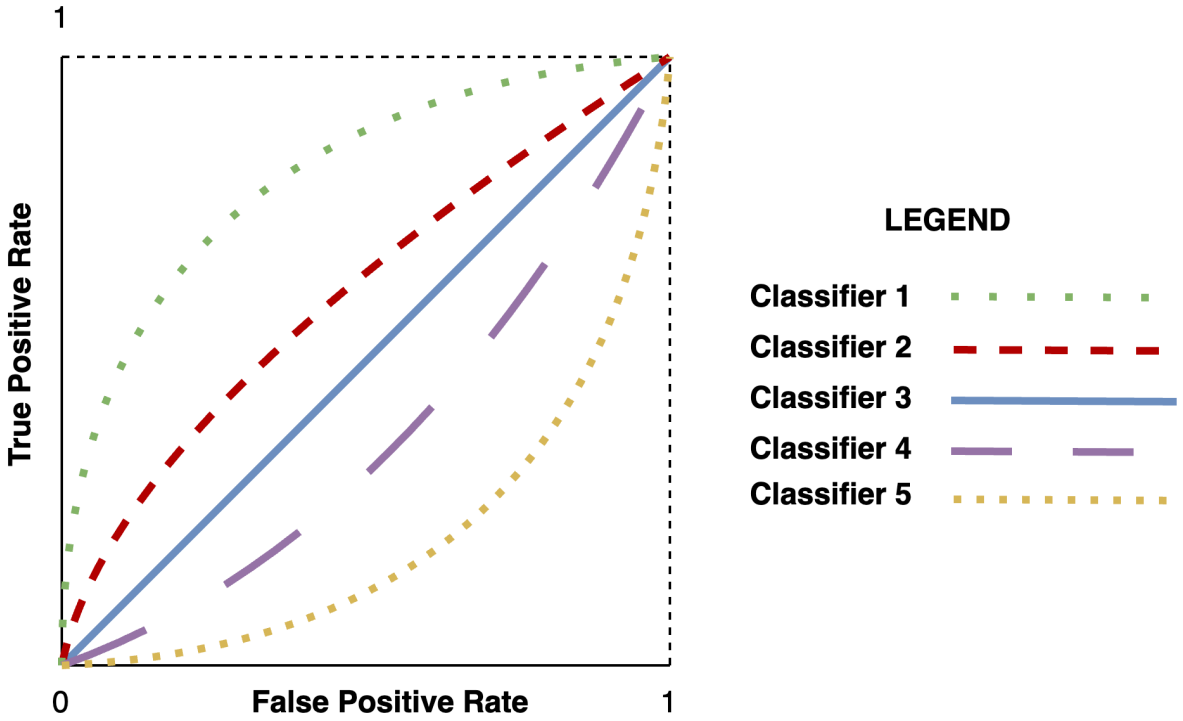
DM581 Introduction to Machine Learning - Exam January 2024

The exam contains 13 questions. All questions ask for an evaluation of five statements with (yes/no) or (true/false) answers. A true answer adds +1 points and a false answer adds -1 points to your exam score. Hence, skipping to evaluate a statement is very likely to be more advantageous for you than making a random guess. Your final score is calculated as (Your point sum)/(13*5)*100. The exam duration is four hours.

$P(\cdot)$ denotes a probability measure.

No assumptions may be made additional to those specified in the question text.

The Receiver Operating Characteristics (ROC) curves for five binary classifiers are given in the figure below.



Evaluate the below statements regarding these five classifiers according to the definitions below:

- outperform: A classifier "outperforms" another classifier if its ROC curve is more favorable.
- random classifier: A classifier is a random classifier if it predicts the output labels following a probability distribution with maximum entropy possible for the sample space in question.
- negating the predictions: If the prediction function of a classifier assigns an input to class 1, flipping it to 0 and vice versa.
- perfect classifier: A classifier is a perfect classifier if it gives 100% accuracy.

