Applied Statistical Methods in Animal Sciences

Peter von Rohr

2022-02-21

Contents

Preface

This document contains the course notes for

751-7602-00L Applied Statistical Methods in Animal Sciences.

General Developments

With the advent of **Big Data** (see [Wikipedia, 2019] and [Mashey, 1998] for a reference), it became clear that the importance of statistical methods to analyze the huge amounts of collected data would increase dramatically. Many modern statistical methods are only applicable due to the vast availability of cheap computing resources. The progress of the development that happens in the hardware manufacturing industry is often referred to by the term **Moore's Law**. This law was stated as a projection as early as 1965 by one of the founders of the Intel cooperation [Moore, 1965]. In a very general term, Moore's law says that the number of circuits that could be placed on a silicon waver would double every 18 months. In a derived version the law was interpreted in a way that the performance of computers would double every 18 months. Together with the high degree of automated production of the building blocks of a computer, the prices for a single unit of computation dropped dramatically. This development made it possible that the possibility to analyze large amounts of data with modern methods can be done by almost everyone. This created very many opportunities which are actively used by many business companies. Statistical methods used to be only used by academic researchers. Nowadays almost all important decisions in business companies are done based on supporting facts that are derived from analyzing market and customer data. With that it is clear that the importance of being able to use statistical methods to analyze data is almost ubiquitous and the knowledge of these methods can be very important in many different jobs or employments.

Where Does This Course Fit In?

This course gives a short introduction to a collection of statistical methods that I believe are relevant for a wide range of topics in Animal Sciences. These methods include

- Multiple Linear Least Squares Regression (MLLSR)
- Best Linear Unbiased Prediction (BLUP) which is called GBLUP when applied in the context of genomics
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Bayesian Estimation of Unknown Parameters (BEUP)

The above listed collection of statistical methods all happen to be illustrated around the same type of dataset. This dataset contains the genetic variants at many locations in the genome for a number of livestock breeding animals. Because there are many genetic locations considered in such a dataset and the locations are distributed across the complete genome, such a dataset is referred to as a **genomic** dataset. This type of dataset does appear in an area of livestock breeding which is called **Genomic Selection** (GS). GS was introduced in a seminal paper by [Meuwissen et al., 2001]. This very same paper is used as a building block to explain some of the statistical methods (MLLSR and BEUP) used in this course. Furthermore the same publication illustrates that some methods (MLLSR) are not suitable for analyzing certain aspects in a genomic dataset.

The time available for this course is just half a semester. This leaves very little time for the introduction of each topic. As a consequence of that each topic can only be presented very superficially and students are expected to work on their own during the exercise hours. Exercises consist of sets of problems related to each topic. Problems are often to be expected to be solved using the R programming language [R Core Team, 2018].

This version of the course corresponds to the fifth edition. With each additional iteration of the course, improvements are sought to be implemented. Hence any input from the students are greatly appreciated.

Course Objectives

The students are familiar with the properties of multiple linear regression and they are able to analyze simple data sets using regression methods. The students know why multiple linear regression cannot be used for problems where the number of parameters exceeds the number of observations. One such problem is the prediction of genomic breeding values used in genomic selection. The students know alternative statistical methods that can be applied in situations where the number of parameters is larger than the number of observations. Examples of such methods are BLUP-based approaches, Bayesian procedures and LASSO. The students are able to solve simple exercise problems applying BLUP-based approaches, LASSO and BEUP. The students are expected to use the statistical language and environment R [R Core Team, 2018].

Prerequisites

Because the data that is used in this course comes from genetics, a basic level of quantitative genetics is useful for this course. All statistical models will be presented in matrix-vector notation, hence some basics of linear algebra helps in understanding the presented material. Introductory chapters to both subjects (quantitative genetics and linear algebra) are included in these course notes, but will not be discussed during the lecture. These chapters are prepared for students who feel that they need more background. But this material is left for self-studying.

Chapter 1

Introduction

According to Wikipedia [Wikipedia, 2019], the term `Big Data` has been used since the 1990s. Some credit was given to John Mashey [Mashey, 1998] for popularizing the term. Nowadays `Big Data` is used in connection with large companies, social media or governments which collect massive amounts of data. This data is then used to infer certain conclusions about behaviors of customers, or followers or voters. The following subsections show a few examples of `Big Data`-applications.

1.1 US Presidential Campaigns

The presidential election campaigns of Barack Obama were examples of how `Big Data` was used to access behaviors of voters [Issenberg, 2013].

1.2 Health Care

A different example is the use of `Big Data` in health care. An overview of the use of `Big Data` in health care is given in [Adibuzzaman et al., 2017]. The collected health data is most likely not only used by research but also by insurance companies.

1.3 Face Recognition

The Swiss TV news show `10 vor 10` showed on the 7^{th} Feb. 2020 how a data journalist managed to build a face recognition system. The general idea how this system works is shown in Figure 1.1.

