Feature selection and intelligent livestock management

Alsahaf, Ahmad

DOI:
10.33612/diss.145238079

Chapter 1

Introduction

This thesis contains two topics that are not directly related. It is therefore split in two parts. Part I, comprising the first two chapters, explores an emerging application area of machine learning and computer vision, namely, their use in the field of livestock breeding and management. In particular, we study the predictability, with machine learning models, of two size-related traits of domesticated pigs: their growth rate and their muscularity. Notwithstanding the specifics of pig production, and the traits of interest therein, we intend to present these chapters as part of a larger trend of using machine learning with livestock animals, and more generally, in agriculture. Therefore, in the first part of this introductory chapter, we broadly discuss these developments without emphasis on the case study of pigs.

Part II, the final chapter of the thesis, proposes a new feature selection method for classification problems. The proposed method, named FeatBoost, is a wrapper-based forward selection method which, as its name may suggest, uses boosting as a primary component in its functioning. We explain the basics of using boosting, or sample re-weighting, for feature selection in section 1.2.1 of this chapter. In chapter 4, the method is presented in detail and compared against existing approaches of feature selection on 17 benchmark datasets. In section 1.2 of this chapter, we argue generally for the importance of feature selection in modern machine learning applications, and discuss its connection to the concept of explainable AI (XAI).

What is common to both parts of the thesis, trivially, is the underlying use of machine learning models. Another slight similarity is the emphasis in Part I on model interpretability, and assessing the predictive ability of different groups of features (e.g. phenotypes and genotypes). This was done by means of analyzing feature importance scores. Beyond those elements, we do not draw more parallels between the two topics. Therefore, in the remainder of this introductory chapter, we treat each of those parts separately.

1.1 Machine learning in livestock science

Nowadays, the use of machine learning methods in scientific and engineering domains is very widespread. Interestingly, these methods are also gaining popularity within domains which used alternate approaches of predictive modeling in the past, or ones where large amounts of data are routinely collected overtime but not fully utilised. The breeding and production of livestock animals fits both of these descriptions.

Throughout the lives of farm animals - such as pigs, cattle, or poultry - large numbers of phenotypic, genetic, and auxiliary data are collected. Phenotypes, which are the observable and measurable traits of an animal, can include weight measurements at different stages of production, health status with respect to various diseases, litter information,[1] milk quality in dairy cattle, and egg size in laying hens. Genetic data includes any information that could quantify the genetic potential the animal, including pedigree records, genetic markers, and Estimated Breeding Values (EBVs).

Depending on farm practices, auxiliary data could be available for individual animals or groups, such as feed type and quantity, environmental factors, and farm conditions. The collection of this information, which has become commonplace in large-scale animal farming, creates large amounts of heterogeneous data that has the potential of being used to predict future animal performance with machine learning.

In the past century, the practice of livestock breeding relied heavily in its progress on developments in population and quantitative genetics [Oldenbroek and van der Waaij, 2014]. More recently, molecular genetic techniques such as marker-assisted selection, genomic selection, and genome editing have also become usable in livestock breeding [Yang et al., 2017]. Hence, the application of machine learning to livestock data does not signify the first occurrence of advanced data use in the field. However, as we shall show later, it represents a promising development.

1.1.1 Estimated Breeding Values

To appreciate the importance of data in livestock breeding, it is useful to understand its role in the core practice of breeding, which is to estimate the genetic potential of an individual animal relative to a population. A breeder's objective is to rank animals' fitness with respect to a certain trait, such as body weight, and then allow only the best ones to parent subsequent generations. The difficulty lies in the fact that most traits of interest in animal production are complex, and determined by

[1] This can include information about the litter itself, such as its size, or how an individual animal compares to its litter-mates, e.g., its size relative to the litter average.

multiple genes and external factors. Thus, they are not fully heritable. If a farmer were to choose the best animals strictly on a body weight criterion, there would be no guarantee that this will produce the heaviest offspring possible. Because this way of selection disregards other factors that may have caused the other animals to be underweight, like underfeeding or disease. Another issue is that some phenotype observations are not available for certain animals. For instance, a bull's genetic potential for siring cows with a high milk yield cannot be determined by measuring the bull's own performance [Oldenbroek and van der Waaij, 2014].

