Started on	Thursday, 31 March 2022, 12:38 PM
State	Finished
Completed on	Thursday, 31 March 2022, 2:38 PM
Time taken	2 hours
Grade	**27.00** out of 100.00

Information

Eigenständigkeitserklärung / Declaration of Independence

By proceeding further, you confirm that you are completing the exam alone, without other resources than those that are authorized. Authorized resources include the ML1 course material, personal notes, online APIs documentation, and calculation/plotting tools.

Question 1

Incorrect

Mark 0.00 out of 5.00

Which of the following is **True**: Let k be a Mercer (PSD and symmetric) kernel and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ be an unlabeled dataset. A Gram matrix K of size $N \times N$ associated to this kernel and dataset always satisfies:

● a. $KK^\top = I$. ✗

○ b. $K^\top = K^{-1}$.

○ c. $\forall\, \boldsymbol{u} \in \mathbb{R}^N :\ \boldsymbol{u}^\top K \boldsymbol{u} \geq 0$.

○ d. $\forall_{i=1}^N \forall_{j=1}^N : K_{ij} > 0$.

Your answer is incorrect.

The correct answer is:
$\forall\, \boldsymbol{u} \in \mathbb{R}^N :\ \boldsymbol{u}^\top K \boldsymbol{u} \geq 0$.

Question 2

Incorrect

Mark 0.00 out of 5.00

Which of the following is **False** : PCA finds directions in input space for which:

○ a. The projection of non-centered data has maximum variance.

○ b. The projection centered data has maximum variance.

○ c. The projection of non-centered data has maximum sum-of-squares.

● d. The projection centered data has maximum sum-of-squares. ✗

Your answer is incorrect.

The correct answer is:
The projection of non-centered data has maximum sum-of-squares.

Question 3

Correct

Mark 5.00 out of 5.00

Which of the following is **True**: In explainable machine learning, Shapley values:

○ a. can be computed in the order of a single forward/backward pass.

● b. requires an exponential number of function evaluations to be computed. ✓

○ c. requires $O(d)$ function evaluations, where d is the number of input dimensions.

○ d. is a self-explainable model that must be trained alongside the actual model of interest.

Your answer is correct.

The correct answer is:
requires an exponential number of function evaluations to be computed.

Question 4

Correct

Mark 5.00 out of 5.00

Which of the following is **True**: Layer-wise relevance propagation (LRP) is a method for explainable AI that:

○ a. can be applied to any black-box machine learning model.

● b. assumes that the machine learning model has a neural network (or computational graph) structure. ✓

○ c. requires $O(d)$ function evaluations, where d is the number of input dimensions, in order to produce an explanation.

○ d. can be applied to any black-box model, with the only condition that the gradient w.r.t. the input features can be computed.

Your answer is correct.

The correct answer is:
assumes that the machine learning model has a neural network (or computational graph) structure.

Information

Assume you would like to build a neural network that implements some decision boundary in $\mathbb{R}^2$. For this, you have at your disposal neurons of the type

$$a_j = \sum_i a_i w_{ij} + b_j$$

where $\sum_i$ sums over the indices of the incoming neurons. Denote by a_1 and a_2 the two input neurons (initialized to the value x_1 and x_2 respectively). Denote by a_3, a_4, a_5 the hidden neurons. The output of the network implements

$$f(x) = v \cdot \max(a_3, a_4, a_5)$$

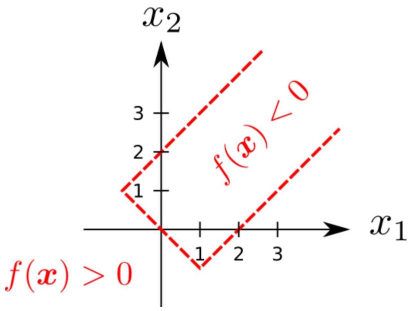
where v is an extra weight parameter.

Question 5

Complete

Mark 0.00 out of 10.00

Give the weights, biases and parameter v associated to a neural network with the structure above and that implements the decision boundary below:



Comment:

Question 6

Complete

Mark 0.00 out of 5.00

Give (i.e. evaluate) the derivative of the function implemented by your network w.r.t. the parameter v when the function is evaluated at $(x_1, x_2) = (3, 0)$.

