Tutorial 2: Model Selection and Linear Models

P-Exercise 2.1

In the context of linear regressions, analyze in what situations the bias term b is not needed in:

$$f(x) = \boldsymbol{\omega}^\top \mathbf{x} + b.$$

C-Exercise 2.2

Use the watermelon data set 3.1 to perform the following steps for a Logistic Regression and Linear Discriminant Analysis:

a) Explorative Data Analysis

   i) Detect missing values in the data frame and exclude the corresponding rows.
   ii) Check your data for class imbalance.
   iii) Calculate summary statistics for all applicable columns.

b) Data Preprocessing

   i) Create dummy variables for the respective columns.
   ii) Choose two columns to plot your data set and colour the samples according to their label.
   iii) Remove the mean and scale features to uni variance where applicable. Then create another plot of your scaled data.

c) Fit a Model

   i) Instantiate and fit a model for classification. Use at least two different parameter settings.

d) Model Diagnostics

   i) Calculate the:
      - Accuracy
      - Precision
      - Recall
      - F_1-score
      - Confusion matrix
      - ROC curve

e) Plot the Model

   i) Plot the confusion matrix as a heatmap.
   ii) Plot the model that performed better by displaying which IDs were predicted correctly.

Tutorial 3: Decision Trees

Exercises

P-Exercise 3.1

One of the most commonly used measures for purity is information entropy, or simply entropy. The following data set includes 17 training samples, which are used to train a decision tree classifier for predicting the ripeness of uncut watermelons, where $|\mathcal{Y}| = 2$. In the beginning, the root node includes all samples in D, where p_1 of them are positive and p_2 of them are negative.

ID	color	root	sound	texture	umbilicus	surface	ripe
1	green	curly	muffled	clear	hollow	hard	true
2	dark	curly	dull	clear	hollow	hard	true
3	dark	curly	muffled	clear	hollow	hard	true
4	green	curly	dull	clear	hollow	hard	true
5	light	curly	muffled	clear	hollow	hard	true
6	green	slightly curly	muffled	clear	slightly hollow	soft	true
7	dark	slightly curly	muffled	slightly blurry	slightly hollow	soft	true
8	dark	slightly curly	muffled	clear	slightly hollow	hard	true
9	dark	slightly curly	dull	slightly blurry	slightly hollow	hard	false
10	green	straight	crisp	clear	flat	soft	false
11	light	straight	crisp	blurry	flat	hard	false
12	light	curly	muffled	blurry	flat	soft	false
13	green	slightly curly	muffled	slightly blurry	hollow	hard	false
14	light	slightly curly	dull	slightly blurry	hollow	hard	false
15	dark	slightly curly	muffled	clear	slightly hollow	soft	false
16	light	curly	muffled	blurry	flat	hard	false
17	green	curly	dull	slightly blurry	slightly hollow	hard	false

a) Calculate the entropy of the root node.

Then, we need to calculate the information gain of each feature in the current feature set {*color, root, sound, texture, umbilicus, surface*}. Suppose that we have selected *color*, which has three possible values {*green, dark, light*}. If D is split by *color*, then there are three subsets: D^1 (*color* = green), D^2 (*color* = dark), and D^3 (*color* = light).

b) Calculate the entropy of the three child nodes.

P-Exercise 3.2

a) Calculate the information gain of splitting by *color*.

Similarly, we calculate the information gain of other features:
Gain(D,*root*) = 0,143; Gain(D,*sound*) = 0,141; Gain(D,*texture*) = 0,381;
Gain(D,*umbilicus*) = 0,289; Gain(D,*surface*) = 0,006.

b) Which feature should be chosen as the splitting feature? Justify your answer.

c) Calculate the intrinsic value of *color*.

d) Calculate the gain ratio of *color*.

P-Exercise 3.3

Pruning is the primary strategy of decision tree learning algorithms to deal with overfitting. The general pruning strategies include pre- and post-pruning. Given the watermelon data set in the table above, suppose the samples are randomly partitioned into a training set {1, 2, 3, 6, 7, 10, 14, 15, 16, 17} and a validation set {4, 5, 8, 9, 11, 12, 13}. Suppose we use the information gain criterion for deciding the splitting features, then the following figure shows the decision tree trained on the data set.

a) The pre-pruning process
Pre-pruning decides by comparing the generalization abilities before and after splitting. Prior to splitting, all samples are in the root node.

i) Which samples of the validation set are correctly classified (ripe = true)? Which samples are misclassified (ripe = false)? Calculate the validation accuracy.

ii) After splitting the root node by *umbilicus*, the samples are placed into three child nodes:

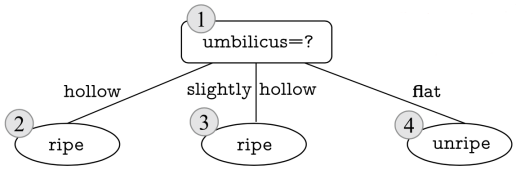


Figure 1: The pre-pruned decision tree

Is the validation accuracy improved? Is the splitting using *umbilicus* adopted?

b) The post-pruning process
Post-pruning re-examines a fully grown decision tree. A node is replaced with a leaf node if the replacement leads to improved generalization ability.