Therefore, breeding programs rely on genetic-statistical models of the animal which incorporate large volumes of data. These models take into account not only the animal's own performance, but also the performance of animals related to it, including its progeny, and environmental factors. They may also include the animal's own genotype and those of its relatives, as in the case of genomic selection [Oldenbroek and van der Waaij, 2014]. These models result in what is referred to as Estimated Breeding Values, or EBVs. The more accurate these estimates are, the better the outcome of the breeding program will be.

Estimated Breeding Values are typically modelled using mixed linear models, which in turn are solved with Best Linear Unbiased Prediction (BLUP) [Robinson et al., 1991; Garrick, 2010]. The purpose of said models and corresponding solution is to estimate on one hand the systematic effects on the observed phenotype - which can include herd, year, sex, age, and various environmental factors - and on the other hand, to estimate the additive genetic effects that contribute to variation of the same phenotype [Robinson et al., 1991; Garrick, 2010]. Such a model can be represented by Eq. 1.1.

$$y = X\beta + Zu + e \tag{1.1}$$

where y is a vector of measurements of a random variable, similar to the dependent variable (output) of a standard linear regression model. Vector β represents a set of parameters assumed to have fixed values (fixed effects), similar to the model coefficients in a linear regression model. Vectors u and e are unobservable random variables, or random effects, with zero mean, and variances given by Eq. 1.2.

$$\text{Var}\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \sigma^2 \tag{1.2}$$

Matrices X and Z are the incidence matrices that relate the effects in β and u, respectively, to y. The BLUP estimates to Eq. 1.1, $\hat{\beta}$ and $\hat{u}$, are given as the solution of the simultaneous equations in Eq. 1.3 [Robinson et al., 1991].

$$\begin{aligned} X^T R^{-1} X\hat{\beta} + X^T R^{-1} Z\hat{u} &= X^T R^{-1} y \\ Z^T R{-}1 X\hat{\beta} + (Z^T R^{-1} Z + G^{-1})\hat{u} &= Z^T R^{-1} y \end{aligned} \tag{1.3}$$

Table 1.1: An example of milk yield records [Robinson et al., 1991]

Herd	Sire	Yield
1	A	110
1	D	100
2	B	110
2	D	100
2	D	100
3	C	110
3	C	110
3	D	100
3	D	100

This model differs from a standard linear regression model - or a fixed effects model - by the inclusion of the random effects term, whose parameter values are to be estimated. In both models, the term e represents the random residuals, or error. The ordinary least squares solution to the fixed effects model, $y = X\beta + e$, is given by Eq. 1.4.

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{1.4}$$

An example of the difference between the two models, when applied in a breeding context, is given by Robinson et al. [1991]. The goal therein is to model the lactation yield of a group of dairy cows, as a function of the additive genetic effects of their sires, and the effects of the herds they live in; and to estimate the relative contribution of each of those effects. In the example, a cow is sired by either of four bulls, and comes from one of four herds, as shown in Table 1.1. Using Eq. 1.1 to model the yield, the genetic merits of the sires are represented by u, while the effects of the herd, environmental or otherwise, are represented by β. Therefore, the genetic effects are treated as random effects, and the herd effects are treated as fixed.

Robinson et al. [1991] showed that the BLUP estimates of Eq. 1.1 (given in Eq. 1.5) differ from the ordinary least squares estimates (Eq. 1.6), which would result if the genetic effects were considered to be fixed as well.

$$\begin{aligned} \hat{\beta}_1 &= [105.64, 104.28, 105.46]^T \\ \hat{u_1} &= [0.4, 0.52, 0.76, -1.67]^T \end{aligned} \tag{1.5}$$

$$\begin{aligned} \hat{\beta}_2 &= [100, 100, 100]^T \\ \hat{u_2} &= [10, 10, 10, 0]^T \end{aligned} \tag{1.6}$$

Intuitively, the estimates of the sire effects given by the ordinary least squares solution, $\hat{u_2}$, reflect that the cows sired by bull D, according to the available observations, have an average yield which is 10 units less than the cows from each of the other sires [Robinson et al., 1991]. The BLUP estimate, $\hat{u_1}$, takes additional information into account. For instance, the fact that bull C is represented by two cows results in a higher estimate of its genetic merit than bulls A and B, despite the yield averages of the cows sired by all three being equal.

