Istanbul Technical University- Fall 2018
BLG527E Machine Learning
Homework 4

Purpose: Getting ready for the final exam.
Total worth: 5% of your grade.
Handed out: Saturday, Dec 29 2018.
Due: Monday, January 7, 2019 23:00. (through ninova!)
Instructor: Zehra Çataltepe (cataltepe@itu.edu.tr),
Assistant: Fulya Çelebi Sarıoğlu (sarioglu16@itu.edu.tr)
Policy: Collaboration in the form of discussions is acceptable, but you should write your own answer/code by yourself. Cheating is highly discouraged for it could mean a zero or negative grade from the homework.

If a question is not clear, please let us know (via email, during office hour or in class).
Submission Instructions: Please submit through the class ninova site.
Please upload all your files using filename studentID_HW4.docx or .pdf.
This homework aims to prepare you for the final exam, so there are 10 questions. However you need to provide answers only for the **5 questions.**

Q1a) Consider a classification problem for a data set $\{\Phi_n, \mathrm{t}_n\}$, n = 1, . . . , N, where $\mathrm{t}_n \in \{0, 1\}$ and $\Phi_n = \Phi\ (x_n)$. Let $\mathrm{t} = [\mathrm{t}_1, .., \mathrm{t}_N]$, $a_n = \mathrm{w}^T \Phi_n$,$y_n = p(C_1|\Phi_n) = \sigma(a_n)$ where $\sigma(.)$ is the logistic sigmoid function: $\sigma(a) = 1/(1 + \exp(-a)\ )$.

Show how you would minimize the error function for the logistic regression classification model:

$$E(\mathbf{w}) = -\ln \mathbf{p}(\mathbf{t}|\mathbf{w}) = -\ln \prod_{n=1}^{N} y_n^{t_n}\ \{1 - y_n\}^{1-t_n}$$

Provide all the details of your work.
Hint: Make use of the fact that: $d\sigma(a)/da = \sigma(a)(1 - \sigma(a))$

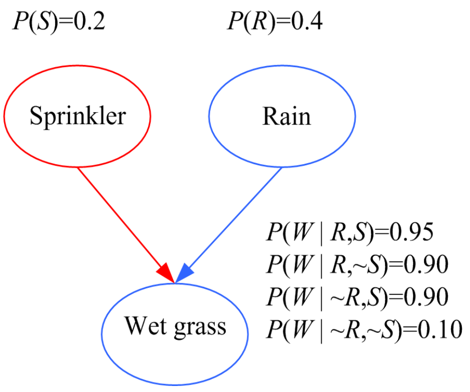
Q2)
Q2a) How do you control an SVM's model complexity?
Q2b) How do you train (i.e. decide on the best weights of) a multilayer perceptron? How can you decide on the best complexity?
Q2c) How do you use Parzen windows for density estimation, classification and regression?
Q2d) What are the differences and similarities between logistic regression and multilayer perceptron classifier?

Define the following using at most three sentences. You can use your own notation to explain the concepts clearer:
Q2e) Support vector
Q2f) Bayes rule
Q2g) mRMR feature selection
Q2h) Area Under the ROC Curve (AUC)

Q3)

Q3a) Consider the Bayesian network given below and compute P(R|W).

Q3b) Consider the Bayesian network given below. Is S independent of R given W? Prove your answer.



Q4)

The following are the actual outputs and outputs produced by 7 classifiers for a classification problem. The outputs on 3 training instances and 2 test instances are given.

Use **bagging** to compute the outputs for the test instances. What outputs do you produce and what is the confusion matrix for the test samples? Show all steps of your solution.

	Actual	$g_1(x)$	$g_2(x)$	$g_3(x)$	$g_4(x)$	$g_5(x)$	$g_6(x)$	$g_7(x)$
Train	0	0	1	0	0	0	1	1
	1	0	1	1	0	1	0	0
	0	1	0	0	1	1	0	0
Test	1	1	1	1	0	0	1	0
	0	0	1	0	1	0	0	0

Q5) Given an HMM λ = (п,A,B) with state transition probability matrix A, emission probabilities B, initial state probabilities п, and two states and two symbols red and green,

п = $[0.3 \ 0.7]^T$

$$A = \begin{vmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{vmatrix}$$

$$B = \begin{array}{|cc|l} \text{red} & \text{green} & \\ 0.7 & 0.2 & \text{State1} \\ 0.1 & 0.1 & \text{State2} \end{array}$$

What is the Pr(O| λ) where O = {red, red, green}

Q6) Produce a univariate decision tree using using the entropy as the impurity criterion for the dataset below. The maximum height for your decision tree is 3.

