Guide sheet summaries DAM

Contents

[Guide sheet 1 2](#_Toc528253826)

[Statistical significance 2](#_Toc528253827)

[Feature thresholding 2](#_Toc528253828)

[Guidesheet2 3](#_Toc528253829)

[Linear/Quadratic discriminant classifiers 3](#_Toc528253830)

# Guide sheet 1

## Statistical significance

We classify data by analysing samples at selected features. But how do we select these features? For a feature to be useful, we need class-respective feature data distributions to be significantly different. Naturally, since we need to examine class-respective distributions, we need to select features using training data (sample classification is given).

Approach to ensure statistical significance:

1. Histograms allow us to visually expect data distributions. Note that since data does not necessarily contain equal amounts of samples from each class, it is important to normalize histograms prior to comparing them.
2. Boxplots are another way of visually inspecting a data distribution. Boxplots information includes: median (different from mean-> middle value not average), 25&75 percentiles, outliers. Notched boxplots are especially interesting because a notch in the box provides the 95% confidence interval of the median.

When examining our data, we can compare class-respective notched boxplots at a given feature to determine whether their medians are significantly different (no overlap in confidence intervals).

1. Student’s T-test is useful because it allows us to compare the means of class-respective datasets at a given feature. Note that t-test operates under the assumption that the datasets we compare are normally distributed and of equal variance. This prevents us from using the T-test ubiquitously.

T-testing returns a p-value which is the probability of observing the given result under the assumption that dataset means are identical. Therefore, a very low p-value is indicative of a significant difference between datasets. Usually we refute the null hypothesis that dataset means are identical if p-value is below 0.05.

Using this approach, we were able to determine that class-respective data differed significantly (high-discriminability) at feature 712.

## Feature thresholding

At a feature displaying high-discriminability, we can define a threshold used to separate data from each class. A simple classification scheme can then be established: Given a sample, check whether its value at FOI (feature of interest) is greater or less than threshold.

Since this model is simplistic and probably not very accurate, we need to be able to evaluate it. We use the following accuracy quantifiers:

; ;

Note that class weights must sum to 1 and that they must reflect the composition of training dataset.

Class error is more advantageous than classification error when training is not balanced (more of one class than another). Say we have 99% class A and 1% class B and we systematically predict A, then model would have classification error of 1% despite it being very bad! For such a model class error would be 50%! However, if 50% was class A and 50% was class B then classification error and class error would be equal (50%).

The weights in class error formula are useful because we can attribute importance to different types of error. For example, we could choose to ignore a certain type of error or emphasize another. Why emphasize a particular error class? Once again to reflect dataset. If we are using a big weight on a class comprised of a few data points it doesn’t make sense. It would allow anecdotal data to override the general consensus.

In practice, we established a 1-D model based on thresholding at feature 712. Initially threshold was chosen visually using histograms. Then we plotted model errors as a function of this value and chose the value that minimized class error as the optimal threshold.

We decided that we could scale up the model but that we needed to choose additional features carefully to avoid adding noise. This means repeating the process that we went through to select 712.

# Guidesheet2

## Discriminant classifiers

We wish to sort our dataset comprised of 597 observations (each comprised of around 2’000 variables) into 2 groups. A discriminant function calculates a weighted average of values of independent variables for each observation. The weights are selected so that the resulting average separates the observations into groups. High values of the average come from one group, low values of the average come from another group. The problem reduces to one of finding the weights which, when applied to the data, best discriminate among groups according to some criterion.

How do we determine these weights? We start with sums of squares matrices: We determine two matrices Stotal; Sgrouped representing the cumulative distance of observations from overall mean and group means respectively. We then define a third matrice Sdiff = Stotal - Sgrouped . The weight coefficients we seek are contained in the eigenvectors of Sgrouped -1 Sdiff .

Other matrices of interest are: overall covariance matrix T, within-group covariance matrix W, between-group covariance matrix A. These are all simply calculated based on the S matrices previously mentioned.

## Important considerations regarding discriminant analysis

Data needs to be screened prior to discriminant analysis. Checklist:

1. We need more observations per group than independent variables. Sorting 597 observations into 2 groups, since error group has prior 30%, say we expect around 200 observations in error class. This means we need to reduce our 2’000 features (variables) down to sub-200. We reduce data set dimensionality by factor of 10!
2. If class frequency in training Data is not representative of prior probabilities, we need to explicitly specify them.
3. Check training data for outlier observations in each class. Try to eliminate them because they can could disrupt classifier building.
4. Singular covariance matrices disturb classifier building. When this arises, it means one predictor variable (feature) is linearly dependent on the others. Solution is to reduce dimensionality of predictor variables used (features).

## Linear Discriminant Classifiers

* Equal class covariance matrices (W). W can be either full or diagonal. Diagonal means that only variance of features is considered whereas full means that we account for covariance between features.

## Quadratic Discriminant Classifiers

* Different class covariance matrices (W). Same full/diagonal implications as for linear case.

## Application

In practice,