CS-E4710 - Machine Learning: Supervised Methods D, Lecture, 6.9.2022-12.12.2022

Started on	Tuesday, 21 February 2023, 9:06 AM
State	Finished
Completed on	Tuesday, 21 February 2023, 11:59 AM
Time taken	2 hours 52 mins

Quiz navigation

1 2 3 4 5 6 7 8 9 10 11 12 13 14

Finish review

Question **1**

Flag question Marked out of 5.00 Complete

Which assumption is correct in a general learning setup:

a. The learner has access to the the function that generates the true labels, but not to the underlying distribution of the data

b. The learner does not know the underlying distribution of the data or the function generating the true labels

c. The empirical error computation needs to have access to the underlying distribution generating the data

Question **2**

Flag question Marked out of 5.00 Complete

Consider a consistent classifier with hypothesis set size $|\mathcal{H}| = 52$ and sample size m. How many samples are needed to keep the generalization error the same, if the confidence level is changed by changing δ from 0.05 to 0.1?

a. $0.63m$

b. $0.90m$

c. $1.11m$

d. $2.07m$

Question **3**

Flag question Marked out of 5.00 Complete

Which of the following claims is true?

a. Rademacher complexity depends on the distribution generating the data

b. Rademacher complexity measures the performance of the learning algorithm in the worst-case scenario of assigning labels to samples in adversarial way

c. It is not possible to estimate Rademacher complexity empirically

Question **4**

Flag question Marked out of 5.00 Complete

Generalization error of a hypothesis class can be partitioned as

$$R(h) = R^* + \epsilon_{estimation} + \epsilon_{approximation}$$

into Bayes error R^*, approximation error $\epsilon_{approximation}$ (also known as bias) and estimation error $\epsilon_{estimation}$ (also known as variance).
Which of the following claims is true?

a. The Bayes error term R^* cannot be reduced

b. The Bayes error term R^* can be reduced by choosing a good hypothesis class $\mathcal{H}$

c. $\epsilon_{approximation}$ can be reduced by using more training data

Question **5**

Flag question Marked out of 5.00 Complete

Let x be chosen from the interval $[-1, +1]$, and the labels, y, from the set $\{0, 1\}$. we are given the following conditional probabilities for all x and y.

$$\begin{aligned} pr(1|x) &= \begin{cases} +x+1 & x \in [-1, 0] \\ -x+1 & x \in [0, +1] \end{cases} \\ pr(0|x) &= \begin{cases} -x & x \in [-1, 0] \\ +x & x \in [0, +1] \end{cases} \end{aligned}$$

What is the Bayes error of the Bayes classifier relating to this model if x is uniformly distributed on $[-1, +1]$?

Hint: you might solve the problem by drawing a plot representing the functions of the conditional probabilities.

a. 0.75

b. 0.5

c. 0.3

d. 0.25

Question **6**

Flag question Marked out of 5.00 Complete

Let us consider the perceptron algorithm and a linearly separable training set. Novikoff's theorem states an upper bound on the number of iterations required that for every training example, $(\mathbf{x}_i, y_i)$ satisfies $y_i \mathbf{w}^T \mathbf{x}_i \geq 0$. How does that bound depend on the largest achievable geometric margin?

a. It is independent from that margin.

b. Increases when the margin increases.

c. Decreases when the margin increases.

Question **7**

Flag question Marked out of 5.00 Complete

For a given sample of inputs and outputs, $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, m$, in the dual of the soft-margin SVM, the normal vector of the separating hyperplane, $\mathbf{w}$, can be expressed as $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$, where $\{\alpha_i\}$ are the optimal dual variables. Which of these statement is true?

a. α_i is greater than 0 for all i

b. For a given i, α_i greater than 0 if $y_i = 1$

c. For a given i, α_i is greater than 0 if $\mathbf{x}_i$ has functional margin smaller or equal to 1.

Question **8**

Flag question Marked out of 5.00 Complete

Let the feature embedding $\phi(\mathbf{x})$ underlying the polynomial kernel, $n_{pol}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T \mathbf{z} + c)^q$, be represented explicitly as a vector. What is the dimension of the explicit feature vector for polynomial kernel of degree $q = 3$ when inputs satisfy $\mathbf{x} \in \mathbb{R}^3$.

a. 20

b. 10

c. 15

Question **9**

Flag question Marked out of 5.00 Complete

Which of the following statements about multi-layer perceptrons (MLP) is true?

a. Stochastic gradient descent is guaranteed to find the parameters for a MLP that minimize the zero-one error on training data.

b. MLPs can be used to represent arbitrary boolean functions

c. Using a linear activation function in each neuron in each layer results in a MLP that computes a non-linear function of the inputs.

Question **10**

Flag question Marked out of 5.00 Complete

Assume 17 base learners with independent true risk 0.45 ($\epsilon = 0.45$), are grouped to form a strong learner via majority voting. What is the probability of having an incorrect aggregated prediction?

a. 0.3

b. 0.34

c. 0.18

d. 0.25

Question **11**

Flag question Marked out of 5.00 Complete

In each iteration of AdaBoost, the weights of the examples are modified to increase the weights on the mistakes made by :

a. The last weak learner

b. The ensemble in the previous iteration

c. The ensemble including the last weak learner

Question **12**

Flag question Marked out of 5.00 Complete

Which option is true?

a. For feature selection, backward elimination is computationally not as demanding as forward selection.

b. The performance of a machine learning algorithm can vary a lot based on the feature transformation applied to the data

c. While l_1 norm regularization induces sparsity in the learning problem, computationally the machine learning problem is not convex.

Question **13**

Flag question Marked out of 5.00 Complete

With the following weight matrix $\mathbf{W}$ for a multi-class SVM, to which class (numbering starts from 0) will a new sample $\mathbf{x} = (-2, 1)$ be classified into?

$$\mathbf{W} = \begin{bmatrix} 2 & -3 & 0.5 & 1 & -2.5 \\ 0 & 0.5 & -1 & -2 & 1 \end{bmatrix}$$

a. 0

b. 1

c. 2

d. 3

e. 4

Question **14**

Flag question Marked out of 5.00 Complete

Alice and Bob wish to have a literature circle, and consider Alice's adventures in Wonderland (AAW), Diskworld series (DW), Dune, Foundation series and Pride and Prejudice (PP).
Alice prefers the order $AAW \succ DW \succ PP \succ Dune \succ Foundation$ while
Bob prefers the order $PP \succ Dune \succ AAW \succ DW \succ Foundation$.

To resolve the conflict they decide that they should read the books in order σ that has a small value $d_{max}(\sigma) = \max d_K(\sigma, \sigma_{Alice}), d_K(\sigma, \sigma_{Bob})$, where d_K is the Kendall's distance, meaning that σ is close to both Alice's and Bob's preferred order.

Which of the following orders would be the best for Alice and Bob:

a. $PP \succ DW \succ Dune \succ AAW \succ Foundation$

b. $AAW \succ PP \succ DW \succ Dune \succ Foundation$

c. $PP \succ AAW \succ DW \succ Dune \succ Foundation$

Finish review

Grades after course exam 12.12.2022