Ε.Μ.Π., Σχολή Η.Μ.& Μ.Υ. **Αναγνώριση Προτύπων** Ακαδ.Ετος 2022-23
Ημερ/νία: 21-12-2022 **2ο Σύνολο Αναλυτ. Ασκήσεων** Παραδοτέο: 11-01-2023

Οι αναλυτικές σειρές ασκήσεων είναι ατομικές, και οι λύσεις που θα δώσετε πρέπει να αντιπροσωπεύουν μόνο την προσωπική σας εργασία. Εξηγήστε επαρκώς την εργασία σας. Αν χρησιμοποιήσετε κάποια άλλη πηγή εκτός των σημειώσεων για τη λύση σας, πρέπει να το αναφέρετε. Η παράδοση των λύσεων των αναλυτικών ασκήσεων της σειράς αυτής θα γίνει ηλεκτρονικά στο helios και θα πρέπει να την υποβάλετε ως ένα ενιαίο αρχείο PDF με το εξής filename format, χρησιμοποιώντας μόνο λατινικούς χαρακτήρες: pr22_hwk2_AM_FirstnameLastname.pdf, όπου ΑΜ είναι ο 8-ψήφιος αριθμός μητρώου σας. Σκαναρισμένες χειρόγραφες λύσεις επιτρέπονται, αρκεί να είναι καθαρογραμμένες και ευανάγνωστες. Επίσης, στην πρώτη σελίδα των λύσεων θα αναγράφετε το ονοματεπώνυμο, ΑΜ, και email address σας.

Υλικό για Ανάγνωση:
Βιβλία: [1], [2], [3] και [4]
Διαφάνειες διαλέξεων μαθήματος

Αναλυτικές Ασκήσεις

Άσκηση 2.1: (Feed Forward Neural Networks)
One of the key theorems illustrating the expressive power of neural networks is the Universal Approximation Theorem, which states (simplified): *Given a non-polynomial element-wise (activation) function σ, and parameters $W_1 \in \mathbb{R}^{D_1 \times n}, W_2 \in \mathbb{R}^{D_2 \times D_1}, b \in \mathbb{R}^{D_1}$ of arbitrary width (D_1, D_2), the function $f(x) = W_2 \cdot \sigma(W_1 \cdot x + b), \quad x \in \mathbb{R}^n$ can approximate any function g with arbitrary low error $\|f(x) - g(x)\| < \epsilon, \quad \epsilon > 0$.*

1. Consider the network $f(x) = f^L(f^{L-1}(f^{L-2}(\dots f^1(x))))$, where $f^l = \sigma(W_l \cdot x + b_l)$. Show that if we choose the linear activation function $\sigma(x) = x$, then for any number of layers L, the network $f(x)$ is equivalent to a single-layer network $g(x) = Wx + b$

2. Consider two scalar variables x_1 and x_2, and the multiplication function $f(x_1, x_2) = x_1 \cdot x_2$, and the network $\tilde{f}(x) = W_2 \cdot \sigma(W_1 \cdot x + b_1)$. Let:

$$W_1 = \begin{bmatrix} \lambda & \lambda \\ -\lambda & -\lambda \\ \lambda & -\lambda \\ -\lambda & \lambda \end{bmatrix} \quad b_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix} \quad W_2 = \begin{bmatrix} \mu & \mu & -\mu & -\mu \end{bmatrix} \tag{1}$$

   where $\mu = (4\lambda^2 \ddot{\sigma}(0))^{-1}$. Write the expression of $\tilde{f}(x_1, x_2)$ wrt to $\sigma, \lambda, \mu, x_1, x_2$. ($\ddot{\sigma}(0)$ is the second derivative of σ at $x = 0$).

3. Show that $\tilde{f}(x) \longrightarrow f(x)$ as $\lambda \longrightarrow 0$. Hint: Use the second order Taylor approximation for σ around the origin.

Άσκηση 2.2: (Recurrent Networks)
Consider the vanilla RNN in Eq. (2), trained using the loss function L.

$$h_t = f(W \cdot h_{t-1} + U \cdot x_t) \tag{2}$$
$$o_t = f(y_t) = f(V \cdot h_t) \tag{3}$$

In the above Equation, f is an activation function, e.g. tanh.

1. Use the chain rule to write the derivative $\frac{\partial L_t}{\partial V}$ of the loss function L at time-step t wrt to the weight V. The final expression should be in terms of L_t, o_t, y_t, h_t.

2. Write the partial derivative $\frac{\partial L_t}{\partial W}$, of L_t wrt the weight W, in terms of $L_t, h_t, o_t, W, f(\cdot)$.

3. Argue, when t is large, the gradient $\frac{\partial L_t}{\partial W}$ can explode / vanish. You can ignore the effect of the activation function for your argument (assume f is the identity function). Hint: use the power iteration method to argue about the product W^k, as $k \longrightarrow \infty$

Άσκηση 2.3: (SVM)
Consider the problem of separating a set of training vectors into two classes. The training data can be written in the form $\{(x_i, y_i)\}$, where the feature vectors are $x_i \in \mathbb{R}^m$ and the labels of the classes $y_i \in \{1, 1\}$.
In the case where the training data are not linearly separable (for example via a decision rule $sign(wx + b)$ for some w, b), then the problem needs to be formulated using slack variables $\{\xi_i\}, 1 \leq i \leq n$. Therefore, the SVM classifier with the largest margin is obtained by solving the dual problem:

$$L(\mathbf{w}, b, \alpha, \xi, \beta) = \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left[(\mathbf{w} \cdot \mathbf{x}_i + b)\, y_i - (1 - \xi_i)\right] - \sum_{i=1}^{n} \beta_i \xi_i$$

where C is a constant, $\alpha_i, \beta_i \geq 0, \forall i$ are the Lagrangian multipliers, and $_i \geq 0$ are the slack variables.