x_1^t	x_2^t	x_3^t	r(t)	predicted(t)=?
0	1	0	1	
0	0	1	1	
1	1	0	1	
0	1	0	1	
0	0	1	0	
0	1	1	0	
1	0	0	1	
1	0	0	0	

Q7)

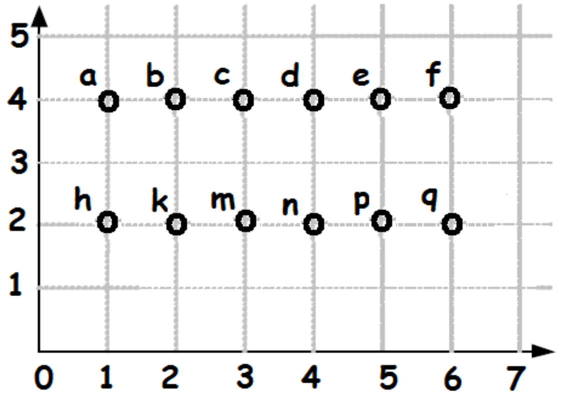
Show the resulting clusters when

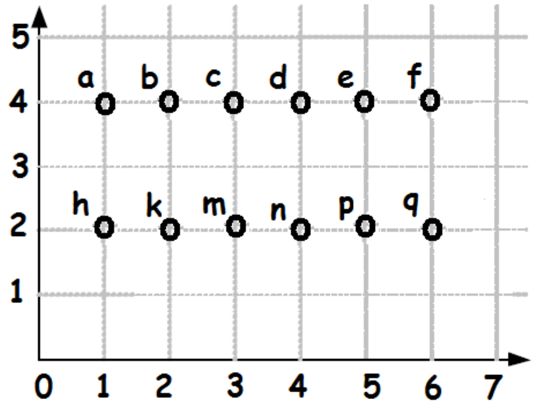
a) (left graph) Kmeans clustering is used with K=2 and **a** and **f** as the initial points.

b) (right graph) Agglomerative clustering with single linkage is used (i.e. the distance between two clusters is the distance between the two closest data points.

c) Which clustering is more suitable for this dataset?

Use the l_1 norm as the distance measure, i.e. dist(u,v) = $|u_1-v_1|+|u_2-v_2|$





Q8) For Classifiers 1 and 2, the 5 fold cross validation errors are given below.

Q8a) Use k-fold cross validation paired t-test to show if Classifier1 and Classifier2 have different performance.

Q8b) Use ANOVA to show if Classifier1 and Classifier2 have different performance.

CV-Fold	$e_1(x)$	$e_2(x)$
1	0.3	0.2
2	0.2	0.2
3	0.3	0.1
4	0.2	0.1
5	0.1	0.3

Q9)
Compute the change in v and w (i.e. Δv and Δw) if you modified the error function for a neural network as follows.

$$E(\mathbf{W},\mathbf{v}|\mathcal{X}) = \frac{1}{2}\sum_t \left(r^t - y^t\right)^2 + \| w \|^2$$

Hint: Without the regularization term, you would have:

$$y_i = \mathbf{v}_i^T \mathbf{z} = \sum_{h=1}^{H} v_{ih} z_h + v_{i0} \qquad z_h = \text{sigmoid}\left(\mathbf{w}_h^T \mathbf{x}\right)$$

$$\Delta v_h = \sum_t \left(r^t - y^t\right) z_h^t$$

$$\begin{aligned}\Delta w_{hj} &= -\eta \frac{\partial E}{\partial w_{hj}} = -\eta \sum_t \frac{\partial E}{\partial y^t} \frac{\partial y^t}{\partial z_h^t} \frac{\partial z_h^t}{\partial w_{hj}} \\ &= -\eta \sum_t -\left(r^t - y^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t \\ &= \eta \sum_t \left(r^t - y^t\right) v_h z_h^t \left(1 - z_h^t\right) x_j^t\end{aligned}$$

Q10) How do you use Parzen windows for density estimation, classification and regression?

Let the Gaussian Kernel function be : $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right]$

Density Estimation:

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)$$

Classification:

$$\hat{p}(\boldsymbol{x} \mid C_i) = \frac{1}{N_i h^d} \sum_{t=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}^t}{h}\right) r_i^t \qquad \hat{P}(C_i) = \frac{N_i}{N}$$

$$g_i(\boldsymbol{x}) = \hat{p}(\boldsymbol{x} \mid C_i) \hat{P}(C_i) = \frac{1}{N h^d} \sum_{t=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}^t}{h}\right) r_i^t$$

Regression:

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)}$$