Introduction to Machine Learning Assignment 1

Matthijs Prinsen (S4003365)
Marinus van den Ende (s5460484)
Group 54

November 28, 2024

Part I - Feature Type Examples

Continuous features (5)

Continuous features have a continuous range for their values; they are usually real numbers (floats).

- **Age**
  Age is a continuous value that increases over time (in a continuous range). It may be approximated with the current year (which would then be a discrete value).
- **Salary**
  Salary has a range of numbers that are uninterrupted (e.g., one person may earn 10 euros an hour, another may earn 10.12 euros an hour).
- **Price**
  Price is on a continuous range, similar to the aforementioned salary.
- **Height**
  Height is a measure of length represented in centimeters, meters, or feet (in the case of Tarantino).
- **Weight**
  Weight is a measure of mass, represented in grams, kilograms, or pounds.

Discrete features (3)

Discrete features are characterized by gaps between possible values, where there is a distinct separation (step) from one discrete value to another.

- **Birth rate**
  Birth rate is on a discrete range of variables, as there can be no measurement smaller than 1 (no half babies allowed here).
- **Email count**
  The number of emails in an inbox is a counting number {0, 1, 2, 3,...}. You cannot have half an email. This is also true for most sets where you count objects. For example, if you have 10 apples, you could have half an apple, but then no one would want it. So you have 10 apples nonetheless.

- **Number of students present**
  You will always have a counting number {0, 1, 2, 3,...} here, as you cannot have a negative, half, or partial students.

Categorical features (2)

Categorical features are used to group information with similar characteristics using categorical labels.

- **City district**
  City districts refer to predetermined areas and are defined by their boundaries. The values are the names of the areas or a way of representing them (e.g., with magic numbers: "1" could represent "city center").

- **Sex**
  Sex will either be male or female—two categorical values.

Part II - Classifiers vs Regressors

Deciding between using a classifier or regression model depends entirely on the type of target variable you want to analyze. If your data is discrete or categorical, you use classifiers; otherwise, if it is continuous, you use regressors. We will use the example of Housing Prices to argue "only one of the possible options."

We can manipulate housing prices to be either continuous or categorical. If we wanted to predict housing prices, we would keep the data as continuous, but if we wanted to classify the houses into affordability bands, we could have categories like {*Cheap, Moderate, Affordable, Pricey, Expensive, Ludicrous, Continental*}.

Having the data as categorical allows the model predictions to be easier to grasp and intuitive for humans. This is useful for real estate agents and house buyers, as viewing the data as *Cheap* or *Continental* is much easier to interpret than continuous values like $214,206.94.

While regression offers the opportunity to predict exact values, it is more prone to overfitting and demands high-quality data. Classification, on the other hand, provides a practical, low overfitting-risk, lenient-data-quality-requirement way of visualizing and interacting with data compared to regression.

Part III - Model Selection and Differences

For this question, we chose to compare classification models. We will discuss four models: Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVMs).

1. **Logistic Regression**
   Logistic Regression is a classification extension to linear regression. It represents the relationship between independent variables and the target variable by taking the linear regression result and using the sigmoid function to predict probabilities of classes (or categories). The model is most effective when the decision boundaries are linear; thus, it assumes a linear relationship between the features and the log-odds of those features. Logistic regression is rather simple, can be easily implemented, computes quickly, and is interpretable.

   Logistic regression struggles with non-linear relationships and might perform worse on large datasets with complex feature interactions.

2. **Decision Tree**
   Decision trees recursively split data into subtrees or branches based on feature thresholds. This results in a tree where the leaf nodes represent the classification decisions. Decision trees are great if you want an interpretable model, as the decision process is clearly represented in a tree structure. This model can handle all kinds of data, including non-linear relationships and both numerical and categorical data.

   Decision trees have a tendency to overfit on datasets (especially small ones), which makes them less generalizable. To solve this, we can either prune the decision trees to make decisions more clear and general or use ensemble techniques where multiple decision tree models work together, with the hope that the average answer from all the models is correct. This concept is formally defined as a Random Forest model.

3. **Random Forest**
   Random Forest is the improved version of Decision Tree. As mentioned above, this model is not entirely one model but rather an ensemble of multiple decision tree models combined together. In practice, bagging is used to "split up" or divide the dataset into N subsets that are then used to train N instances of Decision Trees. Random Forest is remarkably good out of the box, requiring very little preprocessing, and it typically does not overfit.

   Random Forest models drop in performance if the dataset is too large, and they are also less interpretable. However, this can easily be addressed with tools such as variable importance and partial dependency plots.

4. **Support Vector Machines (SVMs)**
   SVMs focus on finding a hyperplane with the maximum margin to separate classes of features. This model is not used much anymore. They are only effective with datasets where the features are clearly distinguishable. This model can handle both linear and non-linear decision boundaries by using kernel functions to transform the decision equation into a non-linear representation. SVMs handle outliers and noise really well.

   However, they have a high computational cost, and their results are not interpretable.

Summary: Logistic Regression is a classification extension of linear regression. It thrives on linear feature and log-odds relations but struggles with large datasets with nuanced feature boundaries.

Decision Tree models are highly interpretable and flexible, capable of handling both numerical and categorical data as well as non-linear relationships. However, they are prone to overfitting, especially on small datasets, and require pruning or ensemble methods to generalize better.

Random Forest improves on Decision Trees by using an ensemble approach, combining multiple trees trained on random subsets of data to reduce overfitting and improve accuracy. This model is robust, works well out of the box with minimal preprocessing, but can become computationally expensive for large datasets and is less interpretable than individual Decision Trees.

Support Vector Machines (SVMs) focus on finding a hyperplane with the maximum margin to separate classes. They handle both linear and non-linear decision boundaries effectively with kernel functions and are robust to noise and outliers. However, SVMs are computationally expensive and lack interpretability, making them less common today.