AI512 Exam January 2025

The exam contains 12 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)/(12*5)*100. The exam duration is four hours.

denotes a probability measure. denotes expectation. denotes variance.

The symbol denotes matrix or column vector transpose.

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

Page 1

Suppose A and B are events with $0 < P(A) < 1$ and $0 < P(B) < 1$. Evaluate the statements regarding these events.

	True	False
There exists A and B that satisfy $P(A \cup B) = 1$.	○	○
If A and B are independent, they cannot be disjoint.	○	○
If A and B are disjoint, they cannot be independent.	○	○
If A and B are independent, then it is possible that A and $A \cup B$ are independent.	○	○
If $A \subset B$, then it is possible that A and B are independent.	○	○

Evaluate the following statements about the $k-$nearest neighbor classifier. Assume that all the conditions other than the ones specified in the statements are equal. k always denotes only the number of nearest neighbors used for classification.

	True	False
The model has higher overfitting risk if training set size shrinks.		
The underfitting risk of the model increases if k grows.		
The training error is guaranteed to be zero for all $k > 0$.		
Model bias increases if the training set size grows.		
Model variance increases if k grows.		

Consider a coin with a heads probability of $2/3$ being tossed three times. Denote by the random variable C the number of times the coin toss results in a heads. Evaluate the following statements related to the random variable C.

	True	False
$\mathbb{E}[C^2] > \text{Var}[C]$		
$P(C = 1) > P(C = 3)$		
$P(C = 0) < P(C = 3)$		
$\mathbb{E}[C^2] > \mathbb{E}[C]^2$		
$\mathbb{E}[C] > 2$		

Consider the supervised learning problem of predicting the real-valued scalar y from a feature vector x to be solved by fitting a ridge regression predictor

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \alpha||w||_2^2$$

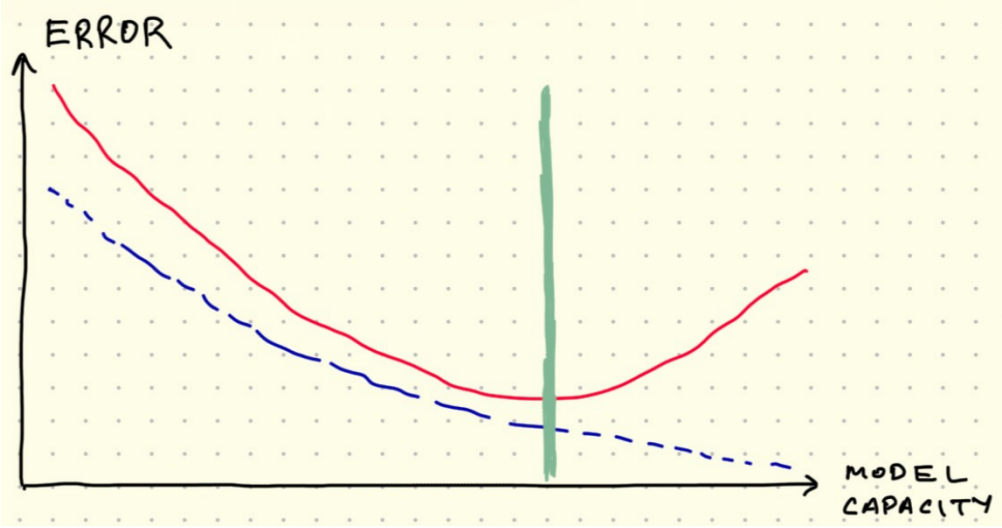
to a data set $D = \{(x_i, y_i) : i = 1, \ldots, n\}$ where w is a parameter vector with appropriate size and a positive constant α. Evaluate the statements below.
Remarks: Model variance is defined with respect to the experiment repetitions. It is the variance that is used in the context of the bias-variance dilemma.

	True	False
The underfitting risk of the model reduces as α decreases.		
Model capacity increases as α increases.		
Model bias increases as n increases.		
Model variance increases as α increases.		
The overfitting risk of the model reduces as α increases.		

```
model = nn.Sequential(
  nn.Linear(500, 100),
  nn.ReLU(),
  nn.Linear(500, 5),
  nn.Softmax()
)
```

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

	True	False
The model has more than 100000 parameters.		
The code builds a neural network with a single hidden layer containing 100 neurons.		
The model is designed to solve a classification task with 5 classes.		
The model uses the sigmoid activation function between the hidden layer and the output layer.		
The model assumes that the input data points have 100 features.		



