به نام خدا

دانشگاه صنعتی امیرکبیر (پلی تکنیک تهران)
دانشکده برق
یادگیری ماشین – نیم‌سال دوم 1401-1402

تمرین تئوری درس یادگیری ماشین

Collaboration Policy

You are to complete this assignment individually. However, you may discuss the general algorithms and ideas with classmates, TAs, peer mentors and instructor in order to help you answer the questions. But we require you to:

- not explicitly tell each other the answers
- not to copy answers or code fragments from anyone or anywhere
- not to allow your answers to be copied
- not to get any answer from the Web

If you have any questions regarding this assignment, please contact Mr. Abdollahpour, Ms. Kordi. (Two last questions: Mr. Aghdasian, Mr. Janani)

Telegram ID: @alirezaap_r9 @yeganehkordi @A_Aghdasian @pooya_9877

Submit by 18th Ordibehesht 1402, 11.59pm

Question 1 - Introduction to LDA [10 points]

Let $p_x(x|w_i)$ be arbitrary densities with means μ_i and covariance matrices Σ_i (not necessarily normal) for i=1, 2,… . Let $y = w^t x$ be a projection, and let the induced one-dimension densities $p(y|w_i)$ have means μ_i and variances σ_i^2.

Show that the criterion function: $J_1(w) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$ Is maximized by

$$w = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)$$

Question 2 - Group scattering [10 points]

Assume we have 2 classes Y_1 and Y_2, with sizes n_1 and n_2 respectively. The expression $J = \frac{1}{n_1 n_2}\sum_{y_i \in Y_1}\sum_{y_j \in Y_2}(y_i - y_j)^2$ measures the total within group scatter. Show that this within group scatter can be written as (m_i and s_i are mean and variance corresponding to class i):

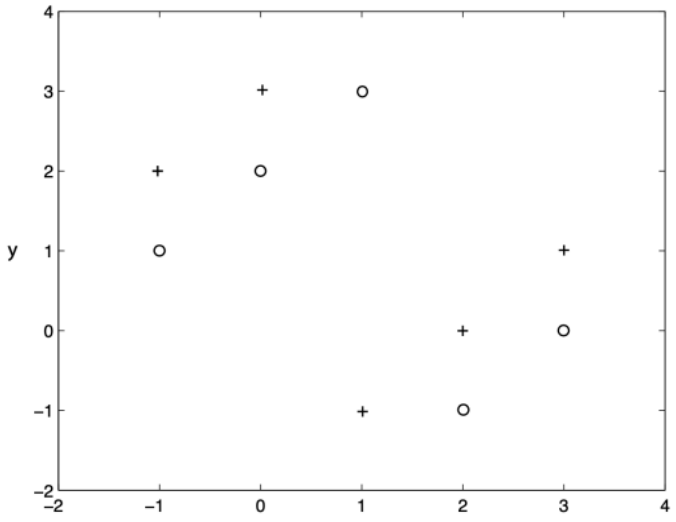
$$J = (m_1 - m_2)^2 + \frac{1}{n_1}s_1^2 + \frac{1}{n_2}s_2^2$$

Question 3 - Naive Bayes on Fashion-MNIST or MNIST [15 points]

By implementing forward selection algorithm, select the optimal number of features for best performance in classification by using Naive Bayes optimal classifier. (Consider Gaussian parametric estimate of pdf's). Plot correct classification rate as a function of the number of selected features to find the optimal number of features.

Question 4 - KNN: [10 points]

Consider K-NN using Euclidean distance on the following data set (each point belongs to one of two classes: + and ◦).



a) What is the leave one out cross validation error when using 1-NN?
b) Which of the following values of k leads to the minimum leave-one-out cross validation error: 3, 5 or 9? What is the error for that k? (If there is a tie, please elaborate)

Question 5 - Naïve Bayes Classifier: [20 point]

In order to reduce my email load, I decide to implement a machine learning algorithm to decide whether or not I should read an email, or simply file it away instead. To train my model, I obtain the following data set of binary-valued features about each email, including whether I know the author or not, whether the email is long or short, and whether it has any of several key words, along with my final decision about whether to read it (y = +1 for "read", y = −1 for "discard").

x_1 know author?	x_2 is long?	x_3 has 'research'	x_4 has 'grade'	x_5 has 'lottery'	y ⇒ read?
0	0	1	1	0	-1
1	1	0	1	0	-1
0	1	1	1	1	-1
1	1	1	1	0	-1
0	1	0	0	0	-1
1	0	1	1	1	1
0	0	1	0	0	1
1	0	0	0	0	1
1	0	1	1	0	1
1	1	1	1	1	-1

In the case of any ties, we will prefer to predict class +1. I decide to try a naïve Bayes classifier to make my decisions and compute my uncertainty.

1. Compute all the probabilities necessary for a naïve Bayes classifier, i.e., the class probability p(y) and all the individual feature probabilities p(xi | y), for each class y and feature xi.
2. Which class would be predicted for x = (0 0 0 0 0)? What about for x = (1 1 0 1 0)?
3. Compute the posterior probability that y = +1 given the observation x = (1 1 0 1 0).
4. Why should we probably not use a "joint" Bayes classifier (using the joint probability of the features x, as opposed to a naïve Bayes classifier) for these data?
5. Suppose that, before we make our predictions, we lose access to my address book, so that we cannot tell whether the email author is known. Should we re-train the model, and if so, how? (e.g.: how does the model, and its parameters, change in this new situation?) Hint: what will the naïve Bayes model over only features x2 . . . x5 look like, and what will its parameters be?

Question 6 - Logistic Regression [15 points]

We have the following two logistic regression models:

$$h_1(x) = \frac{1}{1 + e^{(-6x_0+2x_1+3x_2)}}$$

$$h_2(x) = \frac{1}{1 + e^{(6x_0-2x_1-3x_2)}}$$

a) What is the equation of the decision boundary for each model?

b) How would each model classify the samples of the following Data1 table?

Data 1

x_0	x_1	x_2	Class
0	4	1	1
1	2	0	1
3	4	2	1
2	0	1	0
0	-1	3	0

Question 7 - Linear Regression [20 points]

With the help of linear regression along with regularization, we have trained a model that after convergence of learning, the error of the validation model is high and almost equal to the error of the training data. What was happened? Which of the following solutions can help solve the problem and which can't? Why?

a) Increasing the regularization factor
b) Add new features to data and models
c) Adding to the number of training samples
d) Adding to the number of validation samples
e) Increase the learning rate

Good luck 😊