Classification model for the evaluation of Holstein sires

Bárbara Paola Alcántara Vega

Benjamín Valdés Aguirre

Abstract—The main objective of this report is to break down the analysis behind the model developed to classify levels of acceptance for 17 linear conformation characteristics in bulls within the *SelectSires* database for the Holstein sire breed. The proposed model includes current limitations and proposals for improvement, since the model currently has an average certainty of 85% for the training and testing of each of the classes.

Keywords—*sire, gradient descent, features, classes*

1. Introduction

Sire semen selection is a common practice of dairy farms to ensure progeny with characteristics that maximize productivity while maintaining the health of the herd. This is the reason why companies have developed an extensive catalog with detailed data on the physical and genetic characteristics of each bull they offer, in order to give producers a selection of the bulls that most benefit production based on characteristics already present within the dairy herd. However, many producers don't have the time or knowledge to do an extensive analysis of the characteristics of the bulls they could choose, and end up choosing bulls that, although good, do not maximize the productive performance of the herd, and in the worst case, if only one or two bulls are chosen for the entire herd, the herd ends up with consanguinity problems and further expenses.

Normally, the evaluation and selection of livestock requires a certified expert for the breed that it's intended to work with, but with recent advances in machine learning, it is possible to automate the evaluation process not only through scoring, but also with image analysis, allowing producers to carry out an appropriate evaluation of their livestock and their selection of bulls without having previous experience of the breed in terms of linear conformation traits.

In this context, bull selection with machine learning allows bulls to be classified based on their score; very low, low, high and very high. Considering various attributes such as designations and guidelines to adapt to the needs of each producer; such as optimization in the use of forage feed, grass feeding, gender selection for progeny, and more; the applications of Artificial Intelligence and machine learning are becoming increasingly accessible, allowing producers to make informed and competitive decisions more easily. Not only allowing better selection of genetic material, but also an insight to the factors that best suit them, improving their dairy herd in terms of production and health.

2. Dataset's description

The bull sire database, which serves as the basis of the project, is derived from public access to the sire catalog containing characteristics organized and classified by the company *SelectSires*, for the selection and identification of its sires for the buyer and even for decision making when it comes to getting new genetics for the consumer. They handle a variety of breeds, including the Holstein breed, which is the focus of this project. The reason this company was selected for acquiring the data base is that this database is one of the few open source data available on the internet that contains the linear traits for the Holstein breed as specified in the Holstein Association USA, Inc. Linear Trait Manual, which is the standard in cattle grading and classification. The database relates directly to the attributes that qualify a sire for different needs.

3. Features

The data set includes 161 columns, each representing an important characteristic of the promoted sire. However, in order to simplify the analysis, 17 main characteristics were selected as to universally evaluate livestock and their main identifiers:

3.1. Main identifiers

- **NAAB**: Serial number for the stallion
- **Lineups/designations**: Class to which the stallion belongs

3.2. Main traits

- **PTAT**: Body depth
- **STA**: Stature
- **STR**: Strength
- **DFM**: Dairy form
- **RUA**: Rump angle
- **RLS**: Rear legs from lateral view
- **RTP**: Rear teat placement
- **FTL**: Front teat length
- **RW**: Rump width
- **RLR**: Rear legs from rear view
- **FTA**: Hoof angle
- **FUA**: Front udder insertion
- **RUH**: Rear udder height
- **RUW**: Rear udder width
- **UCL**: Udder cleft ligament
- **UDP**: Udder depth
- **FTP**: Front teat placement

Additionally, it is important to take into account that each characteristic has a percentage of heritability and a level of importance in average health and milk production.

Fig. 7 shows the percentages mainly considered in milk production.

Figure 1. Relevance of each group of traits in health and milk yield [**(Holstein Canada: Services - Classification, 2015)**].

4. Data cleaning

The database contains 161 features, and not all the data they contain is numerical. So the first thing that had to be done is to swap the columns that did not provide relevant or clear information in order to work with numerical data that can be used in learning and training the model. Furthermore, although the database instances are already found randomly, randomization is necessary between each training and test so that the model does not incorrectly fit the data and can predict scores regardless of the order of the data. data provided.

