# Document Information

|  |  |
| --- | --- |
| Document title | Report {{ project.name }} |
| Document version | 1.0 |
| Processing status | Draft |
| Customer |  |
| Project name | {{ project.name }} |
| Project number | xxx xxx xxx |

Release of the document

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Position / Company | Date (DD.MM.YYYY) | Signature |
| Prepared by |  |  |  |
| {{ project.author }} | Data Scientist  Helbling Technik AG | {{ project.date }} |  |
| Reviewed by |  |  |  |
| Surname Name | Project Lead  Helbling Technik AG |  |  |
| Surname Name | Project Lead  Helbling Technik AG |  |  |
| Approved by |  |  |  |
| Surname Name | Project Lead  Helbling Technik AG |  |  |

Revision history

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version | Date | Author | Modified positions | Reason |
| 1.0 | {{ project.date}} | {{ project.author }} | - … | - … |

ReferenceD Documents

|  |  |
| --- | --- |
| Ref. | Description |
|  | https://en.wikipedia.org/wiki/Random\_forest |

Linked Documents

|  |  |
| --- | --- |
| Ref. | Description |
|  | n/a |
|  | n/a |

Attached Documents

|  |  |
| --- | --- |
| **Ref.** | **Description** |
|  | n/a |
|  | n/a |

Table of Contents

[1 Document Information 1](#_Toc508196208)

[1.1 Purpose of the Document 4](#_Toc508196209)

[1.2 Scope 4](#_Toc508196210)

[1.3 Abbreviations and Terms 4](#_Toc508196211)

[2 Classification Problems 5](#_Toc508196212)

[2.1 General concepts 5](#_Toc508196213)

[2.2 Measurements 7](#_Toc508196214)

[2.3 Classification performance estimation 7](#_Toc508196215)

[2.4 Generalization 9](#_Toc508196216)

[3 Illustrations 10](#_Toc508196217)

[4 Annex 12](#_Toc508196218)

[4.1 Additional 12](#_Toc508196219)

[4.2 List of Tables 12](#_Toc508196220)

[4.3 List of Figures 12](#_Toc508196221)

## Purpose of the Document

A purpose statement is a declarative sentence which summarizes the specific topic and goals of this document.

## Scope

A scope gives the reader an accurate, concrete understanding what the document cover and what he/she can gain from reading it.

## Abbreviations and Terms

The following table defines the terms and abbreviations specific for this document.

| Abbreviation | Meaning |
| --- | --- |
| n/a | n/a |

|  |  |
| --- | --- |
| Term | Definition |
| n/a | n/a |

# Classification Problems

## General concepts

This chapter should illustrate general considerations of a classification problem as stated in chapter. These considerations are important for understanding the performance measures used for the candidate model evaluation. Furthermore, it shows the fundamental limits of decidability.

### Feature space

The name “feature” is widely used in the context of classification problems, however not always consistently. Here, it is defined as the values that are obtained by a measurement. This definition is somehow fuzzy but has a general meaning when considering the physical interface of the stated problem (leaving aside the exact technical implementation of sensors, devices, etc.).

For StroKare the features are the sensor readings from the IMU and EMG. This definition leads to a reasonably small number of dimensions of the feature space (around 10)[[1]](#footnote-1).

### Set of positives and set of negatives

Using the definition above all measurements are points in the feature space Ω and correspond to exactly one class in sets of classes (here “paresis” and “non paresis”). As the goal is to build a detector for the “paresis” measurements these are defined to be the “positives” while all measurements for “non paresis” are “negatives”.

Without loss of generality, measurements for both positives and negatives can be taken as being described by a probability distribution onto the feature space.

##### Assumption of uniform density

For simplicity and illustration, it is assumed, that both distributions are uniform distributions with constant density over an arbitrary subset of the feature space and zero on the complement.

(One could also proceed without the assumption and instead using high density regions)

The assumption allows the definition of the following subset:

These sets can be illustrated on a two-dimensional feature space see Figure 2‑1.

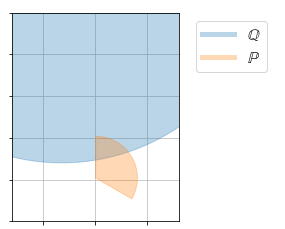


Figure 2‑1 shows a fictive example of the set of positives P and set of negatives Q in a two-dimensional feature space

### Classification performance measures

All classifiers C map every point in to either “positive” (paresis) or negative (non-paresis). Thus, we can define for each classification two sets:

This label assigned by the classifier can either be true or false. So, given a classification every point in (that is every possible measurement) corresponds to exact one of the following classes **true positive**, **false positive**, **true negative** and **false negative**.

The true and false classifications define the performance of the classifier and one can define the following measures.

