URL
Spring Term 2024

Unsupervised Learning Final Exam

Duration: 2.5 hours

14/06/2024
Artificial Intelligence Master

Name: ____

Instructions

This questionnaire is the only thing that you have to deliver at the end of the exam. Put a circle around the correct answer, if you want to change your answer use a cross to invalidate it.

Begin Quiz

This exam is about the different topics of the course. Each question has a value of 1 point, each question has **exactly two correct answers**, each correct answer has a value of 0.5 points, each incorrect answer discounts 0.5 points of the question.

1. Thinking about data preprocessing (multiple answer):

(a) Data normalisation makes the attributes of the dataset to have a different weight depending on their scale of measure.

(b) Power transformations are parametric non linear transformations for normalising data that are applied to obtain a data distribution closer to the uniform distribution.

● (c) The Box-Cox transformation is a parametrical non linear transformation used to transform positive data to a distribution closer to the gaussian distribution.

● (d) Range normalisation is a transformation that makes all the attributes to have the same scale.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because Data normalisation tries to avoid the effect of the scales of measure of the attributes so they have a priori the same importance.

The answer (b) is not correct because Power transformations are used to transform the data to a distribution closer to the gaussian.

■

2. Thinking about data preprocessing (multiple answer):

● (a) Discretization by equal frequency bins consists on choosing a set of intervals that contain the same number of examples.

(b) The quantile transformation is a parametric normalisation method that maps the distribution of the data to the uniform or normal distribution using the empirical quantiles of the data.

● (c) The Yeo-Johnson transformation is a parametrical non linear transformation for normalising data that can be applied to positive and negative data.

(d) Distribution normalisation transforms the data attributes to a specific probability distribution.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because he quantile transformation is a non parametric normalisation method.

The answer (d) is not correct because Distribution normalisation assumes that all attribute follow the same probability distribution and transforms the data so all have the same moments.

■

3. Thinking about data preprocessing (multiple answer):

(a) Data discretization increases the complexity of the data by transforming continuous values to a discrete set of values.

● (b) The use of Kernel Density Estimation for data discretization allows to find the adequate number of discretization intervals for an attribute by finding areas with low data density in the estimated distribution.

(c) One advantage of range normalisation is that it is robust to outliers, so the transformation scales properly the data.

• (d) Discretization by equal sized bins consists on dividing the range of values on a determined number of intervals.

Solution:

The correct answers are (b) and (d).

The answer (a) is not correct because Discretization reduces the complexity of the data obtainign a more coarse grained version.

The answer (c) is not correct because Usual normalisation methods are not robust to outliers, so the distribution of the transformed data can be distorted.

■

4. Thinking about dimensionality reduction (multiple answer):

• (a) Least Squares Multidimensional Scaling obtains a set of new coordinates for the examples by minimizing the square difference of the distances among pairs of examples in the original and new space.

(b) Nonegative Matrix Factorization solves a convex optimization problem applying stochastic gradient descent over the transformation matrices.

• (c) Filtering is an attribute selection method that uses a relevance measure that assesses the importance of each attribute individually.

(d) The Laplacian score is a wrapper attribute selection method based on the K-means algorithm.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because Nonegative Matrix Factorization solves a non convex optimization problem.

The answer (d) is not correct because The Laplacian score is a filter attribute selection method based on the spectral matrix of the graph of k-nearest neighbors.

■

5. Thinking about dimensionality reduction (multiple answer):

• (a) PCA transforms the data to a new space with the same dimensions than the original data.

(b) Multidimensional scaling transforms a dataset by preserving the statistical distribution of the data.

• (c) Local Multidimensional Scaling tries to preserve in the new space the locality of closer neighbors and to increase the distance of the non neighbors.

(d) The variance of the data is evenly distributed over all the components obtained by PCA.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because Multidimensional scaling preserves the relative distances among the examples of the original space.

The answer (d) is not correct because The components obtained by PCA explain different amounts of variance, how much variance is explained is in the proportion of the eigenvalues of the SVD decomposition.

■

6. Thinking about dimensionality reduction (multiple answer):

(a) t-SNE is based on the assumption that distances have a uniform distribution.

• (b) The methods for multidimensional scaling need as a parameter the number of dimensions to be obtained by the transformation.