1. Let $n = 4$ and x 2-dimensional, $< x_i^1, x_i^2 >:< 2, 2 >, < 2.5, 2.5 >, < 5, 5 >, < 7, 7 >$. The SVM is trained using the aforementioned equation. Show that for any labels y that the four training tuples may have, the optimal weight parameter $\tilde{\mathbf{w}} = (\tilde{w}^1, \tilde{w}^2)$ satisfies $\tilde{w}^1 = \tilde{w}^2$.

2. Now consider the SVM training, but this time neglect the bias term ($b = 0$). We use a kernel $\mathbf{K}(\mathbf{u}, \mathbf{v})$ which satisfies $-1 < \mathbf{K}(\mathbf{u}, \mathbf{v}) < 1$ for any $\mathbf{u}, \mathbf{v}$ that belong to the training data. Furthermore, $\mathbf{K}(\mathbf{u}, \mathbf{u}) < 1$. Show that if there exist n training samples and $C < \frac{1}{n-1}$, then all dual variables α_i are non-zero (and therefore all training samples are support vectors).

3. Use the following kernel:

   $$\mathbf{K}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} + 4(\mathbf{u} \cdot \mathbf{v})^2$$

   where vectors $\mathbf{u}$ and $\mathbf{v}$ are 2dimensional. This kernel is equal with the inner product $\phi(\mathbf{u}) \cdot \phi(\mathbf{v})$ for some function ϕ. Which is this function?

Άσκηση 2.4: (LDA)
Let $p_{\boldsymbol{x}}(\boldsymbol{x} \mid \omega_i)$ be some probability densities with mean values (vectors) $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ - not necessarily Gaussian, for $i = 1, 2$. Let $y = \boldsymbol{w}^T \boldsymbol{x}$ be a projection and the corresponding one-dimensional densities $p(y \mid \omega_i)$ have mean values μ_i and variances σ_i^2. Show that the following criterion

$$J_1(\boldsymbol{w}) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

is maximized by the weight vector

$$\boldsymbol{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{4}$$

Άσκηση 2.5: (Bayesian Networks)
Consider the following Figure, which illustrates a car fuel system. You are given the following

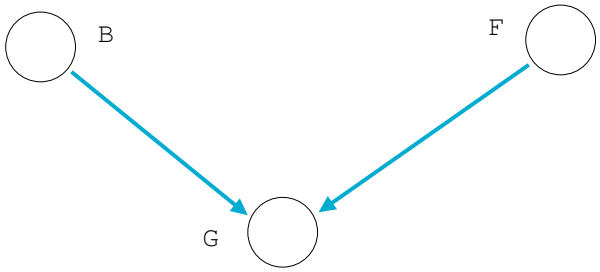

Figure 1: Car Fuel System. Battery (B), Fuel Tank (F) and Gauge (G).

probabilities

$$p(B = 1) = 0.95$$
$$p(F = 1) = 0.8$$
$$p(G = 1 \mid B = 1, F = 1) = 0.95$$
$$p(G = 1 \mid B = 1, F = 0) = 0.3$$
$$p(G = 1 \mid B = 0, F = 1) = 0.25$$
$$p(G = 1 \mid B = 0, F = 0) = 0.2$$

Assume that instead of observing the state of the fuel gauge G directly, the gauge is seen by the driver D who reports to us the reading on the gauge. This report is either that the gauge shows full $D = 1$ or that it shows empty $D = 0$. Our driver is a bit unreliable, as expressed through the following probabilities.

$$p(D = 1 \mid G = 1) = 0.80$$
$$p(D = 0 \mid G = 0) = 0.80$$

Suppose that the driver tells us that the fuel gauge shows empty, in other words that we observe $D = 0$.

1. Evaluate the probability that the fuel tank (F) is empty, given only this observation.

2. Evaluate the probability that the fuel tank (F) is empty, given also the observation that the battery is flat. Which probability is lower? Comment on the results.

ΒΙΒΛΙΟΓΡΑΦΙΑ:

[1] Γ. Καραγιάννης και Γ. Σταϊνχάουερ, *Αναγνώριση Προτύπων και Μάθηση Μηχανών*, ΕΜΠ, 2001.

[2] R. O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification*, Wiley, 2001.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th Edition Academic Pres, Elsevier, 2009. *Ελληνική μετάφραση*: απόδοση-επιμέλεια-πρόλογος ελληνικής έκδοσης Α. Πικράκης, Κ. Κουτρουμπάς, Θ. Γιαννακόπουλος, Επιστημονικές Εκδόσεις Π.Χ. Πασχαλίδης-Broken Hill Publishers LTD, 2012.