The estimates of u in each of those cases are the EBVs of the corresponding sires, and can therefore be used to genetically rank them relative to each other. When this example is scaled up to the proportions of large breeding programs, the reliability of the estimates becomes dependant on a number of factors. For example, the EBVs of sires will be more reliable if their progeny is well represented in each herd or group. Similarly, a good estimate of herd or environmental effects requires that animals of similar breeds exist across herds or environments [Garrick, 2010].

In general, the accuracy of EBVs will depend on the amount of information used to compute them, i.e., the number of related animals whose phenotype measurements are available, the inherent heritability of the phenotype, and the accuracy of phenotype measurements and pedigree records. Even in the best case scenarios, however, an EBV will not fully predict the phenotypic performance of the animal. This is due to the complexity of production-related traits, and since many predictors of such traits occur throughout the animal's life, beyond the influence of its genetics. The linearity assumption in animal mixed models is another obstacle.

Given those limitations in the standard approach of evaluating the genetic potential of an animal, machine learning could improve breeding in at least two different ways. First, non-linear, model-free supervised learning algorithms could substitute mixed-models and BLUP in estimating the breeding values [Shahinfar et al., 2012]. Second, supervised learning could be used to test the efficacy of BLUP estimated breeding values in the task of directly predicting a phenotype. Moreover, it could combine EBVs with phenotypic, environmental, and other auxiliary observations that are recorded throughout the life of the animal to improve said predictions.

In chapter 2 of this thesis, we follow the second strategy. Combining EBVs, phenotype observations, and pedigree records, we use a random forest regression model to predict the age at which an individual pig reaches a desired slaughter weight. The predictive capability of each group of features is also analyzed using feature importance sores and group-wise prediction comparisons.

In addition to improving predictions which pertain to breeding and production objectives, machine learning could solve other issues in the livestock sector, since it is not dependant on inflexible animal models. For example, it could be used to solve predictive problems related to animal welfare and management Liakos et al.

[2018]. In chapter 3, we tackle a problem at pig farms which relates both to farming logistics and animal welfare. Namely, we develop an automatic and non-invasive method for estimating the muscularity of pigs, using a combination of computer vision and supervised learning.

1.1.2 State of the art

Since the advent of precision farming, both in livestock and crop agriculture, a number of studies have used machine learning or computer vision in livestock-related applications. We identified a few non-exhaustive categories of this usage, such as the analysis of sensor data, computer vision-based identification and behavioral analysis, and various production or breeding-related applications, such as disease detection and genomic selection. We highlight a few examples of these categories in the following sections.

Sensor data

The model-free aspect of machine learning models is a convenient approach to utilize data that is generated by a variety of sensors, including those which are proprietary to the livestock industry. Dutta et al. [2015] used GPS collar sensor data of cattle to automatically classify their behavior among a number of behavioral classes, such as grazing, ruminating, resting, or walking. The classes in turn were determined by a framework of unsupervised learning applied to features extracted from time series of the collar data. Gorczyca et al. [2018] used neural networks, gradient boosted machines, and random forest to predict surface and internal temperatures of piglets - which are relevant indicators of stress level - based on a set of environmental factors. With decision trees, Pegorini et al. [2015] used fiber Bragg grating sensor data to classify the chewing patterns of grazing animals. The identification of such patterns can in turn identify the type of forage being consumed by those animals, which has several implications on the production of those animals, and their impact on their pastoral ecosystem.

