(a) (3 Punkte) The Bayes error for classification is:

□ the lowest error achievable by a linear classifier.

□ the lowest error achievable by a nonlinear quadratic classifier

☑ the lowest error achievable by the best possible classifier

□ the lowest error achievable by a classifier assuming Gaussian-generated distributions.

(b) (3 Punkte) Independent Component Analysis can be achieved by:

□ Finding the directions in the input space that maximize the variance of the projected data.

□ Applying a whitening procedure to the data and running PCA on the whitened data.

☑ Finding the directions in the input space that maximize the skewness of the projected data.

□ Finding the directions in the input space that minimize the variance of the projected data.

(c) (3 Punkte) The K-means algorithm:

□ Is a convex algorithm that can be used to cluster the data.

☑ Is a nonconvex algorithm that can be used to cluster the data.

□ Is a kernelized version of the means algorithm where the kernel is Gaussian.

□ Is a kernelized version of the means algorithm where the kernel can be arbitrary.

(d) (3 Punkte) A biased estimator is sometimes used to:

□ Reduce the risk of underfitting the data.

☑ Reduce the estimation error for high-dimensional data.

□ Make the estimation procedure more sensitive to the observed data.

□ None of the above. We should always favor an unbiased estimator.

(e) (3 Punkte) Which is **False**? The Restricted Boltzmann machine is:

☑ A machine learning method that is based on error backpropagation.

□ A machine learning method that can learn initial weights for a neural network.

□ A machine learning method that estimates binary probability distributions for the data.

□ A machine learning method that can learn global and local features in the data.

(a) (3 Punkte) Let $f_1(x) \dots f_N(x)$ be a set of descriminats for classification. The classification decision is given by $c* = \underset{i}{\text{argmax}}\ f_i(x)$ Which of the following sets would produce the same classification as the one above?

- □ $g_i(x) = (f_i(x))^2$
- □ $h_i(x) = log(1 + exp(f_i(x)))$
- □ None of them
- ☑ Both of them

(b) (3 Punkte) Consider a two-class classification problem. A sufficient condition for the Bayes optimal classifier to be linear is:

- □ The data generating distributions for both classes are equivalent except for the mean
- □ The data generating distributions for both classes are Gaussian
- □ The data generating distributions for both classes have the same covariance
- ☑ None of the above

(c) (3 Punkte) The Fisher linear discriminant finds the projection that:

- □ Maximizes the margin between the two data generating distribution
- □ Maximizes the margin between the mean of the two data generating distribution
- □ Maximizes the ratio between the within-class variance and the between-class variance
- ☑ Minimizes the ratio between the within-class variance and the between-class variance

(d) (3 Punkte) Which is **false**? Let k be a Gaussian Kernel. A Gram Matrix K associated to this Kernel always satisfies:

- □ $K = K^T$
- □ $KK^T = I$
- □ All Eigenvalues are non-negative
- ☑ $\forall u \in R^N : u^T K u \geq 0$

(e) (3 Punkte) Error Backpropagation is a technique to:

- ☑ Efficiently compute the error gradient in a multilayer neural network.
- □ Efficiently compute the error gradient in restricted Boltzmann machine.
- □ Efficiently compute the prediction error of a multilayer neural network.
- □ Efficiently compute the prediction error of a restricted Boltzmann machine.

What …. is not a discriminant function:
a) P(w_c|x)
b)P(x|w_c)*P(w_c)
c)P(w_c|x)*P(w_c)^-1
[✓ d)P(x|w_c)^2*P(w_c)^2]

What is likely an overfitted estimator?
a) High bias model
[✓ b) High variance model]
c) Low bias model
d) Low variance model

What does Fisher discriminant optimize?
a)..
b)..
[✓ c) Maximize ratio Within class variance to between class variance]
d) Minimize ratio Within class variance to between class variance

What does the constant C stand for in SWM?
[✓ a) ability of the decision boundary to be out of the margin]
b) number of points not being classified correctly
c)
d)

(a) The Bayes error is

□ the lowest error of a linear classifier.

□ the expected error of a random linear classifier.

■ the error of any nonlinear classifier.

□ the error of a naive BAYES classifier .

(b) The Fisher linear discriminant find the projection $y = w^{\mathsf{T}}x$ of the data that maximises

□ the margin between the two data generating distributions.

