Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	2	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 1

(Abgabe am 09.05.2022)

Task 1

The characteristics of a statistical learning problem are defined by the variable type of the **response**, the **output** variable Y. If the response is a **categorical variable** it is a classification problem, whereas a continuous response is a **regression** problem using **numerical variables**. It can be formalized by:

$$Y = f(X) + \epsilon \tag{1}$$

Both cases are interested in estimating f with given **input** (X). The goal is to (1) make **predictions** for new samples or (2) to make **inferences** about the data-generating process and the association of different **features**. **Training data** is used for learning the model. The learning methods can be categorized in **parametric** and **non-parametric** methods. Non-parametric models do not restrict the form of f. Parametric models make an assumption about the shape of f, such as a linear model. **Test data** is used to compare the fit of the different models. Importantly, you should never use training data to estimate how well the model fits. The best fit of a model is achieved if variance and bias of our model are minimized. When the model is too flexible (low bias but high variance), this could lead to **overfitting**, the other extreme is an inflexible model with low variance but high bias. This increases the risk of **underfitting**. This problem is known as the **bias-variance tradeoff**. The flexibility of a model depends on the **sample size** and the "true" form of f. There is also a distinction between supervised and unsupervised methods. In **supervised learning**, for each sample there exists a corresponding response. In **unsupervised learning**, there is no corresponding response for a predictor, e.g. in **clustering** and dimensionality reduction. [1]

268 words

Task 2

b) The train set has 80 observations, ranging in values from 1 to 110. It has no dimensions and no column names. The test set has 31 observations, with values ranging from 6 to 111. It also has no dimensions and no column names.
Ozone has 111 entries in 4 columns named “ozone”, “radiation”, “temperature” and “wind”. Correspondingly, its dimensions are [111 4]. In total, there are 111 observations.
All the columns seem to consist of numerical continous values. This is because they are physical measurements without discrete steps. However, for the columns "ozone" and "radiation", this is harder to jugde, as the physical basis for their computation is unknown to us.

c) Below in figure 1 you can see the scatterplot matrix for every feature in the dataset. The Pearson correlation coefficient between ozone and radiation is 0.35, between ozone and temperature 0.7 and between ozone and wind -0.61. The correlation between radiation and temperature is 0.29 and between radiation and wind -0.13. The correlation between temperature and wind is -0.5.
In general, the range of the correlation coefficient is -1 to 1. A correlation coefficient of 0 means that there is no correlation between the two variables. Positive values mean positive correlation, negative values mean negative correlation.
In the data we can see that the highest positive correlation is between ozone and temperature, which can be seen in the scatterplot as a distribution with an upward trend. The most negative correlation occurs between ozone and wind, which can be seen in the scatterplot as a downward trend. The next most negative correlation is between temperature and wind, which also show a general downward trend in the scatterplot. The lowest **absolute** correlation value occurs between radiation and wind, which can be visually confirmed by looking at the scatterplot in which the values seem to display no trend. The correlation coefficients between ozone and radiation as well as radiation and temperature are also quite low.

d) The column “ozone” ranges from 1 to 168, with the mean at 42.1 and a variance of 1107.29. The column “radiation” ranges from 7 to 334, with the mean at 184.8 and a variance of 8398.74. The column “temperature” ranges from 57.0 to 97.0, with the mean at 77.79 and a variance of 90.82. The column “wind” ranges from 2.3 to 20.7, with the mean at 9.94 and a variance of 12.67.

e) The mean squared error (MSE) is computed by subtracting the true values from the predicted values, then squaring them and summing them up and subsequently dividing by the number of samples. This was implemented in the function in the