Identification, size, and behavior

At a farm, computer vision can be a powerful tool to identify individual animals, study their behaviours, or determine some of their characteristics, like weight and size. Despite the fact that some of these tasks could be performed by other means, for instance, weighing by scales or assessing behavior by human operators, computer vision has the advantage of decreasing intrusiveness, which can reduce the

stress levels of animals and improve their well-being. More so, in the case of behavioural analysis, the effect of a human observer on the observed behaviors is reduced [Porto et al., 2013].

For instance, computer vision pipelines based on convolutional neural networks have been used for the identification of cattle based on top-view RGB images [Andrew et al., 2017], and side-view RGB and thermal images [Bhole et al., 2019]. Systems of visually detecting behaviors of animals by tracking their movements or postures have been developed for pigs [Ferre et al., 2009; Shao and Xin, 2008], chicken [Kashiha et al., 2013; Leroy et al., 2006; Pereira et al., 2013], and cattle [Porto et al., 2013; Cangar et al., 2008].

Image-based systems for the purpose of determining mass or other size characteristics have been developed for chicken [Amraei et al., 2017], cattle [Kawasue et al., 2017; Nir et al., 2018], and pigs (see section 3.4 for examples). Other applications include those which were originally developed for humans, and later adapted for livestock use, like pain level estimation through facial analysis for sheep [Lu et al., 2017], and facial recognition for cattle, horses, and pigs [Lu et al., 2014; Trokielewicz and Szadkowski, 2017; Hansen et al., 2018].

Breeding and other applications

In addition to supplementing the statistical models used to determine breeding values (Section 1.1.1), machine learning and computer vision can be used to improve livestock production in indirect ways. For instance, in disease detection [Lee et al., 2017; Sharifi et al., 2018; Zhuang et al., 2018], meat quality assessment [Taheri-Garavand et al., 2019; Barbon et al., 2018] , predicting conception outcomes [Shahinfar et al., 2014; Hempstalk et al., 2015], and modelling egg production [Morales et al., 2016; Yakubu et al., 2018].

Lastly, machine learning could be used as an alternative to traditional statistical models in genomic selection. In other words, they could be used to determine genetic markers associated with certain phenotypes. This type of modelling typically suffers from the curse of dimensionality, as the number of markers is significantly higher than the number of genotyped individuals. While traditional statistical models address these issues with shrinkage and regularization, non-parametric machine learning methods can offer other ad-hoc solutions to those issues [Do et al., 2014, and references therein].

1.2 Feature Selection and model interpretability

Modern supervised learning methods often contain implicit or explicit forms of feature selection or dimensionality reduction. For example, models based on decision-trees partition the sample space by recursively finding the best feature splits with respect to an output impurity function. This leads to an internal hierarchy of the input features that could be used to derive a-posteriori feature importance scores [Loh, 2011; Kazemitabar et al., 2017]. In kernel methods, like support vector machines, the decision function depends only on a subset of discriminant samples; the support vectors. Weights of the decision function could also be used for feature selection [Guyon et al., 2002]. Traditional linear models can perform feature selection by means of regularization [Fonti and Belitser, 2017]. Broadly speaking, the success of any predictive model could be explained in part by its ability to discriminate between useful and superfluous data.

Therefore, from a strictly predictive performance point of view, it could be shown that explicit feature selection as a pre-processing step does not provide a significant improvement in a number of supervised learning scenarios [Post et al., 2016]. This is further evidenced by the fact that in feature selection, the selected subsets are often tested on simple learning models to emphasize the positive effect of removing irrelevant or redundant features. Similarly, benefits such as savings in computational time and storage-space are subject to the availability of resources, and could also be debated. Even in such cases, however, the analytic benefits of feature selection remain significant.

In all but few supervised learning problems, there is an underlying task of model or process understanding; the first steps of which are finding the features that affect the prediction the most, or finding a minimal subset of features that define the predictive model[2]. Since modern machine learning applications contain an ever growing number of features, intuitive understanding of prediction outcomes has become virtually impossible before some form of dimensionality reduction. Other aides of model interpretation, like data visualization, are also made easier by reducing the number of features.