□ the within-class variance divided by the between-class variance.

□ the margin between the means of the data generating distributions.

■ the between-class variance divided by the within-class variance.

(c) A biased estimator is used to

■ make the estimator less affected by the sampling of the data.

□ make the estimation procedure more sensitive to the sample data.

□ reduce the risk of underfitting the data.

□ None of the above, an unbiased estimator is always better.

(d) Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be unlabelled observations. Consider a GAUSSIAN kernel and its GRAM matrix $K \in \mathbb{R}^{N \times N}$. Which is always true?

□ $K^T K = I$.

□ $KK^T = I$.

■ $\forall u \in \mathbb{R}^N \; uKu \geq 0$.

□ $\forall u \in \mathbb{R}^N \; uKu \leq 0$.

1. Given two normal distributions $p(x|w_1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $p(x|w_2) \sim \mathcal{N}(\mu_2, \Sigma_2)$ what is a *necessary* and *sufficient* condition for the optimal decision boundary to be linear? (5pts)

   [✓ (a) $\Sigma_1 = \Sigma_2$]

   (b) $\Sigma_1 = \Sigma_2, P(w_1) = P(w_2)$

   (c) ...

   (d) ...

2. We have a classifier that decides the class $\text{argmax}_{w_i} f_i(x)$ for the input x. What is a suitable discriminant functions f_i? (5pts)

   (a) $\sqrt{p(x|w_i)P(w_i)}$

   [✓ (b) $\log(p(x|w_i) + P(w_i))$]

   (c) ...

   (d) ...

3. K-means is (5pts)

   [✓ (a) a non-convex algorithm used to cluster data]

   (b) a kernelized version of the means algorithm

   (c) ...

   (d) ...

4. Error backpropagation gives (5pts)

   [✓ (a) the gradient of the error function]

   (b) the optimal direction in parameter space

   (c) ...

   (d)

Which of the following is false: Assume a boosted classifier consists of weak hypotheses (aka. weak classifiers) that are each of them implemented by a threshold neuron. In that case the boosted classifier:

□ can be viewed as a two-layer neural network.

☑ can be trained by error backpropagation instead of AdaBoost.

□ can represent nonlinear decision boundaries.

□ can represent non-smooth decision boundaries.

Which of the following is true: A Product of Experts:

□ is an extension of a mixture model where each mixture element is allowed to be non-Gaussian.

□ is an extension of a mixture model where each mixture element can be Gaussian with non-isotropic covariance.

☑ allows to learn more global features compared to a mixture model.

□ allows to learn more local features compared to a mixture model.

Which of the following is false: Gaussian kernel ridge regression:

□ is an extension of ridge regression to non-linear models.

□ admits a closed-form solution when minimized for least squares.

□ learns smooth non-linear functions.

☑ assumes that the input data is drawn from a Gaussian distribution.

