2EL1730: Machine Learning
CentraleSupélec

Assignment 2

Instructors: Fragkiskos Malliaros and Maria Vakalopoulou
TA: Enzo Battistella, Yoann Pradat, Jun Zhu

Due: **January 24, 2021 at 23:00**

How to submit: Please complete the first assignment **individually**. *Typeset* all your answers (**PDF** file only). Submissions should be made on **gradescope** (Assignment 2; Entry Code: D5JPBB). Those that have not used their full name while registering in Gradescope, please update this information (**FirstName LastName**). Make sure that the answer to each question is on a **separate page** (questions 1-8).

I. Decision trees

Question 1 [5 points]

Choose the correct answer or answers *with justification*. No partial credit will be given. All the correct answers should be selected.

(a) [3 p] Averaging the output of multiple decision trees helps:

1. Increase bias
2. Decrease bias
3. Increase variance
4. Decrease variance

(b) [2 p] For datasets with high label noise (many data points with incorrect labels), random forests would generally perform better than adaboost.

1. True
2. False

Question 2 [15 points]

Imagine that you were an area chair for one of the oncoming machine learning conferences and you had to deal with 10 different papers. After the review process have finished, the decision (acceptance/ rejection) had been made for these papers using the following criteria: (i) scientific novelty (SN), (ii) clarity of writing (CW) and (iii) reproducibility of the method (RM). From these criteria the CW and RM are nominal (Yes/No) and the SN is ordinal taking the following values $(1, 2, 3, 4, 5)$ with 1 denoting the lower score and 5 the highest. Your dataset is summarised bellow:

Paper Id	CW	RM	SN	Decision
1	Yes	Yes	4	Accept
2	Yes	No	4	Reject
3	No	No	2	Reject
4	No	Yes	3	Reject
5	Yes	Yes	2	Accept
6	No	Yes	5	Accept
7	Yes	Yes	2	Reject
8	No	No	1	Reject
9	Yes	No	1	Reject
10	No	Yes	4	Accept

Using these samples, you want to train a decision tree algorithm to help you decide for the next year submissions. Please address the following questions.

(a) [5 p] Using the Entropy measure find the best threshold for calculating the binary split of the SN attribute. Report the Entropy for each of the splits. (*Hint: read carefully Sections 3.3.1 of Introduction to Data Mining Book* for the definition/ use of the entropy measure).

(b) [7 p] Using the Gini Index, build the best decision tree for this dataset. If needed, use again the Entropy measure to define the best threshold value for the SN attribute (use only binary spliting for this attribute). Please provide the Gini Index of each attribute for each split and your final decision tree model.

(c) [3 p] In a short text comment your results. Is there any problem on the dataset or your model? Would you trust this model for your next year decision? How would you address the possible problem(s)? (*Hint: do all the leafs of the algorithm give you the proper class?*).

II. SVMs

Question 3 [5 points]

Let's assume that you train an SVM classifier using a polynomial kernel $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^p$, for various degrees p ranging in $p = \{1, 2, 3, 4\}$. Let's also assume that parameter C is chosen using cross-validation. Below, you can see different cases about the train and test error. Which one(s) could be found in practice (i.e., are more realistic)? Choose the correct answer or answers *with justification*.

1.

p	1	2	3	4
Train error	0.36	0.31	0.24	0.17
Test error	0.25	0.21	0.18	0.23

2.

p	1	2	3	4
Train error	0.16	0.23	0.28	0.33
Test error	0.29	0.24	0.31	0.35

3.

p	1	2	3	4
Train error	0.24	0.20	0.18	0.13
Test error	0.32	0.26	0.23	0.28

4. All these cases might occur in practice.

Question 4 [20 points]

In the class, we have examined the case of soft-margin SVM, assuming L_1 regularization. Let's consider a different formulation, assuming that L_2 regularization is used:

$$\begin{aligned} &\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \\ &\text{subject to} \quad y^i(\mathbf{w}^\top \mathbf{x}^i + b) \geq 1 - \xi_i, \quad i = 1, \dots, n \end{aligned}$$

(a) [3 p] In the optimization problem above, notice that we have dropped the $\xi_i \geq 0$ constraint (contrary to what we have seen in class about the L_1 formulation of the problem). Discuss why this can be done without affecting the optimal solution of the objective function.

(b) [3 p] Express the Lagrangian $\mathcal{L}(\mathbf{w}, b, \xi, \alpha)$ of this SVM optimization problem. (*Hint: read carefully Sections 7.1 and 7.1.1 of Bishop's book. Your goal is to end up with a formulation similar to Eq. (7.22), but for the L_2 case*).

(c) [7 p] In this subquestion, your goal is to minimize the Lagrangian computed in subquestion (b), with respect to $\mathbf{w}, b$ and ξ. In other words, you have to compute $\nabla_w \mathcal{L}$, $\frac{\partial \mathcal{L}}{\partial b}$, and $\nabla_\xi \mathcal{L}$, and then to set them equal to zero. Here, $\xi = [\xi_1, \dots, \xi_n]^\top$.

