Elements of Machine Learning
Assignment 2

William LaCroix (7038732)
wila00001@stud.uni-saarland.de

Problem 3 (T, 5 Points). So many classifiers.

We now know four different classifers: K-NN, LDA, QDA, and Logistic Regression (LR).

1. [4pts] Which assumptions do each of the models make w.r.t. the data distribution? Depending on the type of decision boundary, which of the respective methods would you recommend?

Assumptions by classifier w.r.t. the data distribution:

- K-NN: completely non-parametric, no assumptions are made about the shape of the decision boundary or the distribution of the observations.
- LDA: Assumes that the observations within each class come from a normal distribution with a class specific mean and a common variance σ^2, and that the log odds of the posterior probabilities is linear in x.
- QDA: Assumes observations are drawn from a multivariate Gaussian distribution, with a class-specific multivariate mean vector and a class-specific covariance matrix, where the log odds of the posterior probabilities is quadratic in x.
- LR: Assumes that the observations within each class do not come from a normal distribution, but maintain a linear decision boundary, such that the log odds of the posterior probabilities is linear in x.

Recommended classifier w.r.t. the type of decision boundary:

- K-NN: When observations come from a normal distribution with uncorrelated predictors and the decision boundary is highly non-linear (provided that n is very large and p is small, ie a lot of observations relative to the number of predictors, such that n much larger than p)
- LDA: Where observations follow a Gaussian distribution, share a common covariance between classes, and the decision boundary is (roughly) linear, LDA has the potential to be quite accurate.
- QDA: Has the potential to be more accurate in settings where interactions among the predictors are important in discriminating between classes, ie if K classes do not share a common covariance matrix. If the decision boundary is non-linear but n is only modest, or p is not very small, then QDA may also be preferred to KNN.
- LR: Where observations share a common covariance between classes, and the decision boundary is (roughly) linear, but the data do not follow a normal distribution, LR can outperform LDA.

2. [1pt] Although LDA and LR often yield similar results, LR is often preferred. Give two reasons for this.

- LR relies on fewer strong assumptions about the underlying nature of the features and the distribution of the data
- LR is more robust to outliers, favoring data near the decision boundary, while LDA depends on all the data, and is more susceptible to the leverage exerted by observations far from the decision boundary.