Furthermore, in certain biomedical applications, knowledge discovery is the primary task of feature selection, more so than improving the prediction outcome or increasing computational efficiency [Borboudakis and Tsamardinos, 2019]. In gene expression studies, feature selection is used to discover the genetic networks associated with diseases [Tabus and Astola, 2005]. In other biomarker discovery studies, the workflow relies heavily on feature selection, as the studies often begin -

[2] An example of supervised learning tasks without inherent analytic questions are those in which humans outperform models; like visual object recognition.

for practical limitation - with large feature, small sample raw data [Christin et al., 2013]. Following successful biomarker discovery with feature selection, larger sample datasets with the reduced number of features could be collected for further study. In those scenarios and others, feature selection is an integral analytic component, and not an optional pre-processing step.

Another benefit of feature selection is in applications where the acquisition of features is costly. In such cases, it is useful to identify if a costly feature happens to be irrelevant or redundant [Early et al., 2016; Bolón-Canedo and Alonso-Betanzos, 2019]. For example, in medical data, a feature could be associated with a clinical test, which could be either expensive, inconvenient to patients, or both.

Moreover, feature selection is a valuable tool in the context of explainable AI (XIA). In XIA research, predictive models are made explainable or interpretable by either of these broad approaches: 1) An ante-hoc approach, in which models that are considered understandable to humans, such as decision trees and linear models, are chosen for a given application; 2) A post-hoc approach, where understandable approximations to black box models are used as the medium of explainability [Lakkaraju et al., 2017; Sokol and Flach, 2020]. In either of those approaches, a parsimony of input features is preferred, if not required, for the models to be truly explainable to humans [Sokol and Flach, 2020]. For instance, a decision tree is easy to understand and visualize only if it was of limited depth.

Due to their importance, a large number of feature selection methods have been developed over the years. And since reducing the number of features in any dataset can be arrived at with principally different approaches, the categorization of feature selection methods is not fixed. For instance, incrementally selecting the most important features, or discarding irrelevant and weakly relevant ones both lead to feature selection. Those differences lead to a categorization of methods that is based on search strategy [Tang et al., 2014]. Popular examples include forward selection, backward elimination, and genetic algorithms.

Another common taxonomy of feature selection methods pertains to how the selection procedure and the associated supervised learning task are connected. Methods that rank features independently of prediction performance are referred to as filter methods, whereas methods that evaluate the performance of feature subsets on a given predictor are called wrapper methods. The former are considerably faster but less accurate than the latter. Embedded methods reduce the number of features in conjunction with solving the prediction problem, by using an objective function that penalizes large numbers of features [Guyon and Elisseeff, 2003; Dash and Liu, 1997; Tang et al., 2014]. Methods that combine elements of wrappers, filters, or embedded methods are known as hybrid feature selection methods.

Feature ranking is a closely related procedure to feature selection. Unlike feature

selection, feature ranking algorithms give a weighted score and a linear order to all features instead of selecting a subset of them. Many feature selection algorithms are based around an auxiliary procedure of feature ranking or weighting. From another perspective, feature selection could be seen as a special case of feature ranking, where the given weights are binary [Molina et al., 2002].

In chapter 4, we propose a novel wrapper feature selection algorithm, which modifies the feature importance scores derived from a tree ensemble model by means of sample re-weighting, or boosting. As a preamble to that chapter, we discuss in the following section the relationship between sample re-weighting, feature importance scores, and feature redundancy.

1.2.1 Sample re-weighting and feature importance

Feature importance scores of tree-based ensembles are powerful starting points for feature selection [Tuv et al., 2009]. This is largely due to the following factors: First, the versatility of those models; being scale invariant, scalable to large datasets, and able to handle numerical, categorical, and missing data. Second, the fact that the intrinsic importance scores can be derived at no additional cost over that of model training.

