Multivariat Notater

Henrik Horpedal

May 2024

Contents

1 PCA - Principal Component Analysis

$$\mathbf{T} = \mathbf{XP}, \quad \mathbf{X} = \mathbf{TP^T} \tag{1}$$

with residuals:

$$\mathbf{T} = \mathbf{XP} + \mathbf{E}, \quad \hat{\mathbf{X}} = \mathbf{TP^T} \tag{2}$$

P: Loadings
T: Scores
X: Original data
$\hat{\mathbf{X}}$: Projected data into reduced dimensions.
E: Residuals

PCA is a linear transformation that finds the directions of maximum variance in the data. It projects the data into a new subspace.

1.0.1 Deciding the number of components

1.0.2 Outliardetection

2 Some definitions from statistics

Covariance: Covariance is a measure of the relationship between two variables. A positive covariance means that the variables tend to increase or decrease together, while a negative covariance means that the variables tend to move in opposite directions. In other words, positive correlation means "if we measure an x that is bigger than its mean, then likely y will be bigger too"

Correlation: The correlation coefficient, denoted by "r", is a value between -1 and 1 that measures the strength and direction of the linear relationship between two variables. A value of 1 means that there is a perfect positive correlation (i.e. a perfect linear relationship) between the two variables, a value of -1 means that there is a perfect negative correlation (i.e. a perfect linear relationship with a negative slope) between the two variables, and a value of 0 means that there is no correlation between the two variables.

3 LS - Least Squares

Least squares is about minimizing the sum of the squares of the differences between the observed and predicted values:

$$\min_\theta J(\theta) = \min_\theta \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i(\theta) - y_i)^2 \tag{3}$$

A common way to predict the value of y is to use a linear model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n \tag{4}$$

3 can solved numerically with **gradient descent**:

$$\theta_{j+1} = \theta_j - \alpha \nabla_\theta J(\theta) \tag{5}$$

where α is the learning rate.
In **Batch gradient descent** we use all the training examples in each iteration.
In **Stochastic gradient descent** we use only one training example in each iteration.

Normal Equations

If we assume a linear model like 4 the least squares problem can be solved analytically with the normal equations. We fist have:

$$\mathbf{y} = \mathbf{X}\theta \tag{6}$$

and

$$\mathbf{X^T}(\hat{\mathbf{y}} - \mathbf{y}) = \mathbf{0} \tag{7}$$

Substiting in $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ we get:

$$\Rightarrow \quad \mathbf{X^T}(\mathbf{X}\theta - \mathbf{y}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{X^T X}\theta = \mathbf{X^T y} \tag{8}$$

assuming the columns of X are linearly independent we get:

$$\theta = (\mathbf{X^T X})^{-1}\mathbf{X^T y} \tag{9}$$

4 Regularization

Regularization is used to punish large values of the parameters θ in order to avoid overfitting. The most common regularization techniques are **L1** and **L2** regularization. When adding regularization you are trading off variance with some bias. (More about that in section xx)

4.1 L1 - LASSO

L1 regularization adds the sum of the absolute values of the parameters to the cost function:

$$\min_\theta J(\theta)_{new} = \min_\theta \quad J(\theta) + \lambda \sum_{j=1}^{n} |\theta_j| \tag{10}$$

4.2 L2 - Ridge

L2 regularization adds the sum of the squares of the parameters to the cost function:

$$\min_{\theta} J(\theta)_{new} = \min_{\theta} \quad J(\theta) + \lambda \sum_{j=1}^{n} \theta_j^2 \tag{11}$$

4.3 Comparison between L1 and L2

The L1 function look like a diamond and the L2 function look like a cone. The L1 function have higher chance of driving θ to zero, which is beneficial reducing model complexity. The L2 function will drive θ close to zero but not exactly zero. The L1 function is more computationally efficient to solve than the L2 function since J will in many cases remain convex.

5 Preformance Metrics

TODO

6 Maximum Likelihood

Maximum likelilihood is about fiting a distrebution to the observed data. It is about choosing both the parameters of the distrebution and the distrebution itself:

$$\theta_{ML} = \arg\max_{\theta} P(X|\theta) \tag{12}$$

The likelihood function is defined as:

$$L(\theta) = P(X|\theta) \tag{13}$$

The maximum likelihood estimator is the value of θ that maximizes the likelihood function. The log-likelihood function is often used since it is easier to work with:

$$l(\theta) = \log(L(\theta)) \tag{14}$$

7 Sampling

Sampling is the process of selecting a subset of individual or items for a larger population for the purpose of analysis or study. We want to get a good represenation of the population with only a subset of the population.

7.0.1 Simple Random Sampling

In simple random sampling, each individual is chosen entirely by chance and each member of the population has an equal chance of being included in the sample. It is impossible to sample truly randomly and it will cause low represenation of some subgroups.

