were correct?***

Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.

Precision:- Accuracy of positive predictions.

Precision = TP/(TP + FP)

**Recall — *What percent of the positive cases did you catch?***

Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall:- Fraction of positives that were correctly identified.

Recall = TP/(TP+FN)

**F1 score — *What percent of positive predictions were correct?***

The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

**Support**

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn’t change between models but instead diagnoses the evaluation process.