For those scores to be used reliably for feature selection, however, one of their main shortcoming, namely that of feature redundancy, must be overcome. In this section, we show an illustration of this redundancy problem, and the way sample re-weighting could be used to mitigate it. For that purpose, we use an artificial dataset with three informative variables, and three linearly separable classes, as shown in Fig. 1.2.1. The green and red classes contain 100 samples each, while the blue class contains 20 samples. The red and green classes are fully separable across either the X_1 or X_2 dimensions, whereas the blue class is separable from the other two across X_3 only. It is easy to show that all classes are fully separable with either of the feature pairs $\{X_1, X_3\}$, or $\{X_2, X_3\}$.

Since there are more red and green samples than there are blue ones, one should expect a feature ranking algorithm to rank either X_1, X_2, or both, higher than X_3. In practice however, this depends on weather the feature ranking algorithm can account for the redundancy between X_1 and X_2. In tree-based ranking specifically, the outcome of the ranking can change depending on the classifier itself, its hyper-parameters, and how feature importance is defined within it.

To examine this, we train different tree-based models with this data set, and observe the resulting feature importance scores and rankings. Namely, a decision tree with a maximum depth of three, an Extra-Trees classifier (a parallel tree ensemble), and an XGBoost classifier (a sequential tree ensemble).

Figure 1.1: Sample dataset with three informative features, two of which, X_1 and X_2, being redundant for classification.

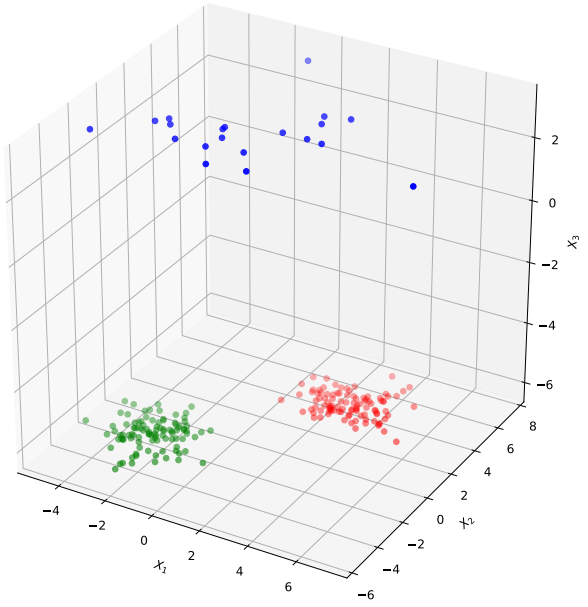

Table 1.2: The feature importance scores from different tree-based classifiers on a sample dataset with feature redundancy.

model name (configuration)	normalized importance scores					
	non-weighted samples			weighted samples		
	X_1	X_2	X_3	X_1	X_2	X_3
Decision tree (max depth = 3)	0	1	0.48	0.36	0	1
Extra trees (20 trees)	0.83	1	0.63	0.55	0.54	1
XGBoost (20 trees)	0.86	1	0.70	0.32	0.08	1

If the resulting importance scores (Table 1.3) were to be used for feature subset selection, e.g., by selecting two out of three features for optimal classification, then the scores from Extra-Trees and XGBoost would lead to sub-optimal choices.

In a decision tree, the importance of a feature is defined as the total value of a node-splitting criterion (e.g. Information Gain, Gini Index) that the feature is responsible for. And unlike XGBoost and Extra-Trees, which optionally employ feature sub-sampling and bootstrapping, a single decision tree is fully deterministic.

For that reason, in a single tree, features X_1 or X_2 have the highest value of the impurity reduction criterion at the tree stump. Then, at the second level, X_3 becomes the most important feature, because the discriminating power of X_1 and X_2 was accounted for at the first split by one of them. This leads to X_2 and X_3 having non-zero importance scores (Table 1.2), which in this simplified example, leads to selecting an optimal feature subset for classification.