• (c) Unsupervised Attribute selection methods select the relevant attributes of the data instead of performing a transformation to a new space.

(d) t-SNE reproduces the distribution of the distances in the original space of the data in a low dimensional space by optimizing the squared difference between the distances among examples in both spaces.

Solution:

The correct answers are (b) and (c).

The answer (a) is not correct because t-SNE is based on the assumption that distances have a gaussian distribution.

The answer (d) is not correct because t-SNE optimizes the KL distance between the distribution of the distances among examples in both spaces.

■

7. Thinking about the Mixture Decomposition/Fuzzy clustering algorithms and its variants (multiple answer):

• (a) In fuzzy clustering, different cluster shapes can be discovered using the adequate objective function.

(b) The maximisation step of the EM algorithm used in Gaussian Mixture Models computes the probability for each example of being from each of the current clusters.
(c) Mixture models and Fuzzy clustering are algorithms that obtain a hard partition of a dataset.
● (d) Fuzzy clustering relaxes the hard partition constraint of K-means by allowing each example to have a membership degree to all the clusters.

Solution:

The correct answers are (a) and (d).

The answer (b) is not correct because This is computed in the expectation step.

The answer (c) is not correct because Mixture models and Fuzzy clustering obtain a soft partition of a dataset.

■

8. Thinking about the Mixture Decomposition/Fuzzy clustering algorithms and its variants (multiple answer):

● (a) An advantage of Mixture decomposition compared to K-means is that it returns a probability for the membership of each example to clusters (soft clustering) instead of performing a hard clustering (only one cluster assignment for example).
(b) The probabilistical model used for determining the clusters needs the number of clusters/mixtures and cannot be included in the estimation process.
● (c) Mixture decomposition assumes that data is generated by a set of probabilistical distributions and that in order to obtain the clusters, the parameters for each distribution have to be estimated.
(d) Using a mixture of Gaussian distributions for Mixture Decomposition, if we assume that the attributes are independent (only variances are estimated, not covariances) the hyperelipsoids that are fitted to the clusters can be arbitrarily oriented in space.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because Dirichlet Process Mixture model includes the estimation of the more adequate number of clusters in the estimation process.

The answer (d) is not correct because Estimating only the variances we assume that the hyperelipsoids are parallel to the coordinate axis (attributes).

■

9. Thinking about the K-means algorithm and its variants (multiple answer):

(a) The K-means algorithm is able to find arbitrarily shaped clusters in data.
(b) The K-means algorithm is a partitional algorithm that finds the optimal centroids by minimizing the square distance among the examples.
● (c) K-medoids is an alternative to K-means that uses one example of the cluster as prototype (medoid) instead of the centroid, as K-means does.
● (d) Bisecting K-means is a variant of the K-means algorithm that begins with one cluster and iteratively adds a new cluster by splitting one of the existing clusters chosen by a quality criteria until the K clusters are obtained.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because The K-mean looks for sphericaly shaped clusters in data.

The answer (b) is not correct because The distance minimized is the one among the examples and the cluster centroids and there is no guarantee that the solution will be optimal.

■

10. Thinking about the K-means algorithm and its variants (multiple answer):

● (a) K-medoids has as disadvantage over K-means that is computationally more expensive because it has to choose for each cluster the example that acts as medoid.
(b) The K-means algorithm is tolerant to outliers and can handle clusters of different sizes.
● (c) The K-means++ initialization strategy for K-means uses a randomized algorithm to maximize the distance among the initial centroids.
(d) Gobal K-means is a variant of the K-means algorithm that begins with one cluster and iteratively adds a new cluster by splitting one of the existing clusters chosen by a quality criteria until the k clusters are obtained.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because Outliers are difficult for K-means and clusters of different sizes can lead to an incorrect clustering of the data.

The answer (d) is not correct because This is what bisecting K-means does.

■

11. Thinking about hierarchical clustering (multiple answer):

(a) COBWEB is a hierarchical clustering algorithm that uses the examples distance matrix to agglomerative build a binary hierarchy with the examples.

● (b) Graph based hierarchical clustering algorithms can be defined as agglomerative or divisive algorithms.

(c) Graph based hierarchical clustering algorithms allow to define a continuum of clustering algorithm based on the link criteria used to merge clusters.

● (d) In Matrix Algebra based hierarchical clustering algorithms the distances to a new formed cluster is recomputed using several different criteria.

