Istanbul Technical University- Fall 2017

BLG527E Machine Learning

Homework 5

Purpose: Getting ready for the final exam.

Total worth: 6% of your grade.

Handed out: Tuesday, May 24, 2017.

Due: Wednesday, June 7, 2017 22:00. (through ninova!)

Instructor: Zehra Cataltepe (cataltepe@itu.edu.tr),

Assistant: Mahiye Uluyağmur- Öztürk (muluyagmur@itu.edu.tr), Hakan Gündüz (hakangunduz@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_HW5.docx or .pdf.

This homework aims to prepare you for the final exam, so there are 17 questions. However you need to provide answers only for the **first 4 questions.**

1. (Ch13)Given 10 training data points (X) and the predefined kernel between them (K) as follows and labels r_i=1 if i<=5 and r_i=-1 otherwise. Train a SVM classifier with given kernel and find all alpha values belonging to classifier. Use these alphas to classify the datapoint x=[0.3, 0.2]

X =

1.6715	2.0347
-0.2075	1.7269
1.7172	0.6966
2.6302	1.2939
1.4889	0.2127
-0.1116	0.4384
-2.1471	-0.6748
-2.0689	-1.7549
-1.8095	0.3703
-3.9443	-2.7115

K =

1.00	0.22	0.36	0.41	0.23	0.15	0.04	0.03	0.06	0.02
0.22	1.00	0.17	0.11	0.16	0.37	0.10	0.06	0.19	0.03
0.36	0.17	1.00	0.46	0.78	0.23	0.06	0.05	0.07	0.02
0.41	0.11	0.46	1.00	0.29	0.11	0.04	0.03	0.05	0.02
0.23	0.16	0.78	0.29	1.00	0.28	0.07	0.06	0.08	0.03
0.15	0.37	0.23	0.11	0.28	1.00	0.16	0.10	0.26	0.04
0.04	0.10	0.06	0.04	0.07	0.16	1.00	0.46	0.45	0.12
0.03	0.06	0.05	0.03	0.06	0.10	0.46	1.00	0.18	0.18
0.06	0.19	0.07	0.05	0.08	0.26	0.45	0.18	1.00	0.07
0.02	0.03	0.02	0.02	0.03	0.04	0.12	0.18	0.07	1.00

2. (Ch15) You are given the following HMM with N=2 hidden states: S1, S2, M=2 possible observations: a,b, and state transition probabilities (A) and observation probabilities (B) and initial state probabilities (P).

a) Compute the probability that the observation sequence O = a,a,b was produced by this HMM.

b) What is the most probable state sequence given O?

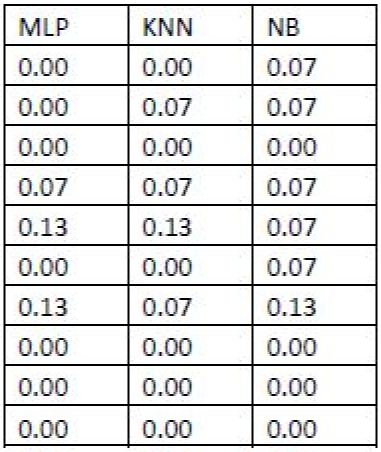
$$A = \begin{bmatrix} 0.8, & 0.2 \\ 0.2, & 0.8 \end{bmatrix} \quad B = \begin{array}{c} \begin{matrix} a & b \end{matrix} \\ \begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix} \end{array} \quad P = [0.9, 0.1]$$

3. (Ch16) Assume that you have a chain of random variables X1 -> X2 -> X3-> X4, with P(X1) = 0.2 and P(Xi|Xi-1) = 0.8. Compute P(X3|X1,~X4) using belief propagation.

4. (Ch19) Assume that you have used MLP, KNN and (NB) Naive Bayes as classification methods and 10 fold cross validation. The errors for each fold and classifiers are given below.

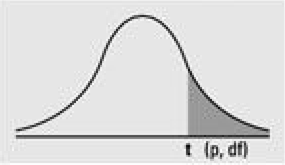
a) With significance level alpha=0.05, would you accept that error of MLP is equal to error of KNN?

b) Use ANOVA to show whether error_MLP = error_KNN = error_NB or not at alpha=0.05 significance level.



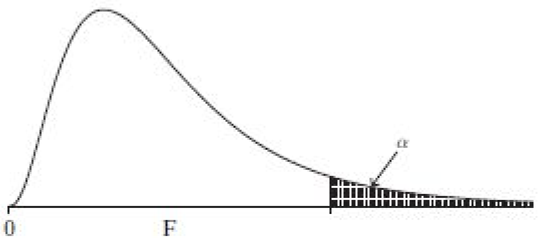
MLP	KNN	NB
0.00	0.00	0.07
0.00	0.07	0.07
0.00	0.00	0.00
0.07	0.07	0.07
0.13	0.13	0.07
0.00	0.00	0.07
0.13	0.07	0.13
0.00	0.00	0.00
0.00	0.00	0.00
0.00	0.00	0.00

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0005**
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

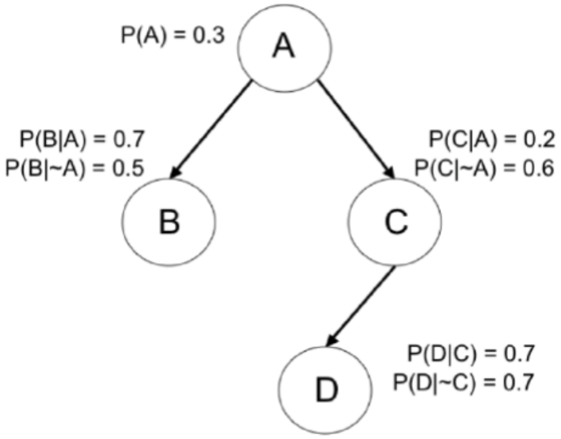
TABLE D: *F* Distribution



	α = .05									
	df$_1$									
df$_2$	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

Source: From Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, published by Longman Group Ltd., London, 1974. (Previously published by Oliver & Boyd, Edinburgh.) Reprinted by permission of the authors and publishers.