Select the correct answers

	True	False
Classifier 2 outperforms a random classifier.	○	○
The area under the ROC curve of a perfect classifier depends on the ratio of the positive and negative labels in the data set used to draw the ROC curve.	○	○
Classifier 5 outperforms Classifier 4 if the predictions of both classifiers are negated.	○	○
A random classifier outperforms Classifier 4 if the predictions of both classifiers are negated.	○	○

Classifier 1 outperforms Classifier 2 if the predictions of both classifiers are negated.	○	○

Consider a neural network that has the following architecture :

- A fully connected layer that linearly maps a scalar input x to a hidden layer $(z_1, \ldots, z_K)$ of K dimensions:
  $z_j = w_j x + b_j, \qquad j \in \{1, \ldots, K\}$ where w_j is the weight and b_j is the bias for the hidden unit j
- Step activation function applied to the linear activation z_j of each hidden unit:
  $h_j = \begin{cases} 1, & z_j > 0 \\ 0, & \text{otherwise} \end{cases}$
- A fully connected layer that maps the nonlinear activations $(h_1, \ldots, h_K)$ to a scalar output y

Can this neural network exactly represent the following function types? In other words, is it possible to express the function types below as a special case of the neural network by choosing its parameters appropriately?

Select the correct answers

	Yes	No
Polynomials of degree one: $y = ax + b$	○	○
Polynomials of degree two: $y = ax^2 + bx + c$	○	○
Piecewise constant functions: $y = \begin{cases} c_1, & \text{if } x < a_1 \\ c_2, & \text{if } a_1 \le x < a_2 \\ & \vdots \\ c_K, & \text{if } x \ge a_K \end{cases}$ for constants $c_1, \ldots, c_K$ and thresholds $a_1, \ldots, a_K$	○	○
Hinge loss: $y = \max(1 - x, 0)$	○	○
The logistic sigmoid function: $y = \frac{1}{1+e^{-x}}$	○	○

You are given a k-Nearest-Neighbor (kNN) classifier evaluated on a particular training and test split of a particular data set for a particular choice of k. Which of the following actions is guaranteed to reduce the variance of a k-Nearest-Neighbor (kNN) classifier?

Select the correct answers

	True	False
Allocating a larger portion of the data set for training	○	○
Building a majority voting ensemble of three kNN classifiers	○	○
Increasing k	○	○
Normalizing the input features	○	○
Enlarging the test split by collecting new data	○	○

Evaluate the following statements about the k-means clustering algorithm.

Select the correct answers

	True	False
Two different initializations of k-means always converge to the same partitioning if the cluster count k is greater than two.	○	○
k-means always finds an optimal partitioning after a finite number of iterations according to an objective optimality criterion.	○	○
For a fixed input dimensionality, the size of the memory k-means requires during the computation of one iteration is smaller than the memory the Expectation Maximization (EM) algorithm requires during the computation of one iteration.	○	○
k-means always converges to a partitioning after a finite number of iterations.	○	○
For a fixed input dimensionality and fixed hardware configuration, the computation time k-means requires for completing one iteration is less than the computation time the Expectation Maximization (EM) algorithm requires for completing one iteration.	○	○

```
1: for epoch in range(100):
2:     for inputs_batch, outputs_batch in train_dl:
3:         predictions_batch = model(inputs_batch.view(-1, 784))
4:         loss = loss_function(predictions_batch, outputs_batch)
5:         loss.backward()
6:         optimizer.step()
7:         optimizer.zero_grad()
```

Evaluate the following statements related to the Python source code given above which uses the PyTorch library.