Equation 4, shows how the data normalization was carried out:

$$X_{norm} = \frac{X - min(X)}{max(X) - min(X)} \tag{1}$$

This way, we keep the range of each feature on a standard scale, since each feature has data whose range is variable with respect to

the others. This numerical stability ensures that each characteristic is within a comparable scale and thus prevents higher values from dominating the training process.

5. Correlation analysis and feature selection

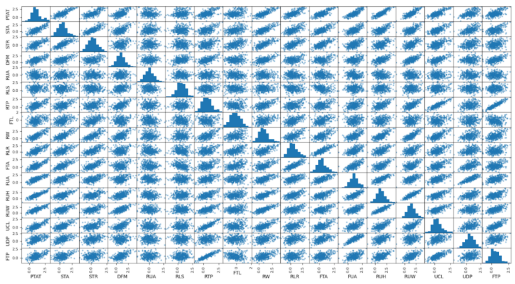

Figure 2. Scatterplot depicting the 17 selected features.

Thanks to a scatterplot of each linear conformation trait, we can observe that the relationship between data involved with udders and legs are closely related, this relation makes sense as these are two dependent characteristics for the development of good breast health in the Holstein breed. This is why farmers tend to focus on these characteristics and tend to be misguided by data such as the one showed on figure 1, however, although it helps a lot in terms of productivity, only focusing on udders and legs could result in other problems, such as lack of strength in the legs and accidental omission of the suspensory tendon, two very important characteristics for the health of the animals as individuals and a greater number of productive days. The characteristics describing ppm solids in milk, pedigree, or dollars per liter of milk on average were not maintained, since in addition to the fact that the Holstein breed is well known as a producer of fluid milk and not of dairy products, the focus on health and productivity is a priority in improving current dairy production nationwide in Mexico.

6. Data set split

Dividing the database into three parts, allows training, validation and testing of the results. For this to work, it is important that the proportions of this division are 60%, 20% and 20% respectively for each part. All this with the aim of training the model with the greatest amount of data available possible and consequently, evaluating the model with respect to its performance as if new data were presented, thus simulating a selection in a real scenario.

7. Base model

The project uses a linear regression model built from its basic elements without using existing machine learning libraries, this with the aim of deepening the understanding of the functioning of these processes during data analysis. This model is mainly used as a binary selection tool, but can also be used for multi-class classification. Due to the number of classes that need to be predicted, the cycle must be repeated for each of the 12 classes. In each cycle, the model uses the strategy called One-vs-Rest.

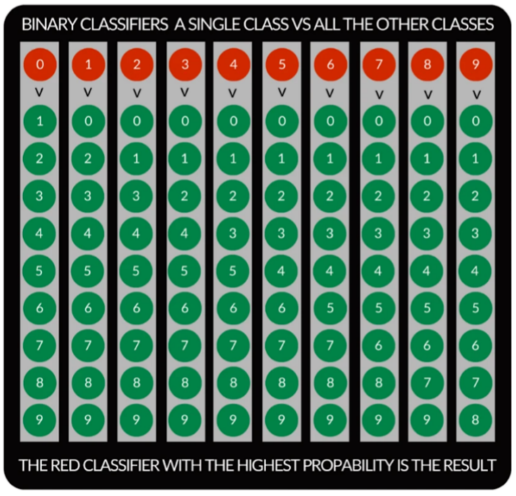

Figure 3. Basic One-vs-Rest depiction [**(Psaltakis et al., 2024)**].

This approach involves training a binary classifier for each class, where the classifier of a specific class is trained to distinguish that class from all other classes combined [**(OneVsRestClassifier, 2025)**]. This way, instead of trying to predict the 12 classes at the same time, it focuses on a single class to treat the rest of the 11 as a class by itself, with the aim of determining whether the data belongs to that class by applying the hypothesis function, and only then, after having repeated the process for three classes, it finds the index of the class with the highest probability of corresponding to the analyzed data.

8. Hypothesis function

The hypothesis function is used to predict the probability that an input data from the data belongs to a specific class or combination of classes.

Equation 4, shows the sigmoid function used:

$$h\theta = \frac{1}{1 + e^{-\theta T x 1}} \tag{2}$$

Because its domain remains in all real numbers, and its range is (0,1), even if a large positive or negative number is received, the result will always be between 0 and 1. It is for this same reason that the One-vs-Rest method is suitable for selection, since it can only determine between 0 and 1.