µ denotes the probability measure defining the sets and in case of a uniform density just the area or the sets.

|  |  |
| --- | --- |
| Measure | Definition |
| True positive rate (TPR), Sensitivity |  |
| False positive rate (FPR) |  |

### Optimal Classifier

From the illustration in Figure 2‑1 it becomes clear that in general there is no single optimal classification. Rather, using again the assumption of uniform density, there is a trade-off to be made between the situations illustrated in Figure 2‑2. The classification mapping for a two-dimensional feature space can be visualized by the means of a separation line (or lines). In the regions outside bot P and Q classification has no meaning.

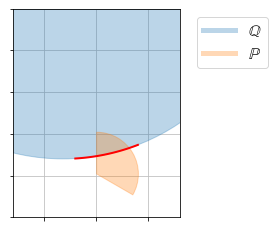
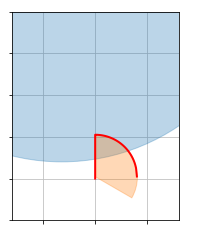


Figure 2‑2 shows the two extreme classifier separation lines for the case (a) maximizing the TPR (left) and the case (b) minimizing the FPR (right). The classifier would map all measurements in the feature space lying below the red line to be positive (paresis).

The trade-off corresponds to a trade-off in the true positive rate (TPR) and false positive rate (FPR) (refer to ). The two extreme cases maximize the TPR (a) and minimizes the FPR (b) respectively. It is straight forward to the derive the values for the optimal classifiers (theoretical limit) and is shown in Equation 1 and Equation 2 where again µ denotes the probability measure (or area) of the sets.

Equation 1 TPR and FPR for case (a) maximizing the TPR

Equation 2 TPR and FPR for case (b) minimizing the FPR

### Decidability

The illustrations in Figure 2‑1 and Figure 2‑2 lead directly the definition of decidability. The two sets P and Q are neither contained one in the other nor distinct. If P and Q are distinct sets obviously a perfect classification with and exists.

If P is contained in Q () even the optimal classifier leads only to very limited performance and the problem might be regarded as undecidable in practice. The case (a) in Equation 1 can only use the relative probability of occurrence of a measurement (naïve Bayes Classification) while the case (b) is does not lead to any useful classification at all (this classifier would label all measurements with negative).

## Measurements

In practical applications like the problem stated in chapter the sets Q and P are not apriori known. To find a good classifier, the two sets (or more precise, their probability density) must be estimated by taking samples.

Within the StroKare project various measurements (studies) have been done or are planned for that purpose. However, this is a very challenging task. Practically it is very difficult to make sure that one samples the correct target distribution, e.g. changing devices, subjects, activity, etc. can strongly alter this distribution.

To illustrate this source of error the two sets (or densities) P’ and Q’ are defined analogously to chapter 2.1.2 describing the measurements.

In StroKare sampling the paresis case (members of P) are undoubtful very costly and one is likely to face different challenges than sampling the non-paresis case (members of Q). Therefore, the two cases and possible sources of errors are discussed separately in the following chapter.

### Paresis Samples

Measurements of patients suffering from a stroke (before and early after it occurred) will not be very numerous for practical and ethical reasons. However, the data is crucial and one might need to come up with a method capable of simulating such samples. This approach basically just relocates the problem to the development of such a simulation method. Although, finding a sufficiently good simulation method that is less expensive to create samples from on the cost of being maybe less precise is probably the only feasible way for the StroKare project.

Measurements for simulated paresis have been acquired in one of the performed study (“simulated paresis”,”study0”) and are for the time being the only source of information about the set P. It must be mentioned that a careful calibration/verification of the simulation method by the means of measurements of real patients has not been done yet and should be treated with high priority.

Let’s assume that the available simulated samples are members of the set P’ that is a “close” to the real set P see

### Time as dimension in the feature space

Form the context of the stated problem it is obvious, that it is almost for sure not possible to encounter a decidable classification problem if one looks only at sensor readings locally in time. Let’s say the sensors have a sampling frequency of ~200Hz the feature space

## Classification performance estimation

### Confusion Matrix

Given the algorithms output labels and the ground truth labels for a specific test data set, the occurrences of matches and mismatches can be counted in a 2x2 confusion matrix.

The performance of the algorithms can now be measured by the numbers of “false positives” and “false negatives” where the best possible performance on the given test data set is represented by both zero “false positives” and zero “false negatives”.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm output | Ground truth | paresis | no paresis |
| paresis | | Number of “true positives” (tp) | Number of “false positives” (fp) |
| no paresis | | Number of “false negatives” (fn) | Number of “true negatives” (tn) |

Table 1: Confusion matrix of classification

### Classification error estimation

From the presented confusion matrix, the following classification error estimations can be calculated. These error estimations can be compared among test data sets with differing total number of samples. *However, it must be taken in consideration that the confidence intervals of these error rate estimations largely depend on the number of samples in the test data (see 2.3.2.1)*

|  |  |
| --- | --- |
| Error name | Formula |
| True positive rate (TPR), Sensitivity |  |
| False positive rate (FPR) |  |
| Positive predictive value (PPV), Precision |  |
| Negative predictive value (NPV) |  |
| Accuracy |  |

Table 2: Classification error estimation

#### Classification error confidence

As mentioned in 2.3.2 the quality of the error estimations largely depends on the number of available samples in the test data set. Under the assumption of k independent samples of true “paresis” events and m independent samples of true “non paresis” events, the numbers of true positives tp and false positives can be used by Bayesian theory to calculated the distribution of the TPR and FPR estimations.