Comment:

Information

We observe fishes arriving sequentially on a conveyor belt. Observations are drawn i.i.d. from an (unknown) distribution consisting of three classes (salmon, sea trout, pike). The sequence of observed fishes is denoted by $\mathcal{D} = (x_1, x_2, \ldots, x_N)$. We consider the probability model:

$$\begin{aligned} P(x_i = \text{salmon} \mid \theta) &= \theta/2 \\ P(x_i = \text{sea trout} \mid \theta) &= \theta/2 \\ P(x_i = \text{pike} \mid \theta) &= 1 - \theta \end{aligned}$$

where θ is a parameter between 0 and 1 that needs to be learned.

Question 7

Complete

Mark 5.00 out of 5.00

Assume we observe the sequence

x_1	x_2	x_3	x_4	x_5	x_6
salmon	sea trout	pike	sea trout	pike	salmon

State the likelihood function $P(\mathcal{D}|\theta)$ measuring the probability of making this sequence of observations as a function of the parameter θ.

$P\left(D|\theta\right) = \prod_{k=1}^{6} P\left(x|\theta\right)$

$P\left(D|\theta\right) = \left(\frac{\theta}{2}\right)^4 . \left(1-\theta\right)^2$

Comment:

Question 8

Complete

Mark 2.00 out of 5.00

Give for this sequence the optimal parameter θ in the maximum likelihood sense.

$\theta = argmax P\left(D|\theta\right)$

$\theta = argmax_\theta \left(4\log\left(\frac{\theta}{2}\right) + 2log\left(1-\theta\right)\right)$

$\bigtriangledown_\theta = \frac{10}{\theta} - \frac{2}{(1-\theta)}$

$\theta = \frac{10}{12}$

Comment:

Right formula, false derivative

Question 9

Complete

Mark 5.00 out of 5.00

Give the probability (according to the learned model) that the next two observed fishes (i.e. x_7 and x_8) are salmons.

$P\left(x_7, x_8|\theta\right) = P\left(x_7|\theta\right).P\left(x_8|\theta\right)$

$P\left(x_7, x_8|\theta\right) = \frac{10}{12}^2$

$P\left(x_7, x_8|\theta\right) = \frac{100}{144}$

Comment:

Follows from the mistake above

Question 10

Complete

Mark 5.00 out of 5.00

We now take a Bayesian view on the problem. We consider the prior distribution for the parameter θ to be

$$p(\theta) = 1,$$

defined on the interval $[0, 1]$. *Give* the equation for the posterior distribution $p(\theta|\mathcal{D})$.

$P\left(\theta|D\right) = \frac{P(D|\theta).P(\theta)}{\int_0^1 P(D|\theta).P(\theta)dx}$

We Know $P\left(D|\theta\right) = \left(\frac{\theta}{2}\right)^4.\left(1 - \theta\right)^2$

$P\left(\theta|D\right) = \frac{\left(\frac{\theta}{2}\right)^4.(1-\theta)^2.(1)}{\int_0^1 \left(\frac{\theta}{2}\right)^4.(1-\theta)^2.P(1)\theta}$

$P\left(\theta|D\right) = 105\theta^4.\left(1 - \theta\right)^2$

Comment:

Question 11

Complete

Mark 0.00 out of 5.00

Under this posterior distribution, give the probability that the next two observed fishes (i.e. x_7 and x_8) are salmons.

$$P\left(x_7, x_8|\theta\right) = \int_0^1 p\left(x_7|\theta\right) . p\left(\theta|D\right) dx . \int_0^1 p\left(x_8|\theta\right) . p\left(\theta|D\right) d\theta$$
$$P\left(x_7, x_8|\theta\right) = \int_0^1 \theta^2 \left[105\theta^4 . \left(1-\theta\right)^2\right] d\theta$$
$$P\left(x_7, x_8|\theta\right) = \frac{5}{12}$$

Comment:

In our Bayes setting as in Ex-Sheet 2 we only assume independence conditional on theta, which leads to $\int_0^1 (\theta/2)^2 \ldots$

Information

Shapley values provide a way of attributing the real-valued prediction $f(\boldsymbol{x})$ of some data point $\boldsymbol{x} \in \mathbb{R}^d$ on the input features. The Shapley values $\phi_1, \ldots, \phi_d$, measuring the contribution of each feature, are defined as:

$$\phi_i = \sum_{\mathcal{S}: i \notin \mathcal{S}} \frac{|\mathcal{S}|!(d-|\mathcal{S}|-1)!}{d!} \left[f(\boldsymbol{x}_{\mathcal{S}\cup\{i\}}) - f(\boldsymbol{x}_\mathcal{S})\right]$$

where $\sum_{\mathcal{S}: i \notin \mathcal{S}}$ sums over all subsets of features that do not contain feature i, and where $\boldsymbol{x}_\mathcal{S}$ is the input $\boldsymbol{x}$ where we have retained the subset of features $\mathcal{S}$ and removed other features. Shapley values assume a reference point which provides a replacement value for features that are removed. In this exercise, we define the reference point to be $\widetilde{\boldsymbol{x}} = \boldsymbol{0}$, i.e. we set features to zero when removing them.

