Machine Learning- Fall 2019

Final Exam

Department of Bioinformatics, IBB, University of Tehran
January 27, 2020

Part 1: problems 1-5; pages 1-8 **Time limit:** 8:30-10:00

Part 2: problems 6-9; pages 9-14 **Time limit:** 10:15-12:15

Total point: 100

A one-sided cheat sheet is allowed. In addition, you can access to the slides and lecture notes in the second part of the exam (problems 6-9). Good luck!

Your name: --------------------------- (please *initial all pages!)*

Disclaimer: *some questions are taken or modified from online resources.*

Please leave the below table empty.

PROBLEM	MAXIMUM POINTS	OBTAINED POINTS
1	14	
2	12	
3	8	
4	8	
5	8	
6	5	
7	10	
8	20	
9	15	
TOTAL	100	

Part 1: problems 1-5

1. Multiple-choice Questions (14 points)

Circle the correct choices. **Note multiple choices or none might be correct.**

i. When the sample size n is extremely large and the number of predictors p is small, a flexible learning method is preferred

a) false b) true

ii. In KNN, K=1 will always give the minimum training error.

a) false b) true

iii. In KNN, K=n will always give the minimum generalization error (n: sample size).

a) false b) true

iv. In statistical learning, the bias always has a bigger contribution to error than the variance.

a) false b) true

v. In random forests, very large number of trees will not result in overfitting.

a) false b) true

vi. In boosting, very large number of trees will often result in overfitting.

a) false b) true

vii. What is the VC dimension for Linear Support Vector Machines in d-dimensional space?

a) 1 b) d c) d+1 d) min(d, n)

viii. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. As we increase s from 0, the bias will:

a) Increase initially, and then eventually start decreasing in an inverted U shape.
b) Decrease initially, and then eventually start increasing in a U shape.
c) Steadily increase.
d) Steadily decrease.
e) Remain constant.

ix. Repeat the previous question for the test error.

a) Increase initially, and then eventually start decreasing in an inverted U shape.
b) Decrease initially, and then eventually start increasing in a U shape.
c) Steadily increase.
d) Steadily decrease.
e) Remain constant.

x. Which of the following models are generative (multiple choices or none can be correct)

a) Logistic regression b) Naïve Bayes c) LDA d) SVM

xi. Bayes error for complex models often is lower than inflexible models.

a) false b) true

xii. Circle the correct choice(s) about “Type-1” and “Type-2” errors?

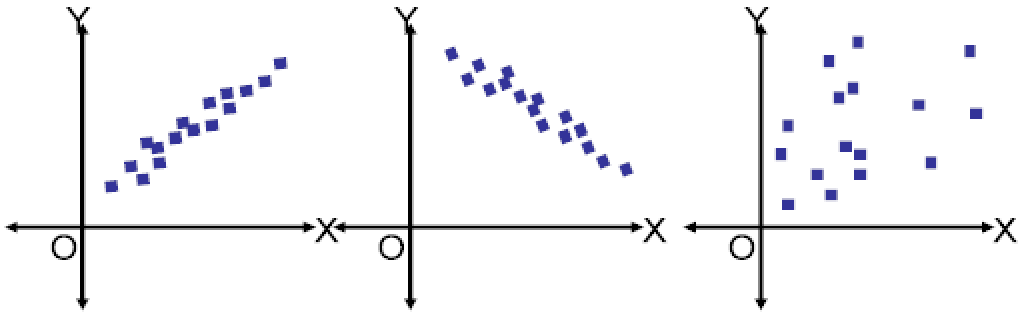
a) Type1 is known as false positive and Type2 is known as false negative.

b) Type1 is known as false negative and Type2 is known as false positive.

c) Type1 error occurs when we reject a null hypothesis when it is actually true.

xiii. Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right), which choices contain multi-collinear features? Features in

a) Left image b) Middle image c) Right image



xiv. In a binary classification problem, circle all choices that are correct according to the below confusion matrix:

a)

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

a) Accuracy is ~0.91

b) Misclassification rate is ~ 0.91

c) False positive rate is ~0.95

d) True positive rate is ~0.95

2. Short Questions (12 points)

a) What is the difference between classification and regression models?

b) Define Bias and Variance of an estimator. What is the bias-variance decomposition?

c) What is the difference between stochastic and batch gradient descent algorithms?

d) Explain random forest algorithm. Why is it a variance reduction algorithm?

e) Explain how maximum likelihood estimation is different for discriminative and generative models (Hint: you need to explain in terms of likelihood function)

f) Explain the difference between parametric and non-parametric models. Give an example for each.