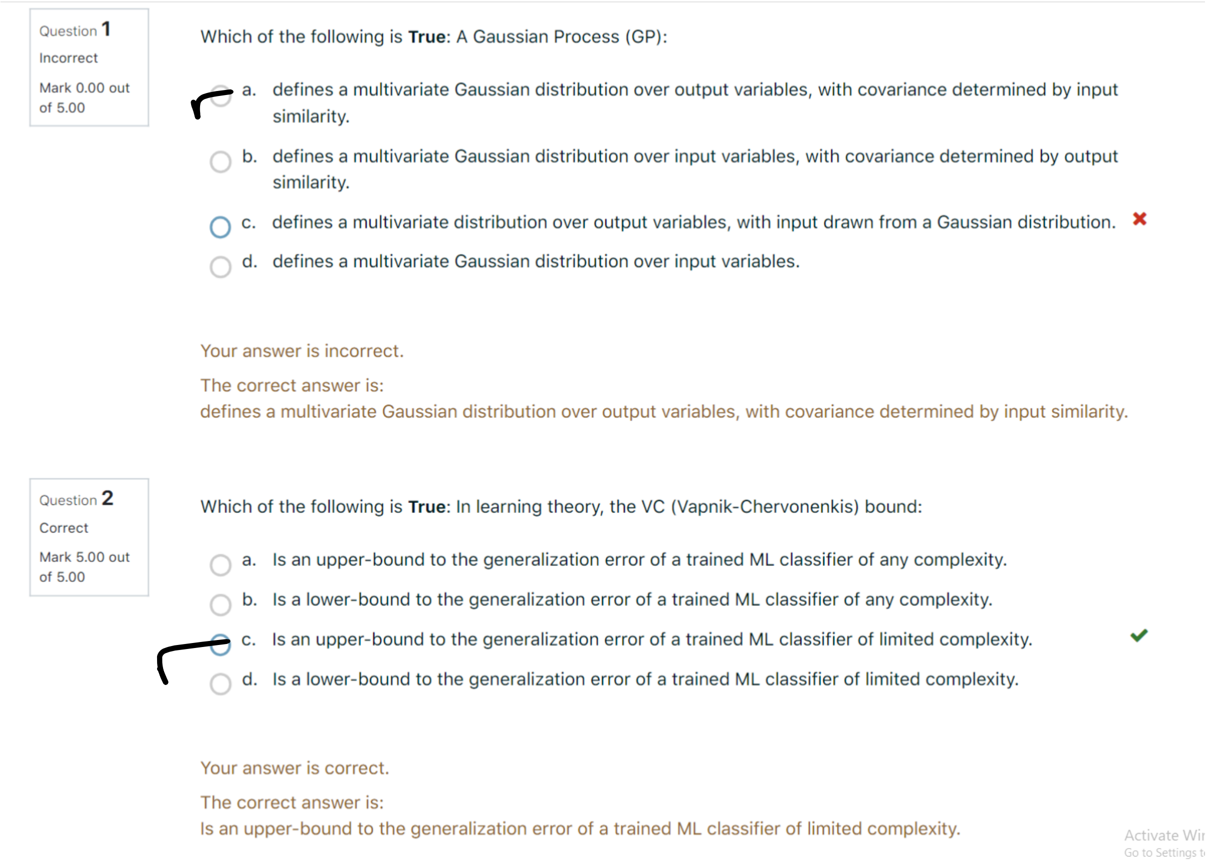
Which of the following is true: In learning theory, the VC (Vapnik-Chervonenkis) bound:

□ is an upper bound to the generalization error of a trained ML classifier of any complexity.

□ is a lower bound to the generalization error of a trained ML classifier of any complexity.

☑ is an upper bound to the generalization error of a trained ML classifier of limited complexity.

□ is a lower bound to the generalization error of a trained ML classifier of limited complexity.



Question **1**

Incorrect

Mark 0.00 out of 5.00

Which of the following is **True**: A Gaussian Process (GP):

○ a. defines a multivariate Gaussian distribution over output variables, with covariance determined by input similarity.

○ b. defines a multivariate Gaussian distribution over input variables, with covariance determined by output similarity.

● c. defines a multivariate distribution over output variables, with input drawn from a Gaussian distribution. ✗

○ d. defines a multivariate Gaussian distribution over input variables.

Your answer is incorrect.

The correct answer is:
defines a multivariate Gaussian distribution over output variables, with covariance determined by input similarity.

Question **2**

Correct

Mark 5.00 out of 5.00

Which of the following is **True**: In learning theory, the VC (Vapnik-Chervonenkis) bound:

○ a. Is an upper-bound to the generalization error of a trained ML classifier of any complexity.

