capbtabboxtable[][]

# Machine Learning - Exam

Winter Term 2017/18

16.02.2018

Prof. Dr. Dr. Lars Schmidt-Thieme

Rafael Rego Drumond

Information Systems and Machine Learning Lab (ISMLL)
Universität Hildesheim

Time: 120 Minutes

Name: _____

MatriculationNr: _____

| Exercise | maximal Points | acquired Points |
|---|---|---|
| 1 | 10 | |
| 2 | 10 | |
| 3 | 10 | |
| 4 | 10 | |
| Bonus | 4 | |
| Total | 40+4 | |

Grade: _____

Name: _____
MatriculationNr: _____

---

**Exercise 1: Supervised Learning: Linear Models and Fundamentals (3+5+2 Points)**

a) Explain:

i. [2pts] What are the advantages and disadvantages of using Gradient Descent to minimize a loss function?

ii. [1pt] What is the main difference between Gradient Descent and the Newton algorithm?

Name: _____
MatriculationNr: _____

b) [5pts] For the data below, learn the weight parameters of the logistic regression model with Gradient Ascent for one iteration and check the accuracy of the model. Initialize the parameters with $\beta_0^T = (0.05 \quad 0.05 \quad 0.05)$ and use a step size $\alpha = 0.6$. Do not forget to include the bias term!

| $x_1$ | $x_2$ | $y$ |
|---|---|---|
| 11 | 3 | 1 |
| 14 | 4 | 1 |
| -16 | -5 | 0 |
| -18 | -6 | 0 |

Name: _____
MatriculationNr: _____

c) [2pts] In the lecture it has been shown that for the simple linear regression

$$\hat{\beta}_1 = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N}(x_n - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

minimize the residual sums of squares (RSS).

Show how these equations can be found by computing the partial derivatives of the loss function. Justify that the given solutions are indeed a global minimizer for the loss.

Name: _____
MatriculationNr: _____

## Exercise 2: Decision Trees (2+6+2 Points)

a) [2pts] Explain and give examples of regularization methods for decision trees. What are the negative outcomes for not using any of them?

b) [6pts] Learn a decision tree to predict whether a vehicle is stolen or not, using misclassification rate as split quality criterion for the following dataset.

| *Color* | *YearsOwned* | *Stolen?* |
|---|---|---|
| Red | 2 | yes |
| Black | 2 | yes |
| Blue | 2 | yes |
| Black | 6 | yes |
| Blue | 6 | no |
| Red | 6 | no |
| Red | 8 | no |
| Red | 8 | no |

c) [2pts] Why is it not advisable to use Misclassification Rate as a split quality criterion for decision trees?

Name: _____
MatriculationNr: _____

---

## Exercise 3: Neural Networks (2+5+3 Points)

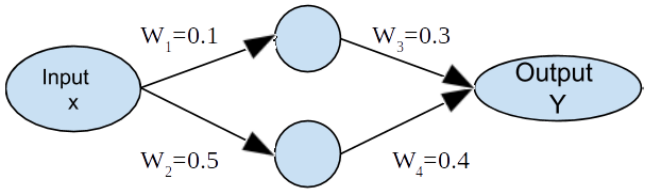
a) Explain:

i. [1pt] How can a feed-forward network model be equivalent to a simple logistic binary linear regression model?

ii. [1pt] What is back-propagation?

Name: _____
MatriculationNr: _____

b) [5pts] Given the model below and considering sigmoid as your activation function, perform a forward pass (input $x = 0.1$), a reverse pass (target $y = 0.1$) and a further forward pass. What is your new error? (For easier calculation, round up your results to 3 decimals). $\mu = 1$



Name: _____

MatriculationNr: _____

---

–

Name: _____
MatriculationNr: _____

c) i. A chef from a bakery has a system to define which food he should recommend based on whether their clients prefer sweet food (1) or salty (0) and if they prefer crunchy (1) or soft food (0). He has four different recipes enumerated from 1 to 4 corresponding to each preference. The Chef built a neural network architecture with one input layer, one hidden layer and one output layer, (where the output is the number of the dish). Consider that the the hidden Layer nodes have a threshold activation function each. Write the value of the weights of the first layer and threshold levels (on the corresponding table). The weights of the final layer are given in the picture below.

