Compulsory exercise 3: Group XX

TMA4268 Statistical Learning V2018

H??vard Bj??rk??y, Oliver Byhring and J??rgen Riseth

27 April, 2018

Task 1 - Classification with trees

Task 2 - Nonlinear class boundaries and support vector machines

a)

The Bayes classifier is a simple rule of how classify a certain test observation with predictor vector x_0. Given an observation, the rule is to assign the observation to the class of largest probability, i.e. selecting the class j for which $Pr(Y = j|X = x_0)$ is largest. Having made some rule from training data, we use this information to compute the conditional probability. This results in that the Bayes classifier is the method that has least probability of misclassifying. It does not matter how many classes we introduce - the largest one always win.

The Bayesian decision boundary arises where probabilities of observing two classes are equal and largest, so it separates regions where we would classify differently. If $Pr(Y = i|X = x_0) = Pr(Y = j|X = x_0)$ are the largest value of all $i, j, i \neq j$, we have a boundary at x_0. Two equal probabilities that are not the largest will not define this boundary. Having two variables this line will always be 50 % probability of observing either class. The Bayes decision boundary draws a multidimensional map for which classifications the classifier make.

Since the Bayes classifier has the lowest probability of misclassifying, the Bayes error rate (BER) is the lowest test error rate there is. We can use the expected prediction values of our whole set to find this test error rate. In a point x_0 we have a probability $p = max_j\{Pr(Y = j|X = x_0)\}$ of classifying correctly, i.e. the misclassification probability is $1 - p$, so for our entire set X, $BER = 1 - E(max_j(Y = j|X))$.

If we know the true Bayes decision boundary there is no need for a test set, because this represents the optimal classification - having the least test error rate. This error rate is analogous to an irreducible error, we can't overcome it. ???

b)

The difference between a support vector machine (SVM) and a support vector classifier (SVC) is that the last one is a generalization of the first. SVC is a SVM that only allows linear decision boundaries, making it a hyperplane, and is also called a soft margin classifier. SVM can take on non-linear forms.

SVM and the SVC creates a classification boundary, both using a soft-margin technique, where observations are separated by a surface trying to maximize its margin to nearby observations. The soft part comes from introducing som slack for misclassification, allowing the boarder to misclassify according to the slack, which is a tuning parameter.

decided from observations nearby the surface (support vectors), and all misclassifications. Misclassifications are being punished, but the amount of punishment is up to tuning. Therefore, both methods make use of

Q14. What are parameters for the support vector classifier and the support vector machine? How are these chosen above? Q15. How would you evaluate the support vector machine decision boundary compared to the Bayes decision boundary?

Task 3 - Unsupervised methods

a)

The biplot is a way of combining info about all observations values, and visualizing them in two dimensions. As we see from the vectors in the biplot is that the first principal component mostly consists of tea, wine and beer, where wine has negative effect, and the other two positive. This is visible from reading off their values on the first axis, corresponding to their weight in PC1. The second principal component is mostly made up from liquer (negative), coffee and cocoa. The last two weigh positive. Now knowing what the two axis explain we can start interprating the placement of each countries by reading off their values of the two components on the opposing axis'.

From inspecting the summary, in particular the vector components of PC1 and PC2, we see that the observations obviously coincide with the observations from the biplot. The two vectors consist of a weighted combination of each variable, but some of them has greater impact on the value of the principal component.

This analysis does really tell us something about drinking habits for certain countries. For instance, Poland, Soviet Union and Hungary all have negative values of PC2, but seem quite neutral in PC1 - wine, tea and beed. The most significant negative factor is therefore liquer, implying that these three countries have a similarity in high liquer consumption. PC1 being neutral could both mean that the consumption is high, but balanced, or a general low consumption. Other countries with negative PC2 values are also most probable placed there because of large liquer consumption. The factors of PC1 have a more similar weight, making us have to generalize more, but there are definately similarities. The negative values of Italy and Spain is probably because of large wine consumption, and Great Britain and Ireland most likely are similar in their consumption of tea and beer.

b)

The distance between clusters for a single linkage is the smallest distance between two clusters, where the cluster can be either one or several points. Here, distance is not defined because there are several methods to measure it, for instance the intuitive euclidean distance, or correlation-based distance. For complete linkage the distance between two clusters is the distance between two elements in different clusters that are the furthest away from each other. We say it uses maximal intercluster dissimilarity. The average linkage makes use of the centroid of each cluster, and the distance between clusters is the distance between their centroids.

We could not decide which figure corresponds to which measure of distance by the lowest level, though comparing the first cluster (Human and chimpanzee) with their link to gorilla, we could tell them apart. From the table we see that the distance between chimpanzee and human is 1, and their centroid is $\frac{1}{2}(0+1) = 0.5$. The new lowest distance would be the one from the centroid to the gorilla, which is obviously 2.5, hence the figure B is the average linkage. The maximal intercluster dissimilarity if from the lowest cluster to gorilla, i.e. $max\{|0-3|, |1-3|\} = 3$, while the minimal intercluster dissimilarity is 2 - distance between chimpanzee and gorilla. Figure A has placed gorilla on height 3, which indicates that this is the complete linkage tree, while figure C has gorilla on height 2, which corresponds to the single linkage measure.

Task 4 - Neural networks