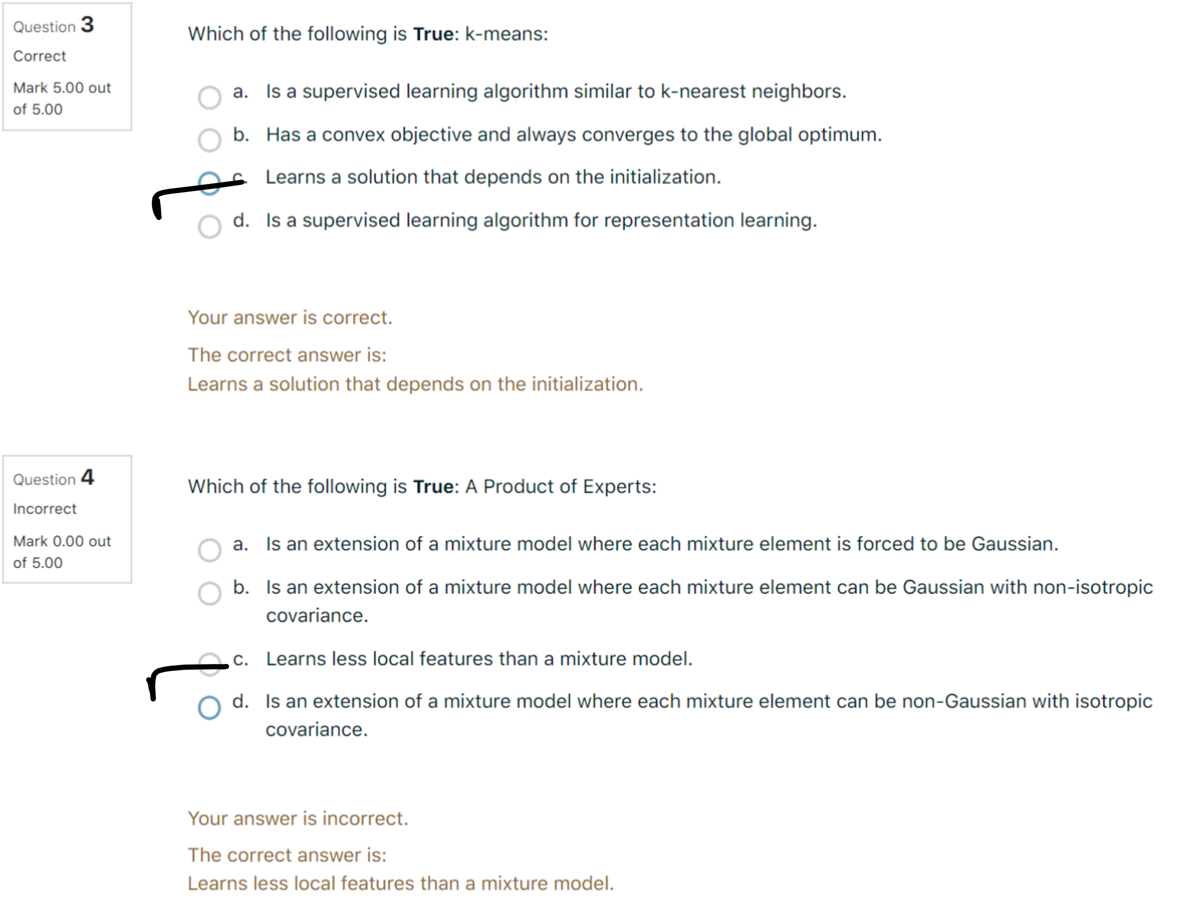
○ b. Is a lower-bound to the generalization error of a trained ML classifier of any complexity.

● c. Is an upper-bound to the generalization error of a trained ML classifier of limited complexity. ✓

○ d. Is a lower-bound to the generalization error of a trained ML classifier of limited complexity.

Your answer is correct.

The correct answer is:
Is an upper-bound to the generalization error of a trained ML classifier of limited complexity.



Question **3**

Correct

Mark 5.00 out of 5.00

Which of the following is **True**: k-means:

○ a. Is a supervised learning algorithm similar to k-nearest neighbors.

○ b. Has a convex objective and always converges to the global optimum.

● c. Learns a solution that depends on the initialization.

○ d. Is a supervised learning algorithm for representation learning.

Your answer is correct.

The correct answer is:
Learns a solution that depends on the initialization.

Question **4**

Incorrect

Mark 0.00 out of 5.00

Which of the following is **True**: A Product of Experts:

○ a. Is an extension of a mixture model where each mixture element is forced to be Gaussian.