7.0.2 Systematic Sampling

In systematic sampling, the list of the population is ordered in some way and then the sample is chosen according to some pattern. This is done by selecting a random starting point and then picking every nth element in the list. Is not recomended if the population has a syclic pattern.

7.0.3 Cluster Sampling

In cluster sampling, the population is divided into clusters and then a simple random sample of clusters is chosen. All individuals in the selected clusters are included in the sample. This is useful when the population is geographically spread out.

8 Bias-Variance Tradeoff

9 DoE - Design of Experiments

10 MLR - Multiple Linear Regression

MLR is used to model the relationship between a dependent variable and one or more independent variables. The model is defined as:

$$\mathbf{y} = \theta_\mathbf{0} + \theta_\mathbf{1}\mathbf{x_1} + \theta_\mathbf{2}\mathbf{x_2} + ... + \theta_\mathbf{n}\mathbf{x_n} \tag{15}$$

The parameters θ are estimated by minimizing the sum of the squares of the differences between the observed and predicted values. This is in practise the same as LS with a linear model; done by solving the normal equations 9. It is worth mentioning that if you have correlated X-variables the regression coefficient will not correspond to the true effect of the variable. (A negative θ_i does not mean that the variable has a negative effect on y). You also have to have at least as many samples as variables in MLR.

11 ANOVA - ANalysis Of VAriance

12 PCR - Principal Component Regression

The main idea of PCR is to use PCA to reduce the dimensionality of the data and then use MLR on the reduced data. The steps are:

1. Do PCA on the X's.
2. Choose the number of components to use.
3. Do MLR on the reduced data.

This is practical because the regression coefficients are interpretable aposed to normal MLR with potentioally correlated X-variables. It is also very practical since you can detect outliears in the predicition phase. However, in the PCA-phase the variance of X is only considered, nothing guarantees that the principal components which explained X optimally could be relevant for the prediction of Y.

13 PLSR - Partial Least Squares Regression

PLS tries to solve the problem of PCR by considering the covariance between X and Y. The model can be described as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{16}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{17}$$

where:

- loadings ($\mathbf{P}$) and scores ($\mathbf{T}$) in the $\mathbf{X}$ space.
- loadings ($\mathbf{Q}$) and scores ($\mathbf{U}$) in the Y space
- loading weights ($\mathbf{W}$), eigenvectors maximizing the covariance between deflated Xa and Ya.
- errors ($\mathbf{E}$ and $\mathbf{F}$) in the $\mathbf{X}$ and $\mathbf{Y}$ spaces.

14 NlPALS - Non-linear Partial Least Squares

15 Validation

16 Outlier Detection

17 Feature Selection

18 SVM - Support Vector Machines

19 Timeseries Forecasting

19.1 Stationarity

A process is said to be **strictly stationary** (or strongly stationary) if its statistical properties are invariant to shifts in time. This means that the joint probability distribution of the process does not change when shifted in time. **Wide sense stationary** is a less strict condition where the mean and variance is constant and the autocovariance function only depends on the time lag. Wide sence stationarity is a necessary condition for beeing able to predict the future.

19.2 Ergodicity

A process is said to be **ergodic** if the time average of one sequence is equal to the ensemble average of another sequence. This means that the statistical properties of the process can be estimated from a single, sufficiently long, realization of the process.

19.3 Decomposition of a timeseries

A timeseries often consists of a trend, cyclic component, seasonale component and a residual. The trend is the long-term increase or decrease in the data. The cyclic component is the periodic fluctuations in the data. The seasonal component is the periodic fluctuations in the data that occurs at fixed intervals. The residual is the part of the data that is left after removing the trend, cyclic and seasonal components. Timese can be made out of addidative or multiplicative components.

19.3.1 FFT - Fast Fourier Transform

FFT can be used to estimate the frequency content of a timeseries. It does assume that the timeseries is infinite and periodic and will therefore only give an approximation of the frequency content of the timeseries. A downside is

that when the signal is converted to s-domain you no longer know at what time events are happening. A method to account for this is **windowed FFT** where you repeatadly do FFT over a sliding window.

19.3.2 Wavelet Transform

Wavelet transform aim to solve the loss of time information in the FFT. The wavelet transform transforms a signal in time domain to a two-dimensional representation in time and frequency. In stead of decomposing the signal into sines and cosines of diffrent frequency, the wavelet transform decomposes the signal into wavelets of diffrent frequency AND diffrent time shift. Wavelets are in contrast to sines and cosines not periodic.