3. Support vector machines (8 points)

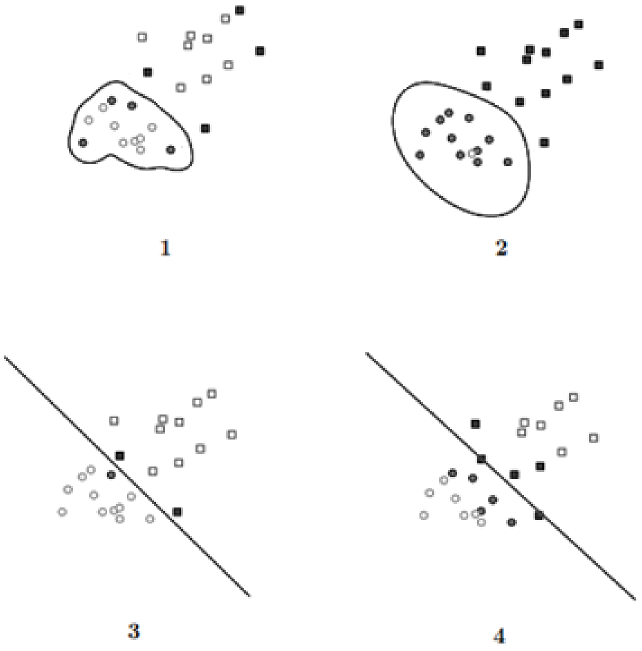
In the below Figure, there are different SVMs with different shapes/patterns of decision boundaries. The training data is labeled as $y[i] \in \{-1, 1\}$, represented as the shape of circles and squares respectively. Support vectors are shaded. Match the scenarios described below to one of the four plots. Each scenario should be matched to a unique plot. Explain in two sentences why it is the case for each scenario.

a) A soft-margin linear SVM with $C = 0.02$.

b) A soft-margin linear SVM with $C = 20$.

c) A hard-margin kernel SVM with $K(u, v) = \exp(-5\|u - v\|^2)$

d) A hard-margin kernel SVM with $K(u, v) = \exp(-\frac{1}{5}\|u - v\|^2)$



4. Model selection (8 points)

For each of the following dataset, explain which learning model or models do you use? (N: number of observations; p: number of feature)

a) Prediction task: $N = 20, \;\; p = 10^6$

b) Knowledge discovery task: $N = 20, \;\; p = 10^6$

c) Prediction task: $N = 10^9, \;\; p = 10^3$

d) Knowledge discovery task: $N = 10^9, \;\; p = 10^3$

5. Stochastic gradient ascent rule for Poisson regression (8 points)

Poisson regression is a generalized linear model where the response variable is assumed to have a Poisson distribution.

Recall a discrete random variable Y is said to have a Poisson distribution with parameter λ > 0, if, for y = 0, 1, 2, ..., the probability mass function of Y is given by $\Pr(\mathrm{Y} = \mathrm{y}) = \frac{\lambda^y e^{-\lambda}}{y!}$. In addition, a distribution is from an exponential family if it can be expressed in the form $P_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp(t(\theta)^T S(y))$ where θ is the parameter set of P and Z, h, t, and S functions are some known functions. In this problem you need to