Keep in mind that:

$$Y = 1 * H_1 + 2 * H_2 + 3 * H_3 + 4 * H_4$$

And each Hidden layer as Threshold activation function as:

$$H_i = Threshold(W_{i,1} * X_1 + W_{i,2} * X_2, \quad level)$$

$$Threshold(x, level) = \begin{cases} 1 & x = level \quad or \quad x > level \\ 0 & otherwise \end{cases}$$

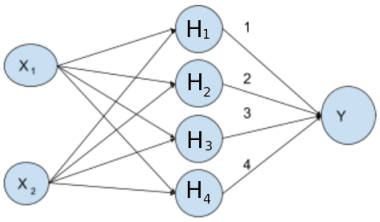
Your weights should minimize the Loss as:

$$argmin \sum_{n}^{4} \begin{cases} 0 & \hat{Y}_n = Y_n \\ 1 & otherwise \end{cases}$$

| Sweet? $X_1$ | Crunchy? $X_2$ | Dish ID $Y$ |
|---|---|---|
| 0 | 0 | 1.Cheese Cream |
| 1 | 0 | 2.Cake |
| 0 | 1 | 3.Spicy Toast |
| 1 | 1 | 4.Peanut Pie |

| Weights for the connection between input and hidden Layer $(W_{h,x})$ | | |
|---|---|---|
| | $X_1$ | $X_2$ |
| $H_1$ | | |
| $H_2$ | | |
| $H_3$ | | |
| $H_4$ | | |

| Node | Threshold Level |
|---|---|
| $H_1$ | |
| $H_2$ | |
| $H_3$ | |
| $H_4$ | |




ii. Explain why we can say that this architecture is similar to the **1 vs. all** approaches . What changes in the network would be necessary to make it similar to a **1 vs. Last** approach?

Name: _____

MatriculationNr: _____

---

-

Name: _____
MatriculationNr: _____

## Exercise 4: Unsupervised Learning (3+4+3 Points)

a) [3pts] Briefly explain (short answers):

i. Give an example of an unsupervised problem.

ii. What is the difference between hard and soft clustering?

iii. Mixture models are said to optimize the expected complete data *log-likelihood*. What does the term "complete" mean in this context?

Name: _____
MatriculationNr: _____

b) [4pts] Use the k-means algorithm and Euclidean distance to cluster the following 8 samples into 3 clusters. The samples' coordinates and the distance matrix based on the Euclidean distance are given below:

| $ID$ | $X_1$ | $X_2$ |
|---|---|---|
| A1 | 2 | 10 |
| A2 | 2 | 5 |
| A3 | 8 | 4 |
| A4 | 5 | 8 |
| A5 | 7 | 5 |
| A6 | 6 | 4 |
| A7 | 1 | 2 |
| A8 | 4 | 9 |

| | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 |
|---|---|---|---|---|---|---|---|---|
| A1 | 0 | $\sqrt{25}$ | $\sqrt{36}$ | $\sqrt{13}$ | $\sqrt{50}$ | $\sqrt{52}$ | $\sqrt{65}$ | $\sqrt{5}$ |
| A2 | | 0 | $\sqrt{37}$ | $\sqrt{18}$ | $\sqrt{25}$ | $\sqrt{17}$ | $\sqrt{10}$ | $\sqrt{20}$ |
| A3 | | | 0 | $\sqrt{25}$ | $\sqrt{2}$ | $\sqrt{2}$ | $\sqrt{53}$ | $\sqrt{41}$ |
| A4 | | | | 0 | $\sqrt{13}$ | $\sqrt{17}$ | $\sqrt{52}$ | $\sqrt{2}$ |
| A5 | | | | | 0 | $\sqrt{2}$ | $\sqrt{45}$ | $\sqrt{25}$ |
| A6 | | | | | | 0 | $\sqrt{29}$ | $\sqrt{29}$ |
| A7 | | | | | | | 0 | $\sqrt{58}$ |
| A8 | | | | | | | | 0 |

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch:

i. Make a sketch of the clusters (i.e. the samples belonging to each cluster).

ii. Compute the centers of the clusters.

iii. Make a new sketch of the clusters which are obtained by running a new epoch. Show the examples which are differently clustered and compute the new centers.

Name: _____

MatriculationNr: _____

---

–

c) [3pts] On K-means clustering, a higher K would not necessarily lead to clusters which represent refinements over the original model, since completely new arrangements could be obtained. What would be the solution if we want to have consistent refined models by varying the number of clusters?