The main goal of the face recognition system was to be able to identify certain persons, in this case the politicians that had a picture on the platform

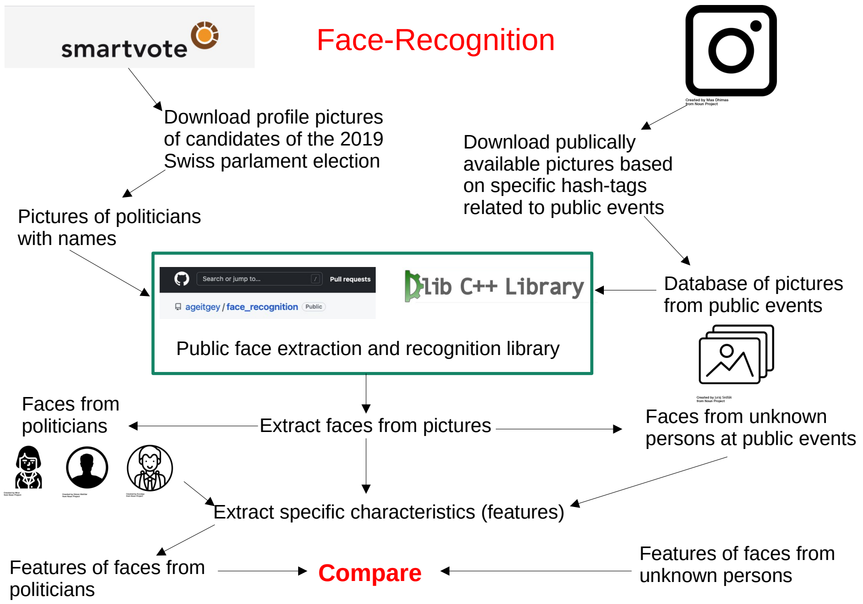
smartvote
Face-Recognition
Download profile pictures of candidates of the 2019 Swiss parlament election
Download publically available pictures based on specific hash-tags related to public events
Pictures of politicians with names
Search or jump to...
Pull requests
ageitgey / face_recognition
Public
Dlib C++ Library
Public face extraction and recognition library
Database of pictures from public events
Faces from politicians
Extract faces from pictures
Faces from unknown persons at public events
Extract specific characteristics (features)
Features of faces from politicians
Compare
Features of faces from unknown persons

Figure 1.1: Design of Face Recognition System

'smartvote', on random pictures obtained from the social media platform Instagram. The data used for the face recognition system can be split into two parts.

1. **Training** set. The training set consists of pictures showing different politicians. This dataset was downloaded from the politics platform smartvote. In a frist step the faces are extracted using a public software library. The training set which consists of the faces extracted from the pictures is used to establish a fixed relation between specific characteristics of the faces of politicians and their names. These specific characteristics of the faces are also called *features*. Most of these features consist of numbers which describe and quantify the faces shown on the pictures. Examples of such features might be the surface of the face, the surface of the hairs shown in the picture, the length of the mouth, etc.

2. **Test** set: The test set consists of 230000 publicly available pictures on the platform Instagram. The content of these pictures is a priori unknown. The same face extraction library as was used for the training data is also used on the public pictures with unknown content to filter out the parts of the pictures containing faces. The question that the face recognition system tries to answer is whether it is possible to identify any of the politicians from the training set on any of the Instagram-pictures. This

question is answered by a comparison of the features of the faces extracted from the instagram pictures to the features that were obtained from the faces of the politicians in the training set. If the feature comparison results in a match, the system suggests that we found a given person on one of the instagram pictures.

The complete story about the face recognition system is available under https://www.srf.ch/news/schweiz/automatische-gesichtserkennung-so-einfach-ist-es-eine-ueberwachungsmaschine-zu-bauen.

1.4 Feed Intake and Behavior Traits of Cows

In the recent past technologies based on computer vision have been introduced into agricultural applications. Two examples of such applications are

1. Estimation of feed intake of cows based on video data as described by [Chizzotti et al., 2015] or based on accelerometer sensors as shown by [Carpinelli et al., 2019]. The company Viking Genetics has developed the CFIT system which uses video camera data to measure cow individual feed intake on commercial farms ([Lassen et al., 2018]).
2. The EU-Interreg project “SESAM” aimed at predicting basic behavior traits from data obtained from sensors and from video recordings.

1.5 Conclusions from Examples

The above shown examples demonstrate that data can be used for very different purposes. Using just one source of data does in most cases not give a lot of insights. But when different sources of information are combined, they can be used to make certain predictions that influences our daily lives. Hence this kind of development is becoming a general interest to all of us. In what follows, we try to show that some of these methods have been applied for a long time in the area of animal science and especially in livestock breeding.