##### Prior distribution

Every knowledge from older analysis can be used (estimations of already performed study runs). However, if no prior knowledge is available, one can use a uniform distribution with values in the interval [0,1] for the rate estimations (TPR, FPR)

##### Model for tp, and fp

Assuming independent samples and an algorithm that classifies the k “paresis” events as such with a probability of TPR, the true positives are distributed according to a binomial distribution with n=k, and =TPR

##### Posterior distribution

Uniform prior distribution for TPR and FPR distribution lead to a posterior beta distribution B(p,q) with the following probability density

The parameters are p=1+tp and q=k-tp+1 and p=1+fp and q=m-fp+1 for the TPR and the FPR estimation respectively.

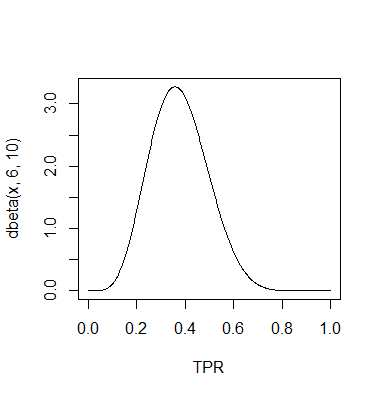
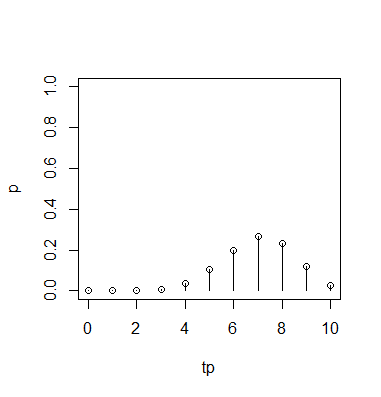


Figure 2‑3: (left) binomial distribution of true positives for a given TPR=0.3 and k=10 true ‘paresis’ events   
(right) beta posterior distribution of the TPR estimation for tp=6 true positives out of k=10 true ‘paresis’ events

##### Estimation and confidence intervals

The Bayesian estimation of TPR and FPR is given by the expectation value of the posterior distribution.

The quantiles can be calculated from tables of the cumulative distribution function for the beta distribution.

## Generalization

# Illustrations

The following chapters list detailed analysis of candidate classifiers and their structure. The used name is a identifier for a fixed compound of method and all parametrization. Further modularization is applied where useful.

### Level 1 modularization

Feature extraction

“Feature Extractor”

“Metric flow”

Classification

“Paresis detector”

“Anomalydetector”

Raw sensor   
data

asynchronous sampling

< 500 Hz

Metrics

synchronous sampling

30s

Decision

synchronous sampling

15min

Computergenerierter Alternativtext:
greclslon 
810 1214 16 18 
random permutations 

Figure 3‑1: Shows the estimations for 20 random choices of input data.[[2]](#footnote-2)

Short\_term\_threshold

arom\_abs

Long\_term\_threshold

v1.7.0

Figure ‑: Illustration of the orthogonal star in the parameter space where the performance measures were evaluated

# Annex

## Additional

## List of Tables

[Table 1: Confusion matrix of classification 8](#_Toc508196222)

[Table 2: Classification error estimation 8](#_Toc508196223)

## List of Figures

[Figure 2‑1 shows a fictive example of the set of positives P and set of negatives Q in a two-dimensional feature space 5](#_Toc508196224)

[Figure 2‑2 shows the two extreme classifier separation lines for the case (a) maximizing the TPR (left) and the case (b) minimizing the FPR (right). The classifier would map all measurements in the feature space lying below the red line to be positive (paresis). 6](#_Toc508196225)

[Figure 2‑3: (left) binomial distribution of true positives for a given TPR=0.3 and k=10 true ‘paresis’ events (right) beta posterior distribution of the TPR estimation for tp=6 true positives out of k=10 true ‘paresis’ events 9](#_Toc508196226)

[Figure 3‑1: Shows the estimations for 20 random choices of input data. 10](#_Toc508196227)

[Figure 3‑2: Illustration of the orthogonal star in the parameter space where the performance measures were evaluated 11](#_Toc508196228)

1. This is only true if the problem could be stated locally in time, see chapter 3.1.2 [↑](#footnote-ref-1)
2. The FPR is largely underestimated by this method [↑](#footnote-ref-2)