HW2

Theory Questions:

1. The most important evaluation metric depends on the original data we have, and specifically the ratio of different classes within it.

For example, if our data contained 900 healthy subjects and 100 with TD1, a model that classifies everyone as healthy will be 90% accurate but will miss the goal intended for the model in the first place.

In this specific case, we have more healthy subjects than TD1 in our data, making performance a more reliable and useful metric of model evaluation.

1. Having more features in the model may increase computation time and include irrelevant parameters in the decision, making the classifier less accurate. The presence of non-informative features can increase the uncertainty of the model, so ideally, we’d like to only include predictive features. Additionally, some predicting variables can be correlated with each other, and therefore one of them will be sufficient and adding the others in unnecessary.

The problem is that in most real-world problems it is not obvious which factors have a predictive value, and it’s very hard to separate them.

Less features could result in a better model with higher certainty, but accuracy is dependent on correct feature selection. The risk in including only 2 out of 5 features is excluding data with important predictive value.

(In the given example, the income level (I), is usually considered to have little if any effect on prevalence of T1 diabetes, given that it is often hereditary. Including this parameter in the model may not increase the prediction accuracy much, but will definitely decrease model certainty. A different problem is including both BMI and level of physical activity. Both features are useful in determining whether someone has T1D or not, but are highly correlated with each other and so including both of them can increase complexity without adding predictive value).

1. A situation where it is difficult to distinguish between classes is usually caused due to extreme or boundary values that make it harder to determine classes using linear of logistic regression. Therefore, in the case of complicated histology images we would prefer SVM methods that can make better differentiation in border samples. Between linear and non-linear SVM models, the decision is based on ease of differentiation between classes in the “original” coordinates. In cases where it is difficult to find a decision line, transformation using non-linear SVM can make a better decision line.
2. Linear regression finds the decision line that minimizes errors from **all** given data points. Linear SVM finds the decision line that minimizes errors for the **extreme border cases**. This means that LR is more reluctant to bias from the number of “easy decisions” in the data, on the expense of borderline cases. On the other hand, Linear SVM is depending on a limited number of samples, which could reduce accuracy.