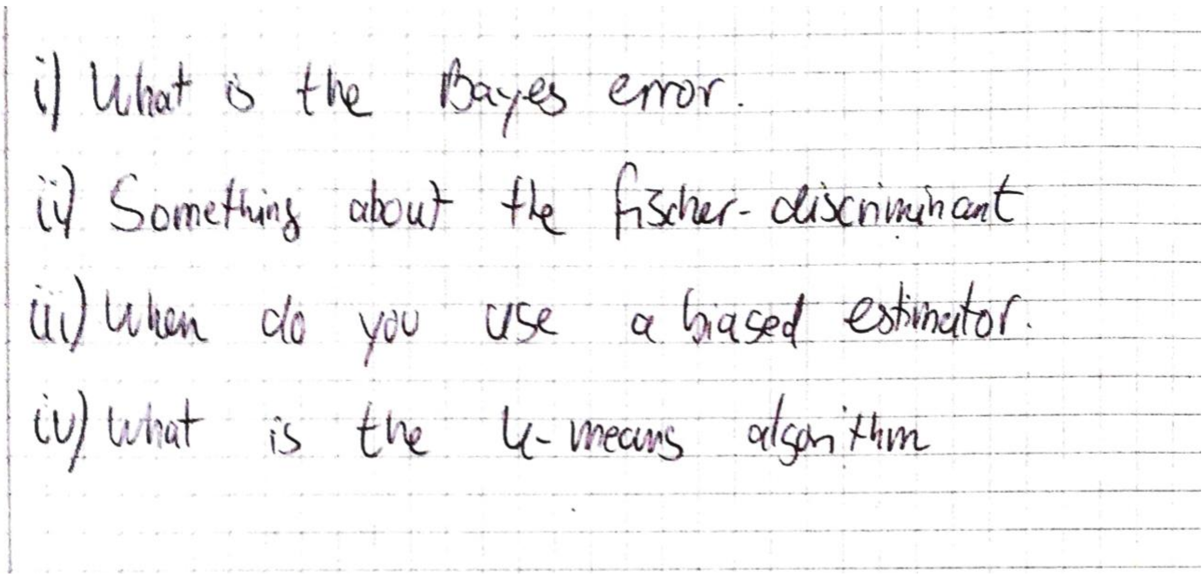
○ b. Is an extension of a mixture model where each mixture element can be Gaussian with non-isotropic covariance.

○ c. Learns less local features than a mixture model.

● d. Is an extension of a mixture model where each mixture element can be non-Gaussian with isotropic covariance.

Your answer is incorrect.

The correct answer is:
Learns less local features than a mixture model.

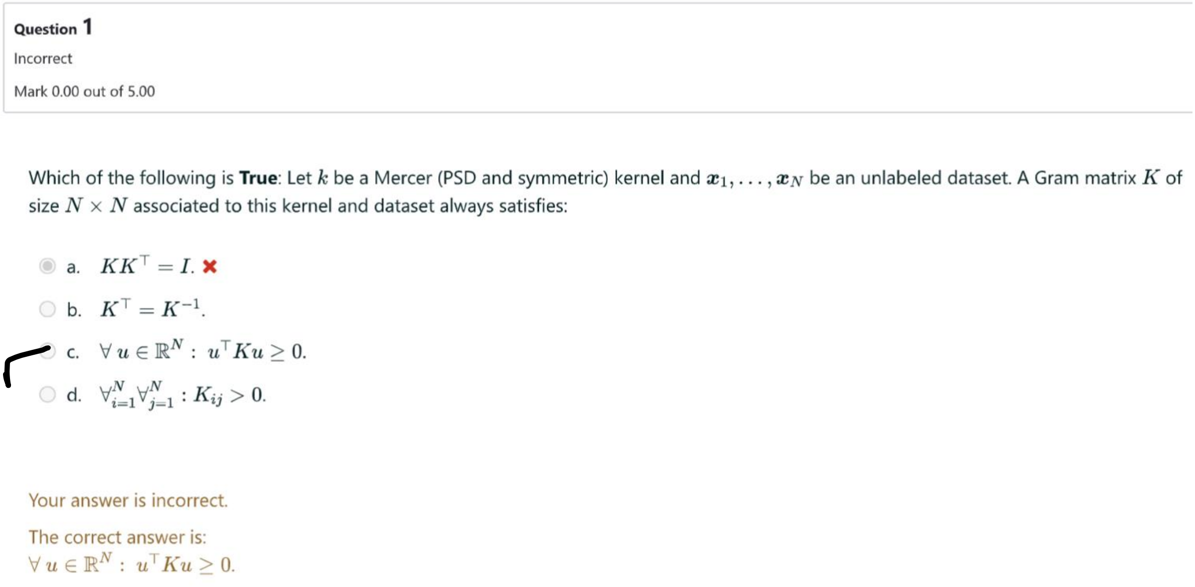


i) What is the Bayes error.

ii) Something about the fischer-discriminant

iii) When do you use a biased estimator.

iv) what is the k-means algorithm



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. $KK^\top = I$. ✖

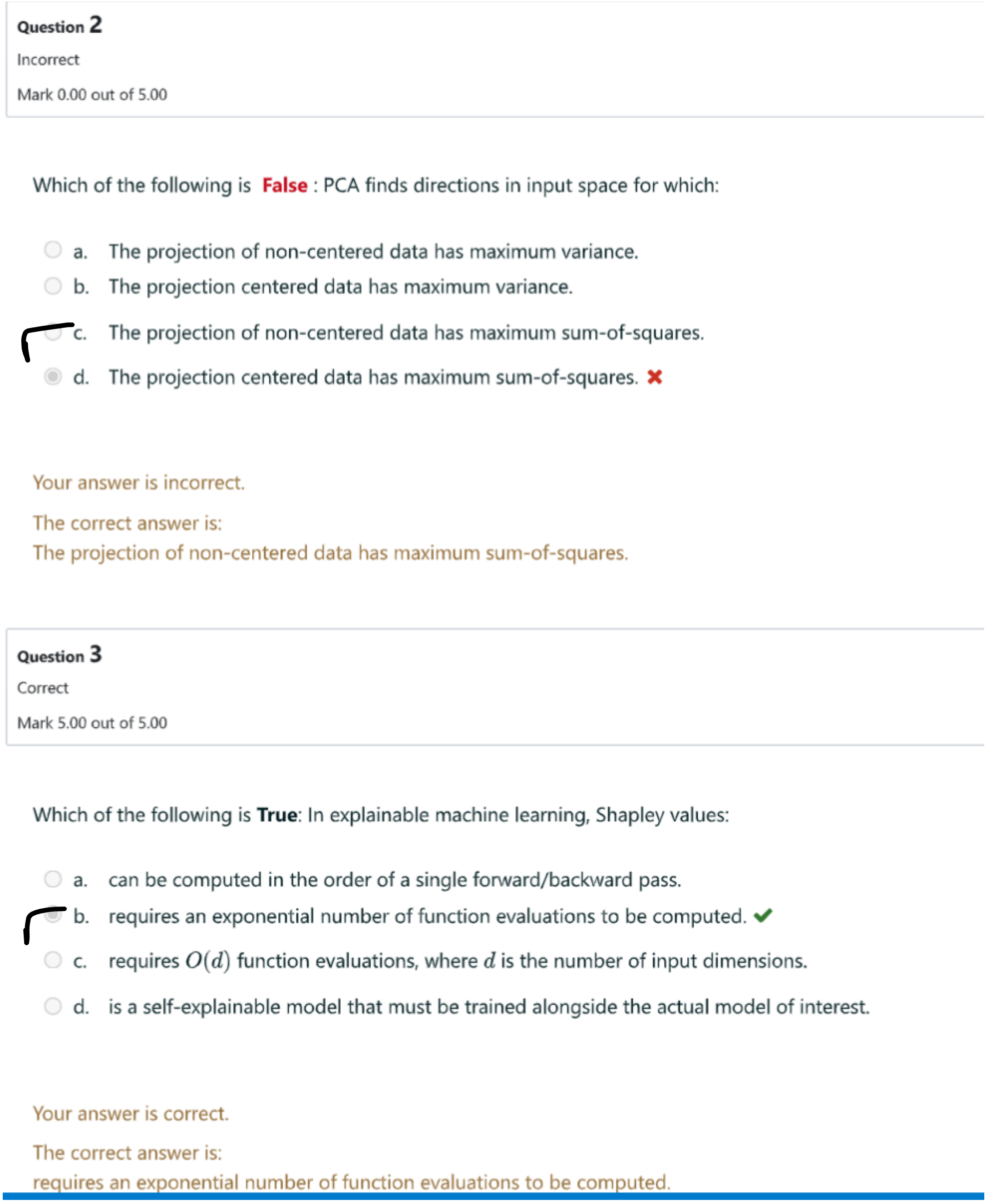
○ b. $K^\top = K^{-1}$.

[✓ ○ 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.