9. Cost function

The cost function quantifies the difference between the predicted probabilities and the current class to which they belong, so if we minimize this difference during training, we allow the model to fit its parameters to improve certainty. Since we have 12 classes, the cost function is used for each of them independently.

Equation 4, shows the cost function used:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta(x^{(i)})\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta(x^{(i)})\right)\right] \tag{3}$$

10. Gradient descent

The gradient descent allows the model to learn from the data by minimizing the cost function. Which means that the gradient of the cost function with respect to the parameters (theta) is calculated first, and subsequently, each parameter is updated in the direction that reduces the cost, thus increasing the accuracy of the predictions.

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left(\left(h_\Theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}\right) \tag{4}$$

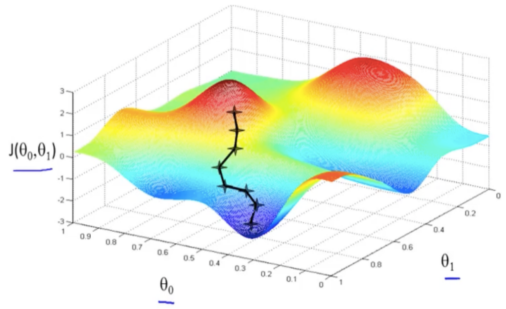

Figure 4. Illustrative depiction of gradient descent [**(DBCerigo, 2018)**].

10.1. Why not use back-propagation?

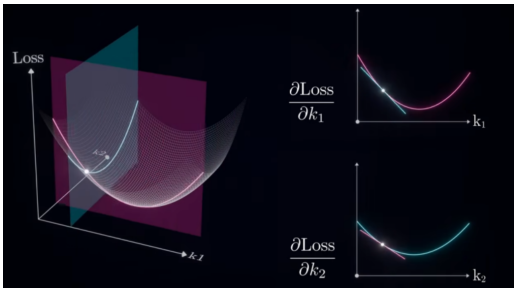

Figure 5. Illustrative depiction of back-propagation [**(Artem Kirsanov, 2024)**].

As it is shown further on, the current model works well with a single perceptron and logistic regression, and thus a more complex model with a neural network could benefit much more from back propagation than this demonstration of theoretical and practical knowledge, as the derivatives slicing each loss in the three dimentional curve would be more heavy to process than what a simple vector with updating parameters (*theta*) for the same result.

11. Training, validation and testing

The model learns by adjusting parameters using gradient descent, minimizing the error in the predictions for each epoch. The One-vs-Rest strategy is used to train the model separately for each class, while the cost is monitored to observe progress.

In addition, the behavior and progress of the model is also evaluated with a separate validation set of training data. This is with the aim of recording how well the model generalizes to data never seen before. In this way, monitoring validation certainty allows the model to be adjusted while preventing overfitting during training.

After training, the model is tested with completely new data (the test set), to evaluate its generalization ability, while the final certainty test tells us how likely it is that the model behaves correctly in a real scenario.

12. Base model results

After training the model and testing it, the results show that the model achieved an average certainty of 85% in training. Furthermore, a reduction in error can be observed by half after 300 epochs.

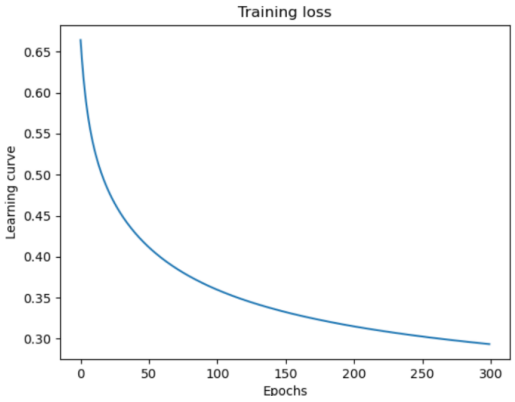

Figure 6. Mean loss curve.