19.3.3 EMD - Empirical Mode Decomposition

EMD decomposes the timeseries into a finite number of **intrinsic mode functions** (IMFs). An IMF is a function that has satisfies:

1. The number of extrema and the number of zero crossings must be equal or differ at most by one.
2. At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

EMD can be described by:

$$x(t) = \sum_{i=1}^{n} IMF_i + r \tag{18}$$

Advantages of EMD:

- works wit short time series.
- Can be used with non-stationary data.
- Can be applied to data representing non-linear processes.
- Works with non-periodic time series.

When preforming EMD you never leave the time domain. The IMFs are highly otrhogonal but not completely. The algorithm works as follows:

1. Identify all the local maxima and minima of the signal.
2. Interpolate between the maxima and minima to get the upper and lower envelope.
3. Calculate the mean of the upper and lower envelope.
4. Subtract the mean from the original signal.
5. Repeat the process until the signal is a IMF.

IMF - Intrinsic Mode Functions

20 Multiway, multiblock and IDLE modelling

21 Clustering, classification and discrimination

22 DTW - Dynamic Time Warping

Dynamic time warping is a similarity meassure between two time series. It is much used in speech and motion recognition. It is mathematically defined as:

$$DTW_q(x, x') = \min_{\pi \in \mathcal{A}(x,x')} \left(\sum_{(i,j)\in\pi} d(x_i, x'_j)^q \right)^{\frac{1}{q}} \tag{19}$$

- π is an **alignment path**, a series of K index pairs $(i_1, j_1), (i_2, j_2), ..., (i_K, j_K)$ where i_k and j_k.
- $A(x, x')$ is the set of all possible π's, satisying the following conditions:
  1. The start and end is matched. $\pi_0 = (0, 0)$ and $\pi_K = (n-1, m-1)$
  2. The index set is monotonic increasing and all datapoints must apear at least once.
- q is the type of norm used to meassure the distance between the two time series.
- d is the distance function.(probably).

$DTW(x, x) = 0$ and $DTW(x, x') > 0$ for $x \neq x'$.
The optimization problem 19 is solved by dynamic programming with complexity $O(nm)$ where n and m is the length of the two time series.
It is worth mentioning that DTW is not a metric since it does not satisfy the triangle inequality.

- It can be used as a **pattern recognition tool** by predefining potential patterns and then compare the patterns to the time series. If the DTW is bellow a certain threshold the pattern is said to be present in the time series.

- It can be used as a **feature extraction tool** by comparing the time series to a set of reference time series. Features can be distances to the reference time series, statistics from the alignment path etc.
- It can be used as a **anomaly detection tool** by comparing the time series to a set of reference anmaly series. If the DTW is bellow a certain threshold the time series is said to be an anomaly.

23 LDA - Linear Discriminant Analysis

LDA can be used to categorze between two or more classes. Assumtions are:

- The classes are gaussian distributed.
- The classes have the same variance matrix Σ.
- The moments of the distributions can be estimated from the data.

Graphically LDA can be thought of as finding the line that maximizes the distance between the means of the classes and minimizes the variance within the classes. The line is called the **discriminant line** (discriminant loading(?)). The projection of the data onto the discriminant line is called the **discriminant score**. Afterwards a threshold c can be set with to classify the data:

$$discriminationtest = \{ \ w^T \mathbf{x} \geq c \rightarrow class1 otherwise class2 \tag{20}$$

24 QDA - Quadratic Discriminant Analysis

If we remove the assumtion of $\Sigma_1 = \Sigma_2$ we get QDA. It is more complex to solve and LDA with quadratic feature engineering often give similar results.

25 Model structure selection

26 PLS-DA - Partial Least Squares Discriminant Analysis

If we have categorical y's PLS can be used as a classifications algorithm the following way:

1. transform y's to $\hat{y}s$ by one hot encoding. (Perhabs use -1 and 1 instead of 0 and 1).
2. do PLS on the $\hat{y}$'s and the X's.
3. use new X's to predict with the model. Choose the class with the highest predicted value.

The nice thing about this algorithm is that you get an associated probability with the prediction.
Another nice thing is that you can study the correlation loadings to see which variables are most important for the diffrent classes.

27 PCA for classification

PCA can be used as a classification algorithm for detecting cancer the following way:

1. Do PCA on the X's (pixels) and generate score plot.
2. Mark a area in the picture where the cancer is present and see where they correspond to in the score plot.
3. Look at neighbouring pixels in the scoreplot and see if they are also cancerous.

28 t-SNE - T-distributed Stochastic Neighbour Embedding

A challenge with PCA isi that it captures the global struchture of the data. t-SNE aims to capture to local data and project it into a lower dimensional space.