	True	False
The optimal model capacity is marked by the green line.		
The dashed blue line shows the model prediction error on the training split.		
Doubling the size of the training data without changing the size of the test data would shift the vertical green line to the left.		
The right-hand side of the solid vertical green line indicates the underfitting region.		
The solid red line shows the model prediction error on the test split.		

A train travels regularly from Paris to Berlin. The train takes the northern route passing through Brussels in 60 per cent of the cases and it takes the southern route passing through Strasbourg otherwise. The train arrives Berlin later than the planned schedule in 30 per cent of the cases when it takes the northern route and in 10 per cent of the cases when it takes the southern route. Evaluate the following statements about this train.

	True	False
Observing that the train is late increases the probability that it took the northern route.		
If the train chose the northern and southern route with equal percentages, then the probability that the train took the northern route given that it arrived Berlin late would be greater than the probability that the train took the southern route given that it arrived Berlin late.		
Probability that the train will arrive Berlin late is greater than the probability that it will arrive Berlin late given that it took the northern route.		
If the train is observed to arrive Berlin late, then the probability that it took the northern route is smaller than the probability that it took the southern route.		
Observing that the train is late increases the probability that it took the southern route.		

```
for epoch in range(20):
    for inputs_batch, outputs_batch in train_dl:
        predictions_batch = model(inputs_batch.view(-1, 5))
        loss = ((predictions_batch - outputs_batch) ** 2).mean()
        loss.backward()
        optimizer.step()
        optimizer.zero_grad()
```

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

	True	False
The command `loss.backward()` updates the learnable parameters of the model.		
The code is usable for the training stage of a regression task.		
The code uses each data point within the set `train_dl` exactly 20 times for training		
The command `optimizer.step()` updates the gradients of the model parameters.		
The input dimensionality is five.		

Consider a predictor that maps a $d-$dimensional input vector x to an output prediction with

$$f(x) = W_2^\top \max(W_1^\top x, 0)$$

for $W_1 \in \mathbb{R}^{d \times h}$ and $W_2 \in \mathbb{R}^{h \times 1}$ and the $\max(\cdot)$ operator applies element-wise to the vector on its argument.

	True	False
The predictor corresponds to a neural network that uses a rectified linear unit as its activation function.		
The predictor can reach zero training error only if the input-output relationship is perfectly linear, i.e. there exists a vector $\beta \in \mathbb{R}^d$ such that $y_i = \beta^\top x_i$ for all training inputs x_i and the corresponding output labels y_i.		
The predictor corresponds to a neural network with a single hidden layer of size h.		
The gradient descent algorithm is guaranteed to converge to the global minimum of the loss function $(y - f(x))^2$ if the learning rate is chosen small enough.		
The predictor cannot map an input outside the $[0, 1]$ interval no matter which values W_1 and W_2 take.		

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 corresponds to a class.

		Prediction			
		A	B	C	
Actual	A	15	5	5	25
	B	10	20	5	35
	C	5	10	25	40
		30	35	35	100

Evaluate the below statements regarding this confusion matrix.

	True	False
The precision for Class C is greater than the recall of Class C.		
The accuracy of the classifier is smaller than 0.5.		
The recall for Class A is greater than the recall for Class B.		
The precision for Class B is equal to the recall of Class B.		
The precision for Class A is greater than the recall of Class A.		

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

	True	False
The algorithm cannot output spiral-shaped clusters when $k > 1$ and the algorithm is run until convergence starting from a random initialization.		
The algorithm can never converge to a clustering before changing the assignment of at least one data point at least once after initialization.		
The algorithm always converges to the same cluster centroids regardless of its initialization.		
The algorithm assigns each data point to a single cluster.		
The computational complexity of the algorithm is not dependent on the data dimensionality.		

Evaluate the statements below regarding a data set $D = \{X_i : i = 1, \ldots, n\}$ that contains n observations X_i that come independently and identically distributed from a Bernoulli distribution with parameter π.

	True	False
$\text{Var}[X_i] > 1$.		
The maximum likelihood estimate for π is $\frac{1}{n}\sum_{i=1}^{n} X_i$.		
$P(X_i = 1 \cup X_j = 0) > P(X_i = 1) + P(X_j = 0)$.		
$\mathbb{E}[X_i] = \mathbb{E}[X_j]$ for all $i \neq j$.		
If $D = \{0, 1, 1, 0, 1\}$, then the likelihood of π is $\pi^3(1-\pi)^2$.		