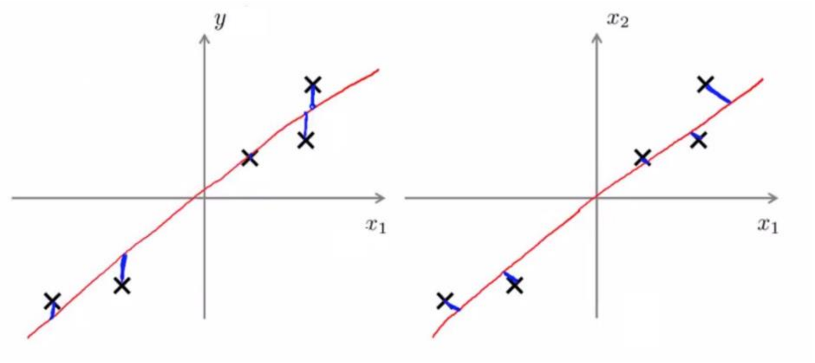
a) Write the Poisson distribution in an exponential family form.

b) Derive the stochastic gradient ascent rule for Poisson regression.

Part 2: problems 6-9

6. PCA and linear regression (5 points)

The first principle component of the PCA model and the linear regression model are drawn in the below figures. Explain which is which and why.



7. Clustering (10 points)

The pseudo code of *BSAS (Basic Sequential Algorithmic Scheme)* and *MBSAS (Modified BSAS)* clustering algorithms are presented here:

BSAS

```
• $m = 1$
• $C_m = \{x_1\}$
• For $i = 2$ to $N$
    —Find $C_k$: $d(x_i, C_k) = \min_{1 \le j \le m} d(x_i, C_j)$
    —If $(d(x_i, C_k) > \Theta)$ AND $(m < q)$ then
        * $m = m + 1$
        * $C_m = \{x_i\}$
    —Else
        * $C_k = C_k \cup \{x_i\}$
        * update representative.
    —End {if}
• End {For}
```

MBSAS

```
    Cluster Determination
• $m = 1$
• $C_m = \{x_1\}$
• For $i = 2$ to $N$
    —Find $C_k$: $d(x_i, C_k) = \min_{1 \le j \le m} d(x_i, C_j)$
    —If $(d(x_i, C_k) > \Theta)$ AND $(m < q)$ then
        * $m = m + 1$
        * $C_m = \{x_i\}$
    —End {if}
• End {For}
Pattern Classification
• For $i = 1$ to $N$
    —If $x_i$ has not been assigned to a cluster, then
        * Find $C_k$: $d(x_i, C_k) = \min_{1 \le j \le m} d(x_i, C_j)$
        * $C_k = C_k \cup \{x_i\}$
        * update representative.
    —End {if}
• End {For}
```

The parameters of both algorithms: Θ: Threshold of dissimilarity, N: Number of samples, q: The maximum allowable number of clusters , m: Current number of clusters

$d(x, C)$= $d(x, m_C)$: d is the distance between sample x and cluster C where m_C is the representative of cluster C. Consider mean as representative. Update m_{C_k} in each step as follows (n is the cardinality of C_k after the assignment of x to it): $\boldsymbol{m}_{C_k}^{new} = \frac{(n-1)\boldsymbol{m}_{C_k}^{old} + \boldsymbol{x}}{n}$

Now, consider the following two-dimensional vectors:

$x_1 = [1,1]^T, x_2 = [1,2]^T, x_3 = [2,2]^T, x_4 = [2,3]^T, x_5 = [3,3]^T, x_6 = [3,4]^T, x_7 = [4,4]^T,$

$x_8 = [4,5]^T, x_9 = [5,5]^T, x_{10} = [5,6]^T, x_{11} = [-4,5]^T, x_{12} = [-3,5]^T, x_{13} = [-4,4]^T, x_{14} = [-3,4]^T$