- **Neighbour Embedding**: The process of mapping high-dimensional data into a lower-dimensional space. The main idea here is to keep the distances between the data points in the lower dimenison as well. This is howerever generally not possible beacuse of whats called the **crowding problem**. Therefore only distances to "neighbouring" points are kept.
-

29 Clustering

In a clustering promblem there are no predefined classes. The goal is to group similar data into clusters. Since we have no "solution" which we try to optimize to, it is an unsupervised algorithm. Clustering is wise to do prior to classification and also as EDA.

Linkage Methods are methods that uses a hierarchical approach to clustering. It is recursivly merging individual datapoints or exisisting clusters into new clusters based on a similarity meassure. The results of hierarcical methods can be visualized in a dendrogram. There clear bias variance tradeoff when choosing the number of clusters. It also important to be aware of "wishfull thinking" since you always can find a cluster if you look hard enough. The most common way to meassure similarity is through the euclidian distance.

29.0.1 K-means

- Simple and easy to implement

- Does not work with non-cicular clusters
- The number of clusters have to be predefined
- The algorithm is sensitive to the initial cluster centers, and must often be run multiple times with different initializations.
- You have to be aware of diffrent scaling on diffrent variables.
- "is bad" -Damiano.

29.0.2 DBSCAN - Density Based Spatial Clustering of Applications with Noise

In DBSCAN you define a neighborhood ϵ which is a distance from a point. Diffrent types of points are defined as following:

- **Core point**: A point that has many number of points within its ϵ neighborhood.
- **Border point**: A point that is within the that only has one neighbour.
- **Outlier**: A point that doesnt have any other neighbours.

The results of DBSCAN is not as dependent of ϵ as K-means is of the number of clusters. DBSCAN is also able to find non-circular clusters.

29.0.3 Visualizing clusters

If you only have one cluster method and few samples you can use a **dendogram**. If you use two clusterers and have few samples you can use a **tanglegram**.

30 ICA - Independent Component Analysis

The fact that the loadings in PCA are orthogonal can sometimes fail to describe the underlying structure of the data. In these cases ICA might be more appropriate. It is much used in signal processing to separate mixed signals. Let us say we have two audio signals x_1 and x_2 recorded at the same time. And that there are two sound sources s_1 and s_2 that are mixed together to form the signals x_1 and x_2. The mixing can be described as:

$$\mathbf{x_1} = \mathbf{a_{11}s_1} + \mathbf{a_{12}s_2} \tag{21}$$

$$\mathbf{x_2} = \mathbf{a_{21}s_1} + \mathbf{a_{22}s_2} \tag{22}$$

This system of equations is underdetermined and we can not solve it directly. But if we assume:

1. The sources are **independent**. $P(s_1, s_2) = P(s_1)P(s_2)$

2. Maximum one of the sources are **non-gaussian**. This is because the central limit theorem states that the sum of independent random variables will be gaussian distributed. Its therefore impossible to separate the sources if they are both gaussian distributed.

3. The A-matrix is **full rank**. This means that the sources are not mixed in the same way. You have to use different recordings.

ICA can be applied. ICA can be interpreted through SVD:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}\mathbf{s} \tag{23}$$

$\mathbf{U}$ is a rotational matrix, $\mathbf{\Sigma}$ is a scaling matrix and $\mathbf{V}^{\mathbf{T}}$ is a final rotational matrix. The angle of $\mathbf{U}$ is found by maximizing the variance. $\mathbf{\Sigma}$ is the covariance matrix and the angle of $\mathbf{V}^{\mathbf{T}}$ is found by maximizing the Kurtosis.

It is wourth mentioning after preforming ICA the order of the components are arbitrary. You also get a scaling ambiguity.

31 Tree-based methods

Tree-based methods require minimal preprocessing. They are computationally efficient and interpretable, but can easily overfit. The cost of predicting a sample with a trained model is the log of datapoints used to train. Decisiontrees can eaisly visualize feature importance of the model. Small variations in dataset will lead to compleatly different trees, and Skewed datasets can lead to biased trees.
In practise they work by asking a series of questions that maximizes the information gain:

$$Information \quad gain = Entropy_{parent} - \frac{1}{N}\sum_{i=1}^{n} Entropy_{child_i} \tag{24}$$

where the entropy is defined as:

$$Entropy = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{25}$$

And can be thought of as a measure of purity. **Random forests** is an ensemble method that uses multiple decision trees to improve the accuracy of the model. The features are selected randomly and the trees are trained on different subsets of the data. The final prediction is the average of the predictions of the individual trees.

32 Logistic regression

Logistic regression is used when the target variable is binary. It is therefore a classification algorithm. Logistic regression can be used for digit recognition. The algorithm is esentially a linear regression model with a sigmoid function applied to the output. The sigmoid function is defined as:

$$\sigma(\mathbf{z}) = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-\mathbf{z}}} \tag{26}$$

33 Neural networks