Elements of Machine Learning
Exercise Sheet 1
Winter Term 2023/2024

William LaCroix - wila00001@stud.uni-saarland.de - 7038732
Philipp Hawlitschek - phha00002@stud.unisaarland.de - 7043167

Problem 1 T, 2 points. Warmup

Suppose we want to apply an appropriate statistical learning model to a given problem. Briefly explain how the following parameters influence our choice,

- labelled vs. unlabelled input data,
  labelled data is typically used in supervised approaches. It allows us to directly estimate our model's parameters, as well as giving feedback on model performance via loss functions which are able to compare predicted values/labels against a gold standard value/label. Unlabelled data may be more suited for unsupervised approaches like clustering methods where labels/categories may be assigned arbitrarily to fit the problem.
- numerical vs. categorical variables,
  numerical variables can be encoded directly into a regression model, whereas categorical variables need to be assigned a value that can turn the category into a value for (such as 0,1 for binary categories). A model that predicts a numerical value typically is called a regressor, while a model that assigns categorical values is typically known as a classifier.
- interpretability vs. prediction task,
  complex and highly non-linear models may outperform simpler models for some tasks, but this can come at the cost of interpretability of the model. Simpler models make it easier to understand what factors contribute to the prediction, and may have direct real-world interpretations, whereas highly complex models may have strange and unexplainable interactions, though model performance is improved.
- fixed vs. flexible number of model parameters,
  having a flexible number of model parameters allows us to filter out variables that may be highly correlated, or unhelpful in prediction. A fixed parameter model will only ever use all of the specified variables, for better or worse. This may incur runtime/compute costs if the model is unnecessarily complex, and it may not be an accurate predictor if the model is too simplistic.