(d) [7 p] Express the dual formulation of the problem, showing the objective $\mathcal{L}(\alpha)$ and the constraints. (*Hint: using the results from subquestion (c), you can eliminate $\mathbf{w}$, b, and ξ, simplifying the Lagrangian, and expressing it as a function of α. In the final formulation, only terms of $\alpha, y, \mathbf{x}$ and C will appear*).

III. Neural Networks

Question 5 [15 points]

The table below is a list of sample points in R^2. Suppose that we run the perceptron algorithm and Rosenblatt's training algorithm, on these sample points. We record the total number of times each point participates in a stochastic gradient descent step because it is misclassified, throughout the run of the algorithm.

x1	x2	y	times misclassified
-3	2	+1	0
-1	1	+1	0
-1	-1	-1	2
2	2	-1	1
1	-1	-1	0

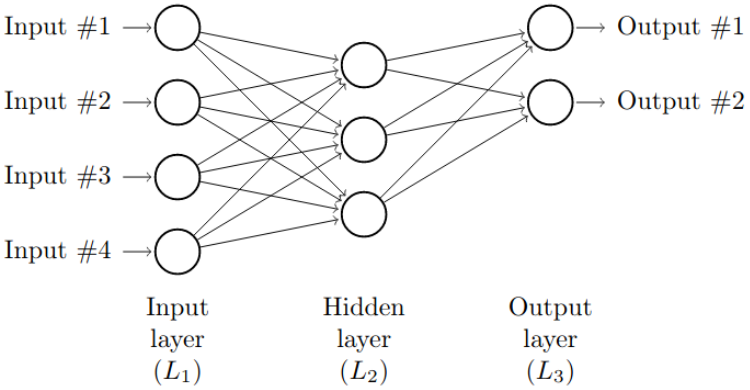
(a) [8 p] Suppose that the learning rate is $\eta = 1$ and the initial weight vector is $w^0 = (-3, 2, 1)$, where the last component is the bias term. What are the final weight vector after the training process is finished?

(b) [2 p] In some cases, removing even a single point can change the calculated weighs by the perceptron algorithm. For which, if any point(s) in our dataset would the learned decision boundary change if we removed it? Explain your answer.

(c) [5 p] How would our result differ if we were to add the additional training point $(2, -2)$ with label $+1$?

Question 6 [20 points]

Consider a three layer fully-connected network with n_1,n_2,n_3 neurons in three layer respectively. Inputs are fed into the first layer. The loss is mean squared error E, and the non-linearity is a sigmoid function. Let the label vector be t of size n_3. Let each later output vector be y_i and input vector be z_i, both of size n_i. Let the weight between layer i and layer $i+1$ be W_{ii+1}. The j-th element in y_i is defined by y_i^j, same for z_i^j. The weight connecting k-th and l-th neuron in i, $i+1$ layers is defined by W_{ii+1}^{kl} (You don't need to consider bias in this problem).



Here is a summary of our notation:

- σ denotes the activation function for L_2 and L_3, $\sigma(x) = \frac{1}{1+e^{-x}}$. There is no activation applied to the input layer.
- $z_i^{(j)} = \sum_{k=1}^{P} W_{i-1i}^{kj} x_{i-1}^{(k)}$
- $y_i^{(j)} = \sigma(\sum_{k=1}^{P} W_{i-1i}^{kj} x_{i-1}^{(k)})$

Now please provide solutions for the following problems.

(a) [6 p] Find the $\frac{\partial E}{\partial z_3^j}$ in terms of y_3^j and t^j

(b) [7 p] Find the $\frac{\partial E}{\partial y_2^k}$ in terms of elements in W_{23} and $\frac{\partial E}{\partial z_3^j}$

(c) [7 p] Find the $\frac{\partial E}{\partial W_{23}^{kj}}$ in terms of y_2^k, y_3^j and t^j

IV. Unsupervised Learning

Question 7 [5 points]

Choose the correct answer or answers *with justification*. No partial credit will be given. All the correct answers should be selected. *[Keep your answer short]*.

(a) [3 p] PCA is sometimes used as a preprocessing step before applying a regression algorithm. Why is this happening?

1. We reduce overfitting by removing poorly predictive dimensions.

2. We reduce the running time of regression.
3. We deal with missing information from the data.
4. We ensure that data will be linearly separable.

(b) [2 p] With the SVD, we have $X = U\Sigma V^\top$. Indicate the matrices that are the eigenvectors of the columns of U.

1. $X^\top X$
2. $XX^\top$
3. $X^\top XX^\top X$
4. $XX^\top XX^\top$

Question 8 [15 points]

Consider 3 data points in the two dimensional space: $(-1,-1)$, $(0,0)$, and $(1,1)$.

(a) [4 p] We perform PCA on the data points. What is the first principal axis (please give the vector; you don't have to explicitly solve the eigenvalue problem to find the vectors)?

(b) [4 p] Assume that we project the points into the one dimensional space defined by the first principal axis. What are the coordinates of the data points in the one dimensional space? In other words, find the first principal component of the data.

(c) [2 p] What is the variance of the projected data?

(c) [5 p] In subquestion (b), you have computed the projected data into the one dimensional space. Now, let's assume that you represent the data back in the original two dimensional space. How close are the points to the original ones (i.e., what is the error)?