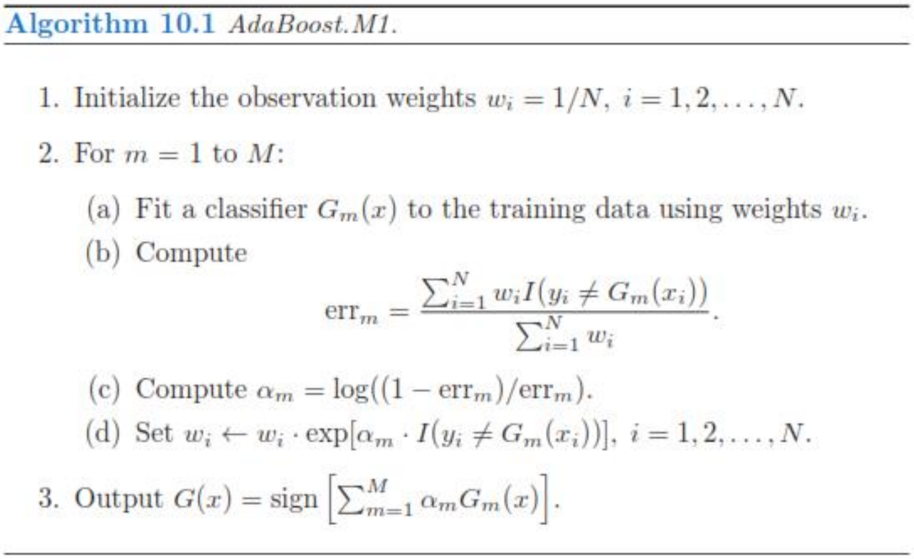
a) Run ***MBSAS*** when the vectors are presented in the given order. Use the Euclidean distance between two vectors and take $\Theta = \sqrt{2}$. Write the final results below and show your work. (**Hint:** first plot the points and then follow the steps of the algorithm).

*MBSAS: C*1 = *{ }, C*2 =*{ }, C*3 = *{ }, C*4 = *{ }* and *C*5 = *{ }*

b) Discuss sensitivity of the *BSAS* and *MBSAS* to the ordering of presentation of the vectors to the algorithm.

8. Boosting (20 points)

In this question, you need to investigate how Adaboost works for the XOR problem. The Adaboost algorithm is as follows

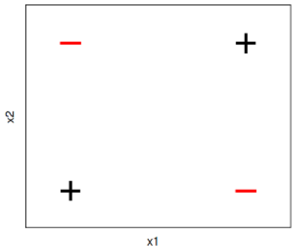


Algorithm 10.1 *AdaBoost.M1.*

```
1. Initialize the observation weights $w_i = 1/N$, $i = 1, 2, \ldots, N$.
2. For $m = 1$ to $M$:
    (a) Fit a classifier $G_m(x)$ to the training data using weights $w_i$.
    (b) Compute
        $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
    (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
    (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))]$, $i = 1, 2, \ldots, N$.
3. Output $G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$.
```

__For the parts *a* and *b*, assume the weak learner is a decision stump (decision tree of with maximum two leaves).__

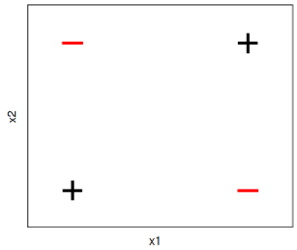
a) Using the below dataset, draw the decision boundary learned by G_1 in the first iteration. What is α_1 ? __*(mark regions as positive/negative assuming that ties are broken arbitrarily.)*__



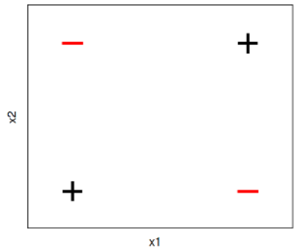
b) What is the minimum achievable training error by Adaboost for this example? Why do you think so?

For the remaining parts, assume the weak learner is a decision tree of with maximum three leaves.

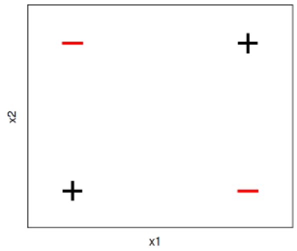
c) Draw the decision boundary learned by G_1. Compute α_1 and weights. Circle misclassified points by G_1.



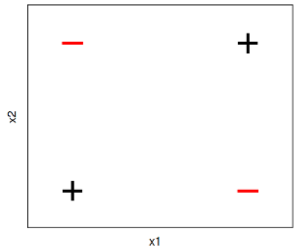
d) Draw the decision boundary learned by G_2. Compute α_2 and weights. Circle misclassified points by G_2.



e) Draw the decision boundary learned by G_3. Compute α_3 and weights. Circle misclassified points by G_3.



f) Indicate how $G = sgn(\alpha_1 G_1 + \alpha_2 G_2 + \alpha_3 G_3)$ classify each of the four points?



h) What is the minimum training error achievable by boosting for this example?