Select the correct answers

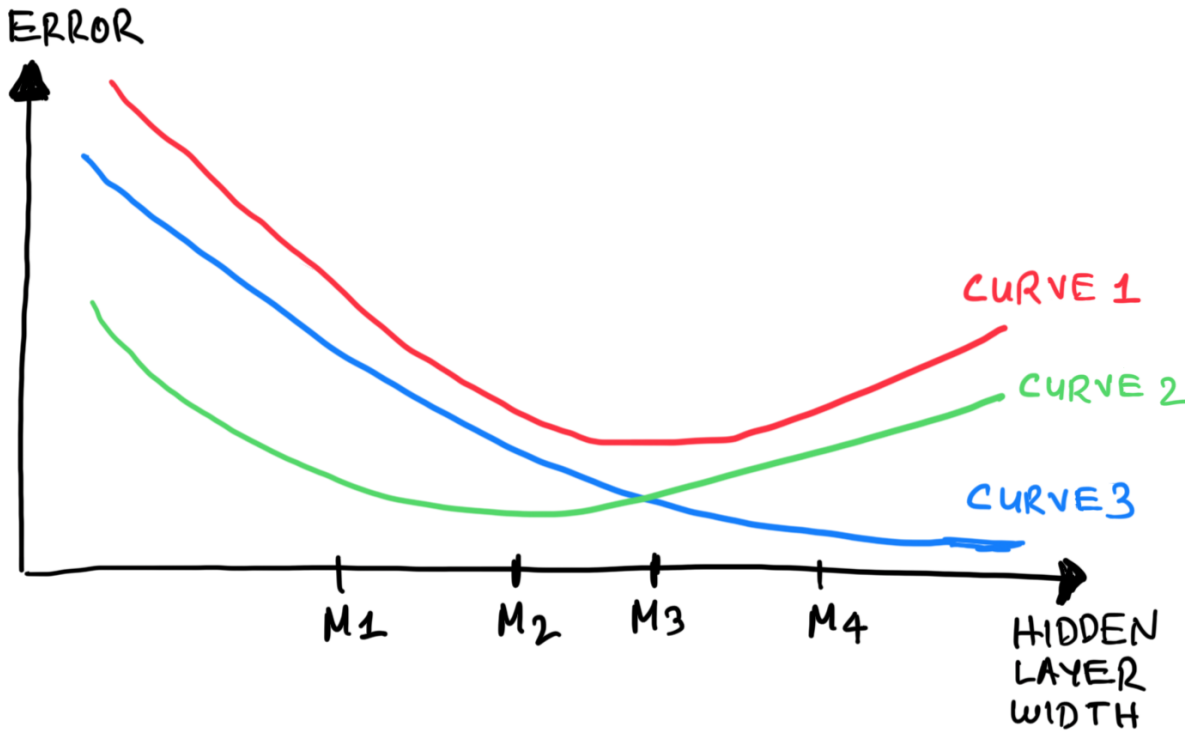
	True	False
The code implements the K-fold cross-validation stage of a supervised learning task.	○	○
Line 5 updates the model parameters.	○	○
The code implements the training stage of a supervised learning task.	○	○
The code performs iterative optimization using stochastic gradient descent.	○	○
Line 7 computes the gradients of the loss function for the drawn minibatch.	○	○

```
1: model = nn.Sequential(
2:     nn.Linear(10, 100),
3:     nn.ReLU(),
4:     nn.Linear(100, 250),
5:     nn.ReLU(),
6:     nn.Linear(250, 125),
7:     nn.ReLU(),
8:     nn.Linear(125, 2)
9:   )
```

Evaluate the following statements about the PyTorch source code given above.

Select the correct answers

	True	False
The neural network has more than 50000 parameters.	○	○
The neural network has three hidden layers.	○	○
The code will run without a dimensionality mismatch error only for 100-dimensional input observations.	○	○
The code will run without a dimensionality mismatch error only on minibatches of 100 observations.	○	○
Line 6 should be replaced by nn.Linear(125, 250) for the code to run without a dimensionality mismatch error.	○	○



Two of the three curves shown above depict how the training and generalization error of a neural network with two fully connected layers with equal width changes on a prediction task as hidden layers get wider (as the number of hidden units at each layer increases) according to the assumptions of the statistical learning theory. The remaining curve does not have a meaning and needs to be discarded. Evaluate the following statements about this figure according to the statistical learning theory.

Select the correct answers

	True	False
Curve 3 (blue) depicts the training error.	○	○
Choosing the hidden layer width smaller than M3 causes underfitting.	○	○
The model with optimal generalization performance is M2.	○	○
Model M4 underfits more than model M1.	○	○
Curve 2 (green) depicts the generalization error.	○	○

Evaluate the following statements about the toss of a fair coin three times.

