Machine Learning - Exam

Winter Term 2017/18

9.04.2018

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 (2+5+3 Points)

a) Explain:

i. [1pts] Why is it important to adjust the learning rate parameter and the weight initialization when performing algorithms such as gradient descent?

ii. [1pt] What are the main properties of Mean Absolute Error and Mean Squared Error?

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 $\beta_0^T = (0 \quad 0 \quad 0)$ 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

c) [3pts] Jane Doe is trying to figure out a simple way of adjusting the learning rate. She decided to try the following: Initialize the learning rate α_0 with some value between 0.1 and 0.01. Initialize the weights with some random value. After every Z iterations the learning rate is updated as following:

$$\alpha_{t+1} = \alpha_t * 0.1$$

. What are the advantages of this approach compared to using a single learning rate value? What are the problems Jane might find? What precautions or improvements should Jane follow? (Consider Z as a hyper-parameter).

Name: _____
MatriculationNr: _____

Exercise 2: Nearest Neighbor Methods (2+6+2 Points)

a) [2pts] Explain what is K-Nearest Neighbors. What is its main advantage and its disadvantage?

Name: _____
MatriculationNr: _____

b) [6pts] There are two types of dance styles: "Overfunk" and "Underfitango". So far, 4 dances have been classified as such:

ID	Dance Steps Sequence	*Class*
1	L-L-R	Underfitango
2	L-R-L	Underfitango
3	T-S-R	Overfunk
4	T-S-T	Overfunk

Given a new dance with steps (L-T-R) use 3-Nearest-Neighbor method to classify it as "Underfitango" or "Overfunk". (Hint: use string distances as your distance metric).

c) [2pts] What is the difference between Kernel Regression and K-Nearest-Neighbor. In what aspect, is it true to say that Kernel Regression solves one hyper-parameter optimization problem for KNN?

Name: _____
MatriculationNr: _____

Exercise 3: Model Selection and SGD (2+5+3 Points)

a) Explain:

i. [1pt] What is the main difference between gradient descent and coordinate descent?

ii. [1pt] What is backward search?

Name: _____
MatriculationNr: _____

b) [5pts] Consider the data:

x_1	x_2	y
1	2	4
-1	1	7
2	1	1

And the model learned from that data: $\beta = (4 \quad -2 \quad 1)$. Perform a backward selection (consider testing only without x_1 and then only without x_2) and indicate using AIC metric if there was an improvement to the current model in both attempts. Hint: Cramer's rule to find the inverse matrix of a 2x2 matrix:

$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad A^{-1} = (1/(ad - bc)) \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Name: _____

MatriculationNr: _____

–

c) Stochastic Gradient Descent (SGD) works very similar to normal gradient descent, the key difference is that only one data instance is used per update, i.e. the (regression) loss function for a single instance resolves to:

$$\mathcal{L}(y, \hat{y}(x)) = \frac{1}{2}(\hat{y}(x) - y)^2$$

For a polynomial regression, i.e.

$$\hat{y}(x) = \beta_0 + \beta_a x_a + \sum_{l=1}^{p} \sum_{j=l}^{p} \sum_{k=l}^{p} \beta_{ljk} x_l x_j x_k$$

Where a is a constant. Compute the update equations of a stochastic coordinate descent for all parameters β_0, β_i and β_{ljk} for the single instance loss!

Name: _____
MatriculationNr: _____

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

a) [2pts] Briefly explain (short answers):

i. The difference between Confidence and Support in Pattern Mining.

ii. The difference between hard and soft clustering.

Name: _____
MatriculationNr: _____

b) [5pts] Consider the following points:

x_1	x_2	Initial Cluster
-3	3	2
3	-3	2
4	-3	2
4	-4	2
-4	3	1
-4	4	1

Consider the initial centers $\mu_1 = (-4, 3), \mu_2 = (-3, 3)$, and the distance metric between a point X_a and a point X_b: $D(X_a, X_b) = |x_{a1} - x_{b1}| + |x_{a2} - x_{b2}|$.

i. Compute the new centers and assign new clusters to the data-points.

ii. Repeat step i until μ_1 and μ_2 stop changing coordinates.

iii. Make a sketch of the clusters before the first update and after the last one (include centers and points).

c) [3pts] Consider a Gaussian Mixture Model, how do we make it equivalent to the K-means method?