Machine Learning Fundamentals
Final Exam
2020-2021

Massih-Reza Amini

Duration: 2 hours, authorized documents: Slides of the course

This exam is not intended to trap you, but to test your knowledge and to see what you have learned in this class. So there is no need to cheat.

Question 1 (4 pt)

1.1 Explain the Occam Rasor principle.

1.2 Explain the backtracking line-search algorithm.

1.3 What are the universal approximators seen in class? Does the property of "universal approximation" ensure that the empirical error on any training set will be equal to 0 (Explain why)?

1.4 Present and explain the three assumptions in semi-supervised learning?

Question 2 (4 pt)
Consider the following binary classification :

$$S = \left\{ \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, +1 \right); \left(\begin{pmatrix} -1 \\ 0 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 2 \\ -1 \end{pmatrix}, -1 \right) \right\}$$

2.1 Draw the points in a orthonormal basis of dimension 2.

2.2 We consider the perceptron algorithm for learning; will the algorithm converge ?

2.3 We initialize the weights and the bias to zero, and we fix the learning rate to 1. We consider that the order of points are taken is anti-clockwise when beginning from the point $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$, i.e. $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ then $\begin{pmatrix} 2 \\ -1 \end{pmatrix}$.... What are the weights found by the algorithm after convergence in this case?

2.4 What is the equation of the decision boundary?

2.5 What is the theoretical maximum number of iterations that ensures the convergence of the algorithm ?

Question 3 (2 pt)
We apply the Adaboost algorithm over a training set of size 5;

$$S = \{(\mathbf{x}_i, y_i); i \in \{1, \ldots, 5\}\} \in (\mathcal{X} \times \{-1, +1\})^5$$

3.1 At step 1, the examples are assigned uniform weights: $\forall i, D_1(i) = \frac{1}{5}$. We suppose that after learning the first classifier $h_1 : \mathcal{X} \rightarrow \{-1, +1\}$ the latter misclassifies 1 example 1 of S. Estimate the error $\epsilon_1 = \sum_{i:h_1(\mathbf{x}_i) \neq y_i} D_1(i)$ and deduce the weight α_1 associated to h_1 found by the algorithm.

3.2 Estimate new weights D_2 of misclassified and well classified examples by h_1.

Question 4 (10 pt)
We consider the CEM algorithm for partitioning a collection $\mathcal{C} = (\mathbf{x}_i)_{1 \leq i \leq N}$ of N examples represented in a vector space of dimension d, into K groups $\mathcal{G} = (G_k)_{1 \leq k \leq K}$.

Classification **E**xpectation **M**aximization[1]

Begin with an initial partition $\mathcal{G}^{(0)}$.
$\ell \leftarrow 0$
while $\mathcal{L}(\mathcal{C}, \Theta^{(\ell+1)}, \mathcal{G}^{(\ell+1)}) - \mathcal{L}(\mathcal{C}, \Theta^{(\ell)}, \mathcal{G}^{(\ell)}) > \epsilon$ **do**
E-step Estimate the posterior probabilities using the current parameters $\Theta^{(\ell)}$:

$$\forall k = \{1, \ldots, K\} \mathbb{E}[t_{ik} \mid \mathbf{x}_i, \mathcal{G}^{(\ell)}, \Theta^{(\ell)}] = \frac{\pi_k^{(\ell)} P(\mathbf{x}_i \mid G_k^{(\ell)}, \theta_k^{(\ell)})}{\sum_{j=1}^{K} \pi_j^{(\ell)} P(\mathbf{x}_i \mid G_j^{(\ell)}, \theta_j^{(\ell)})}$$

C-step Assign to each example $\mathbf{x}_i$ its partition, the one for which the posterior probability is maximum. Note $\mathcal{G}^{(\ell+1)}$ this new partition

M-step Estimate the new parameters $\Theta^{(\ell+1)}$ which maximize$\mathcal{L}(\mathcal{C}, \Theta^{(\ell)}, \mathcal{G}^{(\ell+1)})$

$\ell \leftarrow \ell + 1$
end while

---

[1] Gilles Celeux, Gérard Govaert. *A classification EM algorithm for clustering and two stochastic versions*. Computational Statistics & Data Analysis. 14(3), pp. 315–332, 1992.

Where,

$$\begin{aligned}\mathcal{L}(\mathcal{C},\Theta,G) &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}\log P(\mathbf{x}_i, G_k, \theta_k)\\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}\log \underbrace{P(G_k)}_{\pi_k} P(\mathbf{x}_i \mid G_k, \theta_k)\end{aligned}$$

is the complete log-likelihood; Θ is the set of parameters; and $\mathbf{t}_i = (t_{i1},\dots,t_{ik},\dots,t_{iK})$ is the cluster vector indicator of observation $\mathbf{x}_i$ (i.e. $\mathbf{x}_i \in G_k$ iff $t_{ik}=1$ and $\forall j \neq k; t_{ij} = 0$).

4.1 Explain the algorithm seen in class.

4.2 Show that at each iteration ℓ, the complete log-liklihood $\mathcal{L}(\mathcal{C},\Theta,G)$ increases i.e.

$$\forall \ell \geq 0; \mathcal{L}(\mathcal{C},\Theta^{(\ell+1)},G^{(\ell+1)}) \geq \mathcal{L}(\mathcal{C},\Theta^{(\ell)},G^{(\ell)})$$

4.3 Deduce that the algorithm converges to a local maximum of $\mathcal{L}(\mathcal{C},\Theta,G)$.

4.4 We suppose that samples of different clusters are generated by multivariate normal distributions:

$$P(\mathbf{x} \mid G_k, \theta_k) = \frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}_k|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)}.$$

Further, we suppose the following:

(H1) The covariance matrices of all groups are the identity matrix: $\forall k \in \{1,\dots,K\}; \mathbf{\Sigma}_k = Id_d$;

(H2) The probability of clusters is the uniform probability: $\forall k \in \{1,\dots,K\}, P(G_k) = \frac{1}{K}$.

What is the set of parameters Θ in this case?

4.5 Deduce that the complete log-likelihood writes:

$$\mathcal{L}(\mathcal{C},\Theta,G) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}\|\mathbf{x}_i - \mu_k\|^2 + Constant \tag{1}$$

4.6 For which values of $\mu_k; k \in \{1, \ldots, K\}$, Equation (1) is maximized?

4.7 With assumptions (H1) and (H2), the CEM algorithm reduces to which clustering algorithm seen in class?