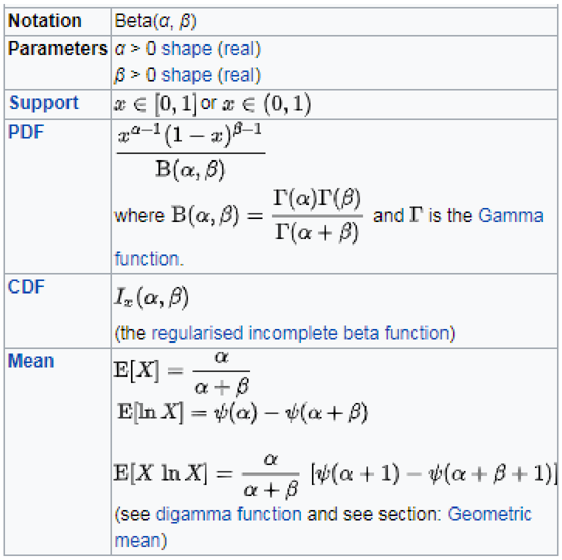
9. Bayesian inference for Naïve Bayes classifier (15 points)

The goal is to provide a Bayesian inference for the Naïve Bayes model with binary features and response variable. Assume the input data is D=$(x[i], y[i])\quad i = 1, \dots, N$ where N is the number of observations and $x[i] = \left[x_1[i], \dots, x_p[i]\right]^T$ (p: number of features).

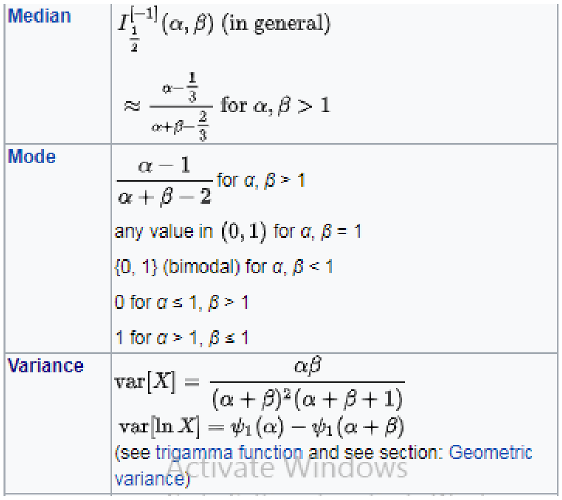
a) Write down the model, its parameters, and likelihood function for p=2.

b) For Bayesian inference, choose the beta prior for parameters (you may choose the same prior for all parameters). Write down the posterior distribution and the MAP estimate for parameters. Necessary formula for Beta distribution are given at the end of this question.

c) Despite the fact that the closed-form expression for posterior distribution is known, for educational reasons, your task is to develop an MCMC algorithm for estimating the posterior distribution.



Notation	Beta(α, β)
Parameters	$\alpha > 0$ shape (real) $\beta > 0$ shape (real)
Support	$x \in [0,1]$ or $x \in (0,1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and Γ is the Gamma function.
CDF	$I_x(\alpha,\beta)$ (the regularised incomplete beta function)
Mean	$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha+\beta)$ $\mathrm{E}[X \ln X] = \frac{\alpha}{\alpha+\beta}\,[\psi(\alpha+1) - \psi(\alpha+\beta+1)]$ (see digamma function and see section: Geometric mean)



Median	$I^{[-1]}_{\frac{1}{2}}(\alpha,\beta)$ (in general) $\approx \frac{\alpha-\frac{1}{3}}{\alpha+\beta-\frac{2}{3}}$ for $\alpha,\beta > 1$
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$ any value in $(0,1)$ for $\alpha, \beta = 1$ {0, 1} (bimodal) for $\alpha, \beta < 1$ 0 for $\alpha \leq 1, \beta > 1$ 1 for $\alpha > 1, \beta \leq 1$
Variance	$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ $\mathrm{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha+\beta)$ (see trigamma function and see section: Geometric variance)