Select the correct answers

	True	False
Observing an even number of heads and observing exactly two heads are equally probable events.	○	○
Observing an odd number of heads and observing less than two heads are independent events.	○	○
The probability of three heads is 1/8	○	○
The probability of observing exactly one heads is 3/8	○	○
Given that you have observed at least one heads, the probability that you observe at least two heads is smaller than 1/2	○	○

Two independent random variables A and B are known to have the following properties:

$$\begin{aligned}\mathbb{E}[A] &= \alpha,\\ Var(A) &= \beta,\\ \mathbb{E}[B] &= \gamma,\\ Var(B) &= \phi,\end{aligned}$$

where $\mathbb{E}[\cdot]$ denotes the expectation and $Var(\cdot)$ denotes the variance of a random variable. $Cov(\cdot, \cdot)$ denotes the covariance of a pair of random variables.

Evaluate the following statements about these two random variables.

Select the correct answers

	True	False
$\mathbb{E}[A + cB] = \alpha + c\gamma$	○	○
$\mathbb{E}[A^2] = \alpha^2 + \beta$	○	○
$Var(cA) = c\beta$ for a positive constant c	○	○
$Cov(A, B)$ cannot be inferred from the information given in the question text.	○	○
$Var(AB) \geq \beta\phi$	○	○

Consider the confusion matrix below obtained from a classifier evaluated on the test split of a data set comprising data points with discrete labels that can take three possible values: A, B, C. Each of these values correspond to a class.

		Predicted			
		A	**B**	**C**	**Total**
Actual	**A**	15	2	3	20
	B	7	15	8	30
	C	2	3	45	50
	Total	24	20	56	100

Evaluate the below statements regarding this confusion matrix.

Select the correct answers

	True	False
The accuracy of a classifier that always predicts Class C on this test split is higher than the expected accuracy of a random classifier that predicts each of the three classes with equal probability.	○	○
The precision for Class B is 0.75.	○	○
The precision for Class A is greater than the recall for Class A.	○	○
The random classifier that predicts each of the three classes with equal probability has an expected accuracy of 50% on this test split.	○	○
False positive rate for the Class B is smaller than the precision	○	○

Consider four independent binary random variables X_1, X_2, X_3, X_4 that follow a Bernoulli distribution $Ber(X|\pi) = \pi^X(1-\pi)^{1-X}$ for identical π. An experiment made on these random variables yield the following data set

$$D = \{X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1\}$$

Evaluate the following statements regarding random variables X_1, X_2, X_3, X_4 and the data set D.

Select the correct answers

	True	False
Performing Bayesian inference on π with respect to data set D using a prior distribution $Beta(\pi\|\alpha, \beta)$ yields the posterior distribution $Beta(\pi\|\alpha + 3, \beta + 1)$.	○	○
The maximum likelihood estimate for π is 0.75.	○	○
The maximum likelihood estimate for π cannot be guaranteed to converge to the true π as the size of the data set D grows without making additional assumptions about the model.	○	○
If $\pi = 0.3$ then $P(X_1 + X_2 + X_3 + X_4 < 3) \geq 0.67$	○	○
$P(X_1 = A, X_2 = B\|\pi) = P(X_1 = A\|\pi)P(X_2 = B\|\pi)$ for any binary pair of values A, B.	○	○

Consider a sequence of random variables X_1, X_2, X_3, X_4 that satisfy the Markov property, where the subscripts denote the order of the occurrence of events in time. Evaluate the following statements about these random variables.

Select the correct answers

	True	False
X_2 andX_4 do not have to be independent random variables if X_1 is observed.	○	○
X_1 and X_3 have to be independent random variables if X_2 is observed.	○	○
X_1 and X_4 have to be independent random variables if X_3 is observed.	○	○
X_2 and X_3 have to be independent random variables.	○	○
X_3 and X_4 have to be independent random variables if X_2 is observed.	○	○

For the three events A, B, and C we know that

- A and C are independent
- B and C are independent
- A and B are disjoint
- $P(A \cup C) = \frac{2}{3}$
- $P(B \cup C) = \frac{3}{4}$
- $P(A \cup B \cup C) = \frac{11}{12}$

Evaluate the statements given below about these three events.

Select the correct answers

	True	False
$P(C) = P(B)$	○	○
$P(A) < P(B)$	○	○
$P(A \cap C) = \emptyset$	○	○
$P(A) + P(B) > P(A \cup B)$	○	○
$P(B) = \frac{1}{2}$	○	○