In Extra Trees (or equivalently Random Forest), which are ensembles of parallel trees, the importance of a feature is defined as the sum or average of the splitting criterion that a feature causes across all trees. Due to feature sub-sampling and bootstrapping, which are integral components behind the generalization ability of such methods, the importance scores of two or more redundant features will be spread among them. In such cases, selecting features from the resulting ranking can lead to undesirable outcomes, namely, by selecting two features that are redundant to each other, X_1 and X_2, and ignoring an informative one, X_3.

In XGBoost (or equivalent boosting methods), importance scores are derived in a similar manner. We observe the same undesirable outcome of X_1 and X_2 having the highest scores. The abundance of configuration options available in XGBoost, or similar packages, could also lead to inconsistencies in the importance scores and corresponding feature ranking. For instance, parameters which randomize the trees, like node or tree-level feature sampling.

In this simplified example, the feature scores of a single decision tree give a result that is consistent with the feature selection objective. However, single decision-trees do not perform well on complex datasets, and are unstable with respect to changes in training samples [Li and Belford, 2002]. Therefore, neither their classification performance nor their feature importance scores can be relied upon in practical scenarios.

A boosting wrapper, like the one in the proposed method, can address the redundancy problem of tree ensemble importance scores, while retaining their other advantages. We demonstrate this by applying the weighting strategy in Eq. 1.7, the same strategy used in multi-class AdaBoost [Hastie et al., 2009]. Namely, after the initial importance scores are obtained, the model is retrained with the best feature, and the misclassified samples are identified and up-weighted. Then, the model is retrained with all features and the newly computed sample weights, and the corresponding feature importance scores are derived. If the top ranking feature from each step, before weighting and after weighting, is added to the selected subset, this leads to the feature pair $\{X_2, X_3\}$ being selected in all cases.

Table 1.3: The feature importance scores from different tree-based classifiers on a sample dataset with feature redundancy.

model name (configuration)	normalized importance scores											
	non-weighted samples						weighted samples					
	X_1	X_2	X_3	cX_1	cX_2	cX_3	X_1	X_2	X_3	cX_1	cX_2	cX_3
Decision tree (max depth = 3)	0	0	0	0	1	0.45	0.85	0	0	0	0	1
Extra trees (20 trees)	0.77	0.83	0.76	0.66	0.57	1	0.76	1	0.70	0.93	0.67	0.95
XGBoost (20 trees)	0.69	1	0.24	0	0	0	0.32	0.08	1	0	0	0

$$\begin{aligned} \alpha &= \log \frac{1 - err}{err} \\ w_j^{i+1} &= w_j^i \cdot exp(\alpha);\ \forall\, j = 1, \cdots, n \\ w_j^{i+1} &= \frac{w_j^{i+1}}{\sum_{j=1}^{n} w_j^{i+1}};\ \forall\, j = 1, \cdots, n \end{aligned} \tag{1.7}$$

where err is the classification error from the previous iteration and ω_j^i is the sample weight for sample j at the i^{th} iteration.

To show this effect under a condition of additional feature redundancy, we repeat the previous experiment with a dataset modified by adding a linear copy of each of the initial variables as such: $\{X_1, X_2, X_3, cX_1, cX_2, cX_3\}$. The results in Table 1.3 show that following that aforementioned rule, the selected subsets with a decision-tree, Extra-Trees, and XGBoost are $\{cX_2, cX_3\}$, $\{cX_3, X_2\}$, and $\{X_2, X_3\}$ respectively, which all lead to optimal class separation.

1.3 Thesis Outline

The remainder of the thesis is divided as follows: Part I, which consists of two chapters, gives two examples on applying machine learning and computer vision to livestock problems. In chapter 2, we use random forest and multiple linear regression to predict the age at which a pig reaches 120 kilograms; a preferred slaughter age. In chapter 3, we apply computer vision, using depth images captured by a Kinect camera, to estimate the muscularity of pigs. Part I, or chapter 4, is the part of the thesis concerned with feature selection, where we introduce a novel boosting-based feature selection algorithm. Finally, chapter 5 is the thesis outlook where we discuss some implications of the research conducted in this thesis and propose future research directions.

Part I

Machine Learning And Computer Vision for Livestock