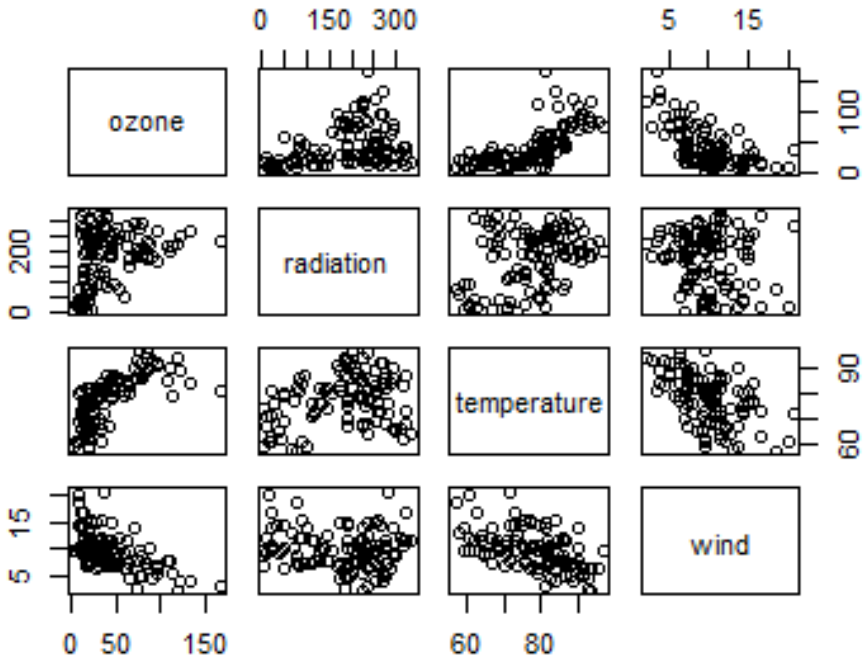


Figure 1: Scatterplot for every pair of variables in the dataset

script.

f) The coefficient values given out for the linear regression model are are: -86.48 for the intercept, 0.03 for radiation, 1.96 for temperature and -3.04 for wind.
The resulting Pearson correlation coefficient is 0.82, so there is a relatively strong positive correlation between prediction and true values.
The MSE for the true values against the predicted values is 327.46. You can see the scatterplot of true ozone values against predicted ozone values in Figure 2 a).

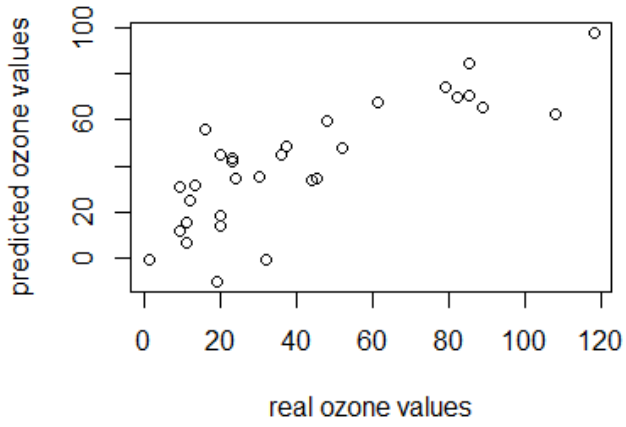
g) Below in Figure 2 b) you see the MSE depending on different values for k. On the left there are the most complex models because the variance is highest for those models. This can allso be seen because in Figure 2 c), the model makes no errors on the train set but the error increases with higher values for k. On the right, the bias would be higher but the variance would decrease. This is known as the bias-variance-tradeoff. We would choose a value of 8 for k because at this value, the MSE is smallest. KNN does not make any assumptions about the underlying data distribution because it just looks for the closest point in each direction, the resulting model can have any form (e.g. linear, sigmoid, ...).

h) The optimal MSE is 329.15, the linear MSE is 327.46, so in principle the two models

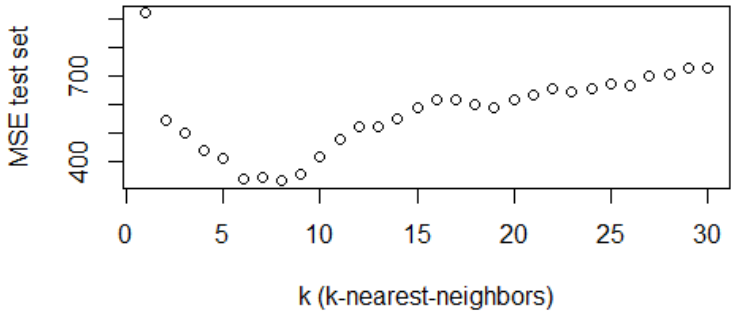
are equivalent for this train set and test set. However, linear models tend to be more robust. The principle of Ockham's razor should be applied: When a simple explanation suffices, you should use the simpler model. This is why the linear model with its lower complexity should be chosen.

References

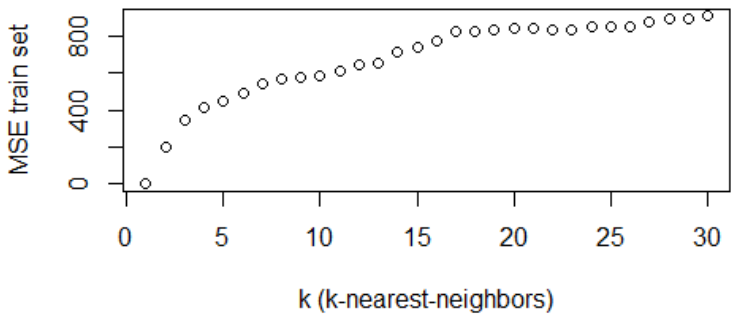
[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics. Springer US, 2021.



(a) Linear model: real vs. predicted ozone values.



(b) Mean Squared Error for the test set depending on different values for k.



(c) MSE for the train set depending on different values for k.

Figure 2: Linear and kNN models for ozone prediction