Question 12

Not answered

Marked out of 7.50

Compute the Shapley values ϕ_1, ϕ_2, ϕ_3 for the function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ evaluated at the data point $\boldsymbol{x} = (3, 4, 0)$.

Question 13

Not answered

Marked out of 7.50

When the data is high-dimensional, computing Shapley values requires evaluating the function f exponentially many times, which is intractable. Instead, a popular explanation technique known as Gradient × Input, attributes according to the formula:

$$\phi_i = [\nabla f(\boldsymbol{x})]_i \cdot x_i$$

where $\nabla f(\boldsymbol{x})$ is the gradient of the function f evaluated at $\boldsymbol{x}$, and $[\cdot]_i$ extracts the ith dimension of the gradient.

Compute the Gradient × Input explanation for the function and data point above.

Question 14

Not answered

Marked out of 5.00

A desirable property of an explanation technique is conservation, specifically, the sum of contributions of all input variables should relate to the function value as

$$f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}}) = \sum_{i=1}^{d} \phi_i.$$

State whether the Shapley value explanation for the function and data point above is conservative, and *state* whether the Gradient × Input explanation for the function and data point above is conservative.

Information

In this exercise, we would like to measure empirically the bias and variance properties of some estimator built from observed data. We consider a sample

$$x_1, \dots, x_N \sim \mathcal{U}([0, 1]),$$

i.e. a collection of real values drawn iid. following a uniform distribution on the interval $[0, 1]$. We assume a function $f : [0, 1] \to \mathbb{R}$. Applying this function to each element in our sample (i.e. $\forall_{i=1}^{N} : \ y_i = f(x_i)$), gives us a new set of random variables

$$y_1, \dots, y_N$$

which we will use to build an estimator.

Question 15

Complete

Mark 0.00 out of 4.00

Let $f(x) = x^2 - x + 1$ be our function defined on the interval $[0, 1]$.

The function reaches its minimum value $\theta = \min_{x \in [0,1]} f(x) = 0.75$, and we consider the following estimator of such minimum value:

$$\hat{\theta} = \min\{y_1, \ldots, y_N\} - \epsilon$$

Write a function that takes a sample $y_1, \ldots, y_N$ as input (a numpy array of size N), and returns the estimation $\hat{\theta}$.

```
def estimate_theta(y):
    X = list.random[0,1]
    fx = x^2-x+1
    est_theta = min(0.75,fx)
        return est_theta
```

Comment:

Question 16

Not answered

Marked out of 4.00

Write a similar function, but that receives K samples (each sample is of size N), and compute the estimator for each sample. Specifically, the function should receive a matrix Y of size $K \times N$ and return a vector of size K containing the estimation $\hat{\theta}$ for each sample. Your implementation should be computationally efficient, i.e. make use of numpy vector operations whenever possible, and avoiding loops.

Question **17**

Complete

Mark 0.00 out of 8.00

Write a function that estimates the bias, variance, and mean square error of this estimator based on K samples.

The function receives as input the function you have implemented above. It should first build the matrix of size $K \times N$ containing the desired samples, then call the estimation function received as input, compare the estimates $\hat{\theta}$ to the true value θ, and return the desired quantities, specifically, a tuple of three real values containing the bias, the variance, and the mean square error of the estimator respectively.

Like for the exercise above, your implementation should be efficient (numpy vector operations, no loops, etc.).

```
def bve(estimate_theta_vec):

fx = numpy.array([predictor.fit(*K).predict(x)

for i in range (est_thetas)])


bias  = (x - f.mean(axis = 0)**2)

Variance = (x -)
      return (bias, variance, error)
```

Comment:

Question **18**

Not answered

Marked out of 4.00

Propose a better estimator. In this question, you are not asked to implement such an estimator, but simply to give a short textual description of the proposed estimator and explain its advantages. *(Hint: your estimator is not restricted to using the values $y_1, \ldots, y_N$, but can also be built on the inputs $x_1, \ldots, x_N$.)*

◄ Written Exam A

Jump to...

Lecture 1 (video) ►