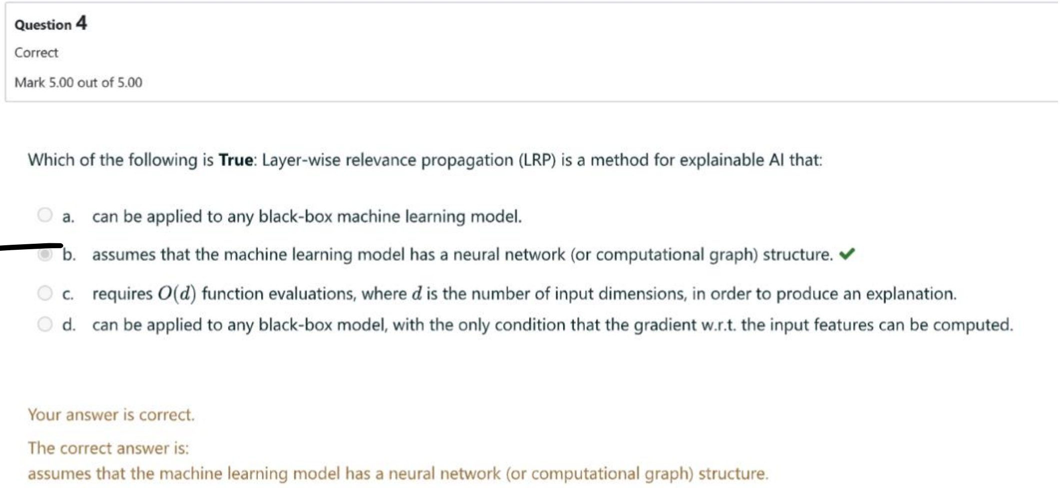
● b. requires an exponential number of function evaluations to be computed. ✓

○ 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.

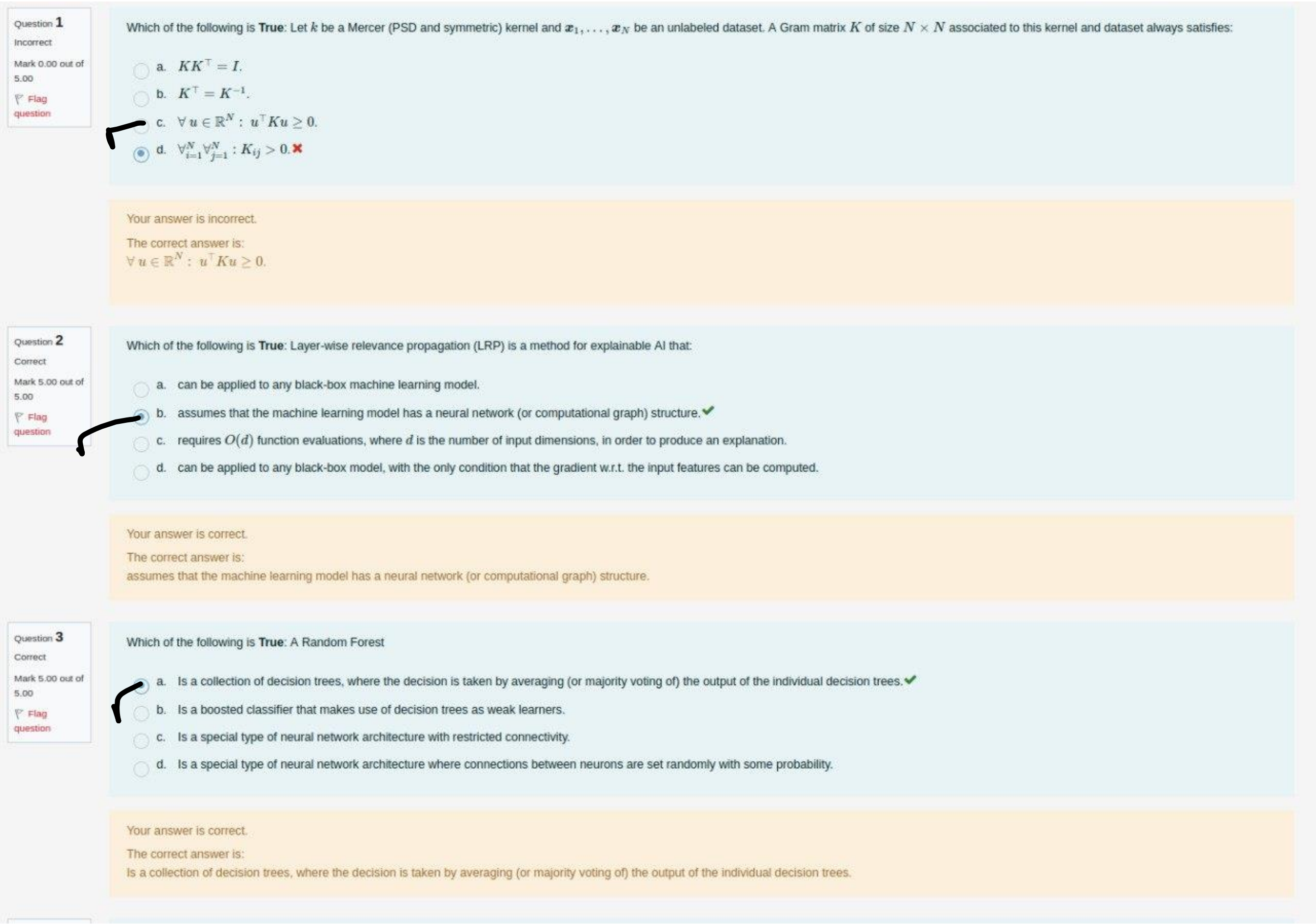
● b. assumes that the machine learning model has a neural network (or computational graph) structure. ✓

○ 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.



Question **1**

Incorrect

Mark 0.00 out of 5.00

Flag question

Which of the following is **True**: Let k be a Mercer (PSD and symmetric) kernel and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ 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. $\forall\, u \in \mathbb{R}^N:\ u^\top K u \geq 0.$

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

Your answer is incorrect.

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

Question **2**

Correct

Mark 5.00 out of 5.00

Flag question

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.

Question **3**

Correct

Mark 5.00 out of 5.00

Flag question

Which of the following is **True**: A Random Forest

● a. Is a collection of decision trees, where the decision is taken by averaging (or majority voting of) the output of the individual decision trees. ✓

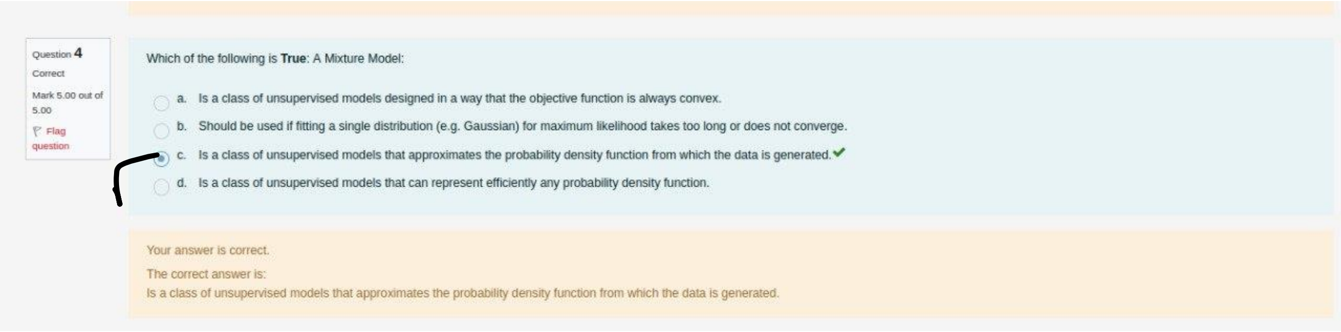
○ b. Is a boosted classifier that makes use of decision trees as weak learners.

○ c. Is a special type of neural network architecture with restricted connectivity.

○ d. Is a special type of neural network architecture where connections between neurons are set randomly with some probability.

Your answer is correct.

The correct answer is:
Is a collection of decision trees, where the decision is taken by averaging (or majority voting of) the output of the individual decision trees.



Question **4**

Correct

Mark 5.00 out of 5.00

Flag question

Which of the following is **True**: A Mixture Model:

○ a. Is a class of unsupervised models designed in a way that the objective function is always convex.

○ b. Should be used if fitting a single distribution (e.g. Gaussian) for maximum likelihood takes too long or does not converge.

● c. Is a class of unsupervised models that approximates the probability density function from which the data is generated. ✓

○ d. Is a class of unsupervised models that can represent efficiently any probability density function.

Your answer is correct.

The correct answer is:
Is a class of unsupervised models that approximates the probability density function from which the data is generated.