Solution:

The correct answers are (b) and (d).

The answer (a) is not correct because The hierarchies built by COBWEB do not necessarily have to be binary, it uses the Category Utility measure for deciding how to cluster examples and and the algorithm does not work only in an agglomerative fashion.

The answer (c) is not correct because Matrix algebra based algorithms are the one that define a continuum of algorithms.

■

12. Thinking about clustering validation (multiple answer):

(a) The Normalized Mutual Information is a variation of the Mutual Information where its value for two random partitions is zero.

● (b) The Normalized Mutual Information is an external clustering validity index based on infomation theory that measures the concidence of the distribution of examples in clusters for two clusterings.

(c) The advantage of using log likelihood to compare soft partitions respect to using distortion in hard partitions is that it can be used to compare partitions with different number of clusters.

● (d) External clustering validity indices compare a clustering with a reference partition.

Solution:

The correct answers are (b) and (d).

The answer (a) is not correct because The Normalized Mutual Information puts the value of Mutual Information in the range (0,1).

The answer (c) is not correct because Log likelihood depends on the number of clusters of a partition just like distortion.

■

13. Thinking about clustering validation (multiple answer):

● (a) Internal clustering validation indices are based on the model of the clusters and measure properties that a good clustering should have.

(b) The Gap statistic is computed comparing the clusters of a partition with partitions of several random gaussian distributed datasets.

● (c) The Davis-Bouldin criteria and the Silhouette index are internal clustering validity indices.

(d) Distortion can be used to compare partitions with different numbers of clusters.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because The Gap statistic is computed comparing the clusters of a partition with partitions of several uniformly distributed datasets.

The answer (d) is not correct because The distortion depends on the number of cluster, so partitions with different number of clusters can not be compared.

■

14. Thinking about clustering in KDD (multiple answer):

(a) CURE is a clustering algorithm that processes batches of data incrementally building a hierarchy of clusters that is compressed when the number of clusters at the deepest level reaches a threshold.

● (b) The one pass strategy for clustering scalability is based on using an incremental clustering algorithm as a preprocessing step of the dataset before using a more accurate algorithm.

● (c) The Compression/Summarization strategy for clustering scalability is based on discarding examples and using sufficient statistics to summarize their information.

(d) One key element for the scalability of BIRCH is that it uses a cheap distance to determine the assignment of new examples to the current clusters.

Solution:

The correct answers are (b) and (c).

The answer (a) is not correct because CURE is a hierarchical clustering algorithm that uses sampling to obtain P batches, clusters each batch separately and then joins all the results by applying a hierarchical algorithm.

The answer (d) is not correct because the key point of the scalability of BIRCH is the logarithmic time indexing datastructure for the clusters.

■

15. Thinking about clustering in KDD (multiple answer):

• (a) Mini-Batch K-means is an on-line version of K-means that processes consecutively small batches of examples to update the current protoypes until convergence.

(b) On many data stream clustering algorithms outliers are detected and discarded when there is no current cluster that is close enough.

• (c) Canopy clustering divides the dataset into several batches by using a cheap clustering algorithm and then clusters each batch separately using different strategies.

(d) Clustering of data streams assumes that data comes only from a stable model that can be summarized by a non incremental clustering algorithm.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because When clustering on data streams an example can not be discarded because it is not currently close to an existing dataset.

The answer (d) is not correct because Clustering of data streams assumes stable or changing generative model and the clustering has to be incremental.

■

16. Thinking about autoregressive models, which of the following statements are correct? (multiple answer):

(a) The advantage of using convolutions in autoregressive networks with respect to attention is that convolutions have a larger receptive field than attention.

(b) In autoregressive models we can use any network to represent the joint probability distribution of the variables, since they all represent probability distributions.

• (c) The advantage of using 2D convolutions for images in autoregressive networks compared to 1D convolutions is that 2D convolutions allow to capture better the spatial relationships between adjacent pixels.

• (d) The advantage of using causal convolutions in autoregressive models compared to the MLP network that uses MADE is that the convolutions are more efficient in time and memory.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because The attention mechanism has an unlimited receptive field, while convolutions have a limited receptive field.

The answer (b) is not correct because In order to represent joint probability distributions with neural networks, they must comply with the rules of probