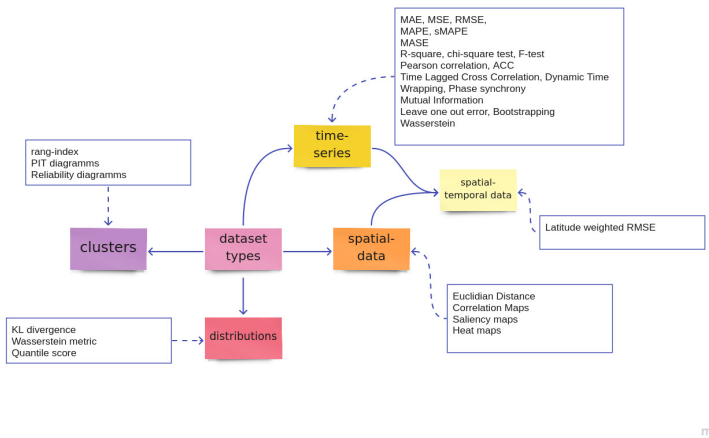
scores_metrics

jakob-schloer edited this page 18 hours ago · 18 revisions

Prediction scores and metrics



Pages 3

Table of contents

Home

Journal Club

Prediction scores and metric

Clone this wiki locally

https://github.com/mlcs

Time series data

Scale dependent errors

The errors are on the same scale as the data, i.e. different data sets cannot be compared

Mean absolute error (MAE)

> $$MAE = mean(|y_i - \hat{y}_i|) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Minimizing MAE leads to prediction of median.

Mean square error (MSE)

> $$MSE = mean(|y - \hat{y}_i|^2)$$

Strongly penalizes large wrong predictions

Root mean square error (RMSE)

> $$RMSE = \sqrt{mean(|y_i - \hat{y}_i|^2)}$$

Minimizing RMSE leads to prediction of mean

Percentage errors

Percentage errors are unit free and therefore allow comparison of different data sets.

Mean absolute percentage error (MAPE)

> $$MAPE = mean(\frac{|y_i - \hat{y}_i|}{y_i})$$

Problems:

- cannot be used for zero values
- puts more weight on negative than positive errors

Symmetric absolute percentage error (sMAPE)

Used to overcome the problems of MAPE

$$sMAPE = mean(\frac{|y_i - \hat{y}_i|}{(|y_i| |\hat{y}_i|) / 2})$$

Scaled errors

Alternative to percentage errors when comparing different datasets

Mean absolute scaled error (MASE)

$$MASE = \frac{mean(|y_i - \hat{y}_i|)}{\frac{1}{N-1}\sum_{t=2}^{N} |y_t - y_{t-1}|}$$

- scale invariance
- symmetric
- less than one if it arises from a better forecast than the average naïve forecast and conversely it is greater than one if the forecast is worse than the average naïve forecast

https://otexts.com/fpp2/accuracy.html

https://scikit-learn.org/stable/modules/model_evaluation.html#

Goodness of Fit

Check if a hypothesis is correct. Often referred to as explained variance scores.

Coefficient of determination (R^2)

Describes the variance (of y) which is explained by the model prediction.

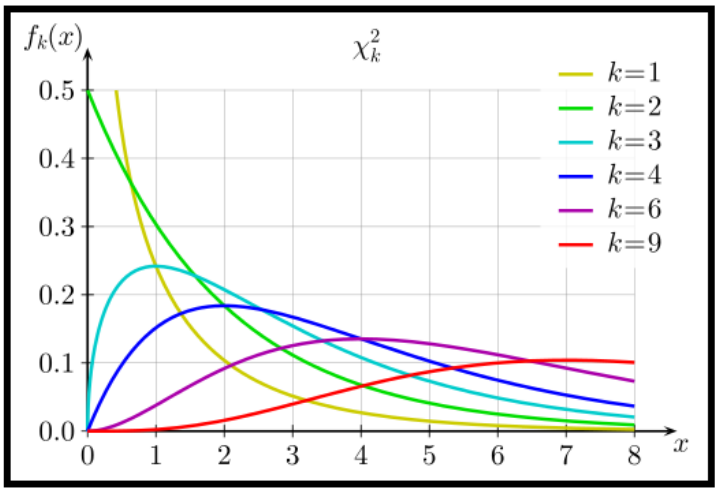
$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

Chi-square test

Statistical test that measures the significance of the hypothesis given the data. Chi-square is obtained by

$$\chi^2 = \sum_i^N \frac{(\hat{y}_i - y_i)^2}{y_i}$$

The sum of square errors follows the chi-square distribution if the errors are independent and normally distributed. The probability of the calculated χ^2, $p(\chi^2)$ is called "p-value". If the p-value will be larger than the threshold α, the hypothesis is significant.



F-test

The F-value expresses how much of the model has improved compared to the mean (null hypothesis) given the variance of the model and data.

For regression, it can be generalized in order to compare the fit quality of a complex model as compared to a simpler version of the same model.

Here, we assume model 1 with k_1 number of parameters and model 2 with k_2 number of parameters, where k_1 > k_2. The F-test is obtained as follows:

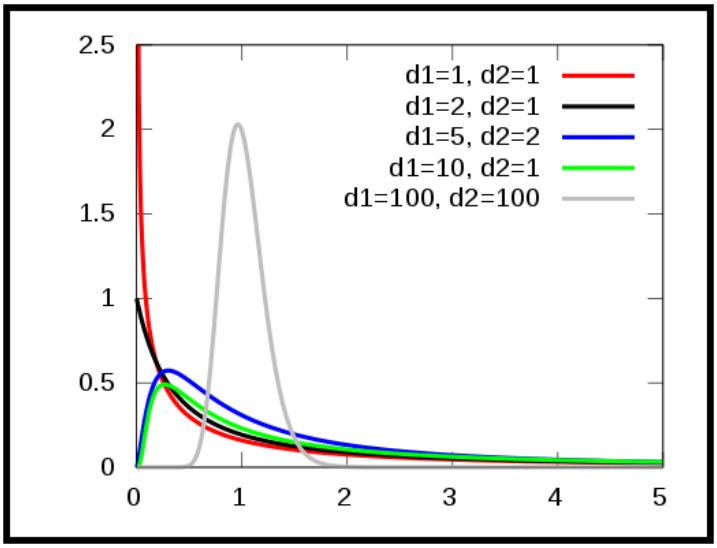
1. Compute the sum of squares of residual errors $RSS = \sum_i (y_i - \hat{y}_i)^2$ for each model

2. The F-statistics of the two regression models is obtained by

$$F = \frac{\frac{RSS_1 - RSS_2}{k_1 - k_2}}{\frac{RSS_2}{n - k_2}}$$

where n is the number of data points.

3. The RSS are random variables described by a probability distribution. The sum of **independent** and **standard normal** variables follow the pdf of the chi-square distribution. This happens to be the case in the equation above. The ratio of the two scaled chi-square distributions is described by the F-distribution.



4. We can evaluate the probability of occurrence p and compare it to our threshold value α (some small number we choose). If $p < \alpha$ we can conclude that model 1 is able to explain the variance in the data better than model 2.

https://towardsdatascience.com/fisher-test-for-regression-analysis-1e1687867259

Correlation and Synchrony

Pearson correlation

Pearson correlation measures how two continuous signals co-vary over time. The linear relationship between these signals are given from -1 (anticorrelated) to 0 (incorrelated) to 1 (perfecly correlated).

The Pearson correlation coefficient for two random variables X_1 and X_2 is:

> $$\rho_{X_1,\,X_2} = \frac{cov(X,\,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X-\mu_x)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

For time-series on can calculate a

- global correlation coefficient: a single value

- local correlation coefficient: determine correlation in a rolling window over time

Caution:

1. outliers can skew the correlation
2. assuming the data is homoscedatic, i.e. constant variances

Anomaly correlation coefficient (ACC)

In climate science and meteorology correlating forecasts directly with observations may give misleadingly high values because of the seasonal variations. It is therefore established practice to subtract the climate average from both the forecast and the verification. The anomaly correlation coefficient is obtained by

$$ACC = \frac{\sum_i (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2 \sum_i (y_i - \bar{y})^2}}$$

In case the climate average $\bar{c}$ is known, it often replaces $\bar{\hat{y}} \text{and} \bar{y}$.

https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2013-nwp/pdf/outline2013_Appendix_A.pdf

Time Lagged Cross Correlation (TLCC)

TLCC is a measure of similarity of two series as a function of displacement. It captures directionality between two signals, i.e. leader-follower relationship.

Idea: Similar to convolution of two signals, i.e. shifting one signal with respect to the other while repeatedly calculating the correlation.

$$(f \star g)(\tau) \triangleq \int_{-\infty}^{\infty} f^*(t) g(t\tau)\, dt$$

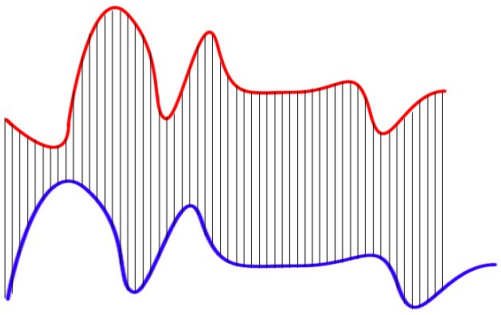
Windowed time lagged cross correlations (WTLCC) are an extension of TLCC where local correlations coefficients are computed for each lag-time which is then plotted as a matrix.

Dynamic Time Wrapping (DTW)

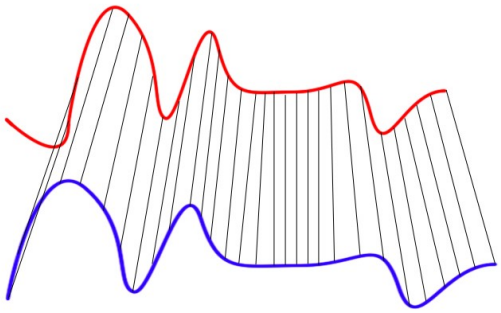
DTW computes the path between two signals that minimize the distance between the two signals. DTW computes the euclidean distance at each frame across every other frames to compute the minimum path that will match the two signals.

Properties:

- deal with signals of different length
- requires interpolation of missing data



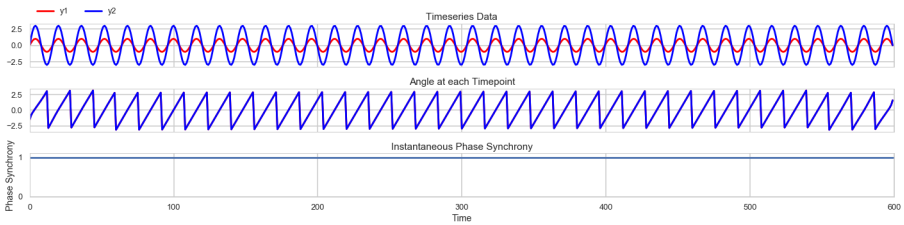
Euclidean Matching



Dynamic Time Warping Matching

Instantaneous phase synchrony

For time series with oscillating properties the instantaneous phase synchrony measures the phase similarities between signals at each timepoint. The phase between signals is refered to as angle which is obtained by a Hilbert transformation of the signals. Phase coherence can be quantified by subtracting the angular difference from 1.



http://jinhyuncheong.com/jekyll/update/2017/12/10/Timeseries_synchrony_tutorial_and_simulations.html

https://towardsdatascience.com/four-ways-to-quantify-synchrony-between-time-series-data-b99136c4a9c9

Information/Entropy measures

Information measures on time series amounts to constructing empirical distributions and apply Shanon Information Measures on them.

Mutual Information (MI)

MI is a measure of the mutual dependence of two random variable. Applied to two time series $Y = \{y_i | i = 1, ..., N_y\}$ and $X = \{x_i | i = 1, ..., N_x\}$, we first construct the empirical distributions > $p(x_i)$, $p(y_i)$ and the joint $p(x_i, y_i)$. The mutual information is defined as

> $$I(X, Y) = \sum_{x_i, y_i} p(x_i, y_i) \log\left(\frac{p(x_i, y_i)}{p(x_i)p(y_i)}\right)$$

which is essential the KL-divergence of the distributions.

https://elife-asu.github.io/PyInform/timeseries.html

Cross-validation

Leave-one out error (LOOE)

Bootstrap

Others

Wasserstein metric

Granger causality

Statistical hypothesis test to determine whether one time series is useful in forecasting another.

Spatial data

- Euclidian distance (ACC)
- correlation maps
- density metrics
- saliency maps: highlights which changes in the input would most affect the output.

- heat maps: highlights which inputs are most important for the prediction

Spatial temporal data

Latitude weighted RMSE

Skill metric used in climate science to compare time-series from different spatial points:

$$RMSE = \frac{1}{N}\sum_{i}^{N}\sqrt{\frac{1}{N_{\text{lat}}}\frac{1}{N_{\text{lon}}}\sum_{j}^{N_{\text{lat}}}\sum_{k}^{N_{\text{lon}}} L(j)(\hat{y}_{i,\ j,\ k} - y_{i,\ j,\ k})^2}$$

with latitude weighting factor

$$L(j) = \frac{\cos(\text{lat}(j))}{\frac{1}{N_{\text{lat}}}\sum_{j}^{N_{\text{lat}}}\cos(\text{lat}(j))}$$

Cluster data

- rang-index
- PIT diagram
- reliability diagram

Probability Distributions

- KL-divergence
- Wasserstein metric