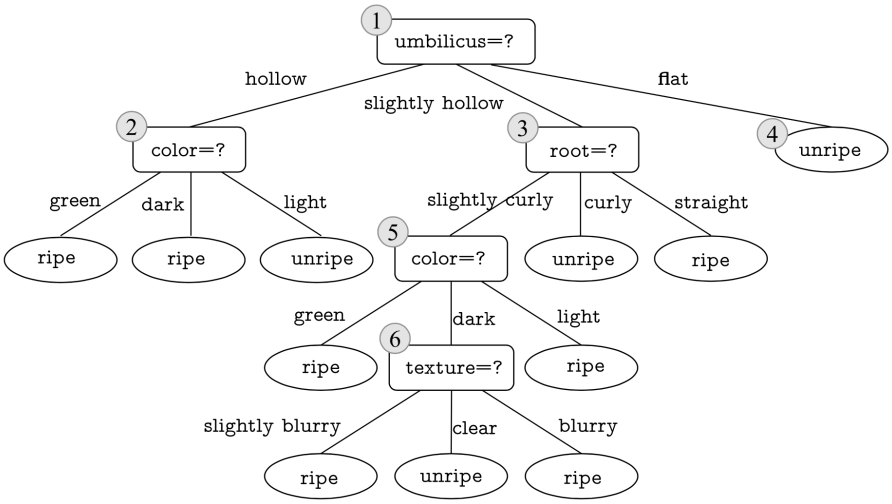


Figure 2: Fully grown decision tree

i) Node ⑥ is the first one examined by post-pruning. If the subtree led by node ⑥ is pruned and replaced with a leaf node, then it includes the samples {7, 15} and its label is set to the majority class *ripe*. The validation accuracy changes from 42,9% to 57,1%. Is the pruning performed?

ii) Which node is examined next in the post-pruning process?

C-Exercise 3.4

Use the bank marketing data set.

a) Explorative Data Analysis

i) Compute the summary statistics for the respective columns.

ii) Count the occurrence of unique values in each column.

iii) Count the education per job in the data frame.

b) Data Preprocessing

i) Create dummy variables for the respective columns.

ii) Split your data into a training and testing set where the testing set is based on 20 % of the data.

iii) Remove the mean and scale features to uni variance where applicable.

c) Fit the Decision Tree

i) Instantiate and fit a decision tree for classification. Use 2 as the maximum depth of the tree and Gini as the measure to evaluate the quality of the split.

ii) Plot the respective tree diagram and generate the tree as a text file.

d) Model Diagnostics

i) Calculate the:

- Accuracy
- Precision
- Recall
- F_1-score
- Confusion matrix

Tutorial 4: Ensemble Learning

P-Exercise 4.1

Explain the differences between hard voting and soft voting classifiers?

P-Exercise 4.2

Describe the advantage of the put-of-bag technique.

C-Exercise 4.3

Use the bank marketing data set.

a) Explorative Data Analysis

   i) Generate a bar plot showing how many clients signed up for a deposit.
   ii) Generate a bar plot, showing how many persons signed up for a deposit within each job.

b) Data Preprocessing

c) Fit the Model

   i) Use the bagging algorithm to classify if a person will subscribe to a new deposit.
   ii) Use the random forest algorithm to classify if a person will subscribe to a new deposit.
   iii) Use the AdaBoost algorithm to classify if a person will subscribe to a new deposit.
   iv) Apply an exhaustive method to evaluate the best parameter set out of a predefined parameter set.
   v) Apply a voting classifier over all previously used model. Use hard voting, soft voting and a weighted average.

d) Model Diagnostics

Tutorial 5: Support Vector Machine

P-Exercise 5.1

Look at the documentation of the sklearn function 'svm.SVC()'. If you would be using a gaussian kernel how is σ defined by default?