The error starts relatively high around 0.66, and as the epochs increase, the loss drops to stabilize around 0.29, showing that the model is successfully minimizing the error. Since the curve does not have oscillations or sharp increases, the model learns and reduces its error with each iteration without presenting obvious signs of overfitting (without increase in loss/error), in addition, the convergence is progressive with a marginal improvement around the 250-300 epochs, which means that training for more epochs would provide minimal improvement in error.

Table 1 Regarding the certainty results for each of the classes:

Table 1. Each feature's accuracy

Feature	Accuracy
PTAT	0.898
STA	0.874
STR	0.882
DFM	0.840
RUA	0.715
RLS	0.785
RTP	0.914
FTL	0.777
RW	0.835
RLR	0.825
FTA	0.869
FUA	0.898
RUH	0.869
RUW	0.887
UCL	0.882
UDP	0.911
FTP	0.882
Mean accuracy	**0.855**

Please note that FTL has a low level of accuracy.

The training graph, together with our observed certainties table, indicate that there is no presence of overfitting. Fortunately, none of the characteristics exceed 95%, which would indicate a suspicious level of accuracy, although it is important to note that the front teat length (FTL) does not have a very high level of accuracy. This makes sense, since nipple length does not usually influence classes, except for *RobotPRO*, where nipple length facilitate but does not determine automatic milking, meaning that having long or short nipples does not completely prevent automatic milking.

The model seems to work without signs of overfitting using 300 epochs, where the training process looks quite successful with a validation certainty level of approximately 91%, and a validation loss of 0.3, value from which, there is no longer a significant difference in the decrease in error.

Fig. 7 shows the comparison between Training Loss/Training Accuracy and Validation Loss/Validation Accuracy.

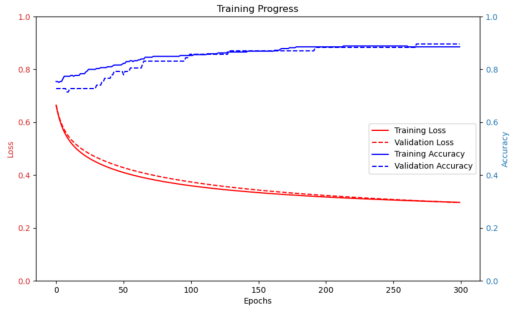

Figure 7. Graph with training progress for validation and loss.

The model learns effectively from training data while generalizing well to new data not seen before (validation set). So there are no signs of overfitting or severe underfitting. Furthermore, the certainty measure shows us that 91% of the time a correct classification is being made for the classes.

The smooth and constant curve shows us that the parameters for the "learning rate" are appropriate, since a very high learning rate can cause disturbances in the curve, and a very low learning rate can cause a much slower convergence.

13. References:

OneVsRestClassifier. (2025). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

Psaltakis, G., Rogdakis, K., Loizos, M., & Kymakis, E. (2024). One-vs-One, One-vs-Rest, and a novel Outcome-Driven One-vs-One binary classifiers enabled by optoelectronic memristors towards overcoming hardware limitations in multiclass classification. Discover Materials, 4(1). https://doi.org/10.1007/s43939-024-00077-7

Select Sires. (2024). Default. https://www.selectsires.com/home

3Blue1Brown. (2017, October 16). Gradient descent, how neural networks learn | Deep Learning Chapter 2 [Video]. YouTube. https://www.youtube.com/watch?v=IHZwWFHWa-w

DBCerigo. (2018). talks-presentations/deep_learning_without_hype/On Why Gradient Descent Is Needed - Without The Hype.ipynb at master · DBCerigo/talks-presentations. GitHub. https://github.com/DBCerigo/talks-presentations/blob/master/deep_learning_without_hype/On%20Why%20Gradient%20Descent%20Is%20Needed%20-%20Without%20The%20Hype.ipynb

Holstein Association USA, the world's largest dairy cattle breed association. (2024). https://www.holsteinusa.com/

Holstein Canada: Services - classification. (2015). Holstein Canada. https://www.holstein.ca/Public/en/Services/Classification/Breakdown_of_Traits

Huang, Z., Zhu, X., Ding, M., & Zhang, X. (2020). Medical image classification using a Light-Weighted hybrid neural network based on PCANET and DenseNet. IEEE Access, 8, 24697–24712. https://doi.org/10.1109/access.2020.2971225