5. (Ch16) For the Bayesian network shown below, compute the following:



a) P(~A,B,C,D) =?
b) P(A|~B) =?
c) P(C|B) =?

6. (Ch19) Schizophrenia is a mental disorder where the patients confuse what is real and what is their imagination. It highly reduces the quality of life for both the patients and for people close to them. It is known that 1 in every 100 individual has schizophrenia. Recently scientists came up with a genetic signature that only 7 percent of schizophrenic people don’t have it. And only 9 percent of healthy individuals have the signature. There is a gene therapy when the fetus is not more than 5 months old but it has risks of course. Even when the fetus will really have schizophrenia the cost of the procedure is decided to be 10 units. But if the fetus will not be schizophrenic the cost of the therapy is 60 units. If we do not apply the therapy and the fetus will have schizophrenia the cost is assumed to be 100 units. Suppose you are faced with a fetus that shows the genetic signature for schizophrenia, would you apply the therapy? Justify your decision.

7. (Ch14) Given that p(x) ~ N(mu, sigma2), and prior distribution for mu as p(mu) ~ N(0,1), and N observations X = {x1,...,xN}

derive the posterior distribution for mu: p(mu|X)

8. (Ch4) Assume you observed one coin tossed 11 times and the observations at time ***t=1 to t=11*** are as follows: **{H,T,H,H,H,T,T,T,H,T,T}**. Is the outcome at toss ***t+1*** independent of the outcome at toss ***t***? Why or why not?

9. (Ch6) a) Name two differences between PCA (Principal Component Analysis) and Backward Feature Selection.

10. (Ch7) Consider the four unlabeled data points **{x1 = (1,1), x2 = (2,1), x3 = (2.5,0.5), x4 = (2,0)}**

a) You need to divide them into 2 clusters. What would be the most reasonable clustering? Give the coordinates of the center of each cluster. (Use **city block distance** as the distance measure: **|a,b|=|a1-b1|+|a2-b2|**)

b)Considering all possible initial cluster centers (assume that 2 different cluster centers are chosen randomly among the four data points), what are the clusterings produced by the k-means clustering algorithm with k=2.

11. (Ch8) Given a classification problem with inputs X = {x1,...,xN} with outputs r={r1,...,rN} write down the Parzen window estimator for the output for a given input x.

12. (Ch9) Given a classification problem and a dataset with continuous outputs, you want to train a decision tree. For each node, instead of splitting on one variable at a time, you want to split using multivariate logistic regression classifiers. Explain how would you train such a decision tree.

13. (Ch10) Given a function of the form g(w1,w2,x) = w1*x*x + w1*w2*x + w2*x + w1 and squared error, derive the partial derivative of the error function with respect to both parameters w1 and w2 and explain how you would use gradient descent to determine the best values of w1 and w2 for a particular dataset with inputs X = {x1,...,xN} with outputs r={r1,...,rN}

14. (Ch13) Compare the error functions used by SVM and MLP for classification (regularized error vs error) and regression (hinge loss vs squared error). Explain why these error functions could help SVM generalize better.

15. (Ch11)
a) Explain how does each of the following help for the training of a MLP: adaptive learning rate, momentum, L2 regularization.
b) Assume that you have a regression problem with one dimensional inputs and you know the following hint about your problem: if (x<5) the output should increase with x and if x>=5 the output should decrease with x. You are given a training data set X of N labeled instances also. Explain how could you train a MLP taking into account both the hint and the dataset X.

c) Write down and compare the error functions and MLP architectures that you would use to train a MLP for a regression problem and for a classification problem.
d) How do you initialize a MLP and why?
e) Assume that you have received a training dataset X1 and trained an MLP using that dataset. Let the resulting MLP be MLP1. A while later you received another dataset X2. You do not have access to dataset X1 anymore but you have MLP1. Explain how could you train an MLP so that you could learn X1 and X2 at the same time.

16. (Ch13) Compare the error functions used by SVM and MLP for classification (regularized error vs error) and regression (hinge loss vs squared error). Explain why these error functions could help SVM generalize better.

17. (Ch18) a) Give two differences between Bagging and Adaboost. b) Assume that you combined K classifiers using stacking with a MLP with as many layers as you need. Prove that the Adaboost combiner can be implemented with MLP-stacking. c) Under what conditions (of, for example, problem difficulty, number of instances, number of features, number of base classifiers) could you prefer Adaboost over MLP-stacking?

18. (Ch4)
1a) Suppose you are given a financial regression dataset generated from **a polynomial of degree of 4**. Indicate whether you think the bias and variance of the following models would be relatively high (H) or low (L) considering the true model. Polynomial of degree 1,
Polynomial of degree 4,
Polynomial of degree 10
Polynomial of degree 10 trained with regularization
1b) Given a multivariate binary classification problem and assumption of normally distributed d dimensional inputs, what are the most complex and least complex classifiers that you could produce? Explain in detail your assumptions to arrive at those classifiers and the number of parameters needed to be estimated from training data.