C-Exercise 5.2

Use the columns 'density' and 'sugar' from the watermelon 3.0 data set to perform the following steps to build SVM models using the linear, polynomial and Gaussian kernel. Compare their support vectors.

a) Explorative Data Analysis

b) Data Preprocessing

c) Fit the Model

d) Model Diagnostics

e) Plot the Model

C-Exercise 5.3

Use the columns 'balance' and 'duration' from the bank marketing data set to perform the following steps to build SVM models using the linear, polynomial and Gaussian kernel.

a) Explorative Data Analysis

b) Data Preprocessing

c) Fit the Model

d) Model Diagnostics

e) Plot the Model

Tutorial 6: Bayes Classifier

C-Exercise 6.1

Use the watermelon data set 3.0 and classify the T1 sample in Sect. 7.3. Perform the following steps to implement a naïve Bayes Classifier.

a) Explorative Data Analysis

   i) Plot the data set (for the categorical and numerical columns) and add the T1 sample in a different colour.

b) Data Preprocessing

c) Fit the Model

d) Model Diagnostics

e) Plot the Model

C-Exercise 6.2

Use the bank marketing data set to classify if a person will subscribe to a new deposit. Perform the following steps to implement a naïve Bayes classifier.

a) Explorative Data Analysis

b) Data Preprocessing

c) Fit the Model

d) Model Diagnostics

C-Exercise 6.3

Use the bank marketing data set to classify if a person will subscribe to a new deposit. Implement a 10-fold cross validation to compare the accuracy of the following methods:

a) Logistic Regression

b) Linear Discriminant Analysis

c) Decision Tree

d) Random Forest

e) Support Vector Machine (with a gaussian kernel)

f) Naïve Bayes classifier

Tutorial 7: Clustering

P-Exercise 7.1

Below you can see a data snippet from the bank marketing data set with only the numerical attributes.

	age	balance	day	duration	campaign	pdays	previous
x1	33	390	9	665	2	-1	0
x2	32	311	12	757	2	-1	0
x3	35	414	13	504	4	-1	0

a) Calculate the Minkowski distance, for all combinations of samples, for $p = 1, 2, 4$.

b) Calculate the weighted Minkowski distance, for all combinations of samples, for $p = 1$, $w_{\text{age}} = 0.05$, $w_{\text{day}} = 0.90$ and all other weights $w_i = 0.01$.

C-Exercise 7.2

Use the watermelon data set 4.0 in Sect. 7.4.1 to perform the known steps to:

a) Implement and run the k-means algorithm.

   i) Use three different initial centroids. Discuss what kind of initial centroids can lead to good results.

   ii) Discuss how the KMeans function from the sklearn packages chooses the centroids by default.

   ii) Use three different k values. Discuss your results.

b) Implement and run the Learning Vector Quantization algorithm. (Suppose, c_2 is the class label of the samples with ID 9-21, and c_1 is the class label for the rest of the samples.)

   i) Use a setting with and prototype per class and two prototypes per class.

   ii) Implement a corresponding Voronoi tessellation for the model with two prototypes per class and plot your results.

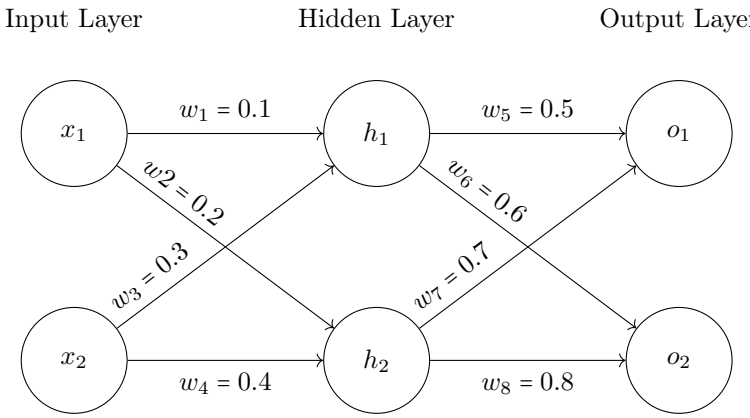
   iii) Compare it to your results for the k-means algorithm.

c) Implement and run the DBSCAN algorithm.

   i) Use three different ϵ values and three different numbers of $MinPts$. Discuss what type of neighborhood-parameters can lead to good results.

Tutorial 8: Neural Networks

P-Exercise 8.1



Assume the inputs for x_1 is 0.1 and for x_2 0.5 and the expected output is 0.05 for o_1 and 0.95 for o_2. Further, use 0.25 as initial bias for the hidden layer and 0.35 for the output layer. As activation function, use the Sigmoid function.

a) Calculate the forward pass.

b) Calculate the backpropagation with respect to all weights.

C-Exercise 8.2

Use the bank marketing data set.

a) Explorative Data Analysis

b) Data Preprocessing

c) Fit the Model

   i) Instantiate a neural network classifier with one hidden layer encompassing 5 neurons. Further, use the Sigmoid function as activation function, a constant learning rate, a regularization technique and the stochastic gradient descent method for the optimizer.

   ii) Consider the resulting weights and biases.

   iii) Use Gridsearch for Hyperparameter tuning and increase the number of hidden layers to 3.

d) Model Diagnostics

   i) Visualize the loss decay for each epoch.

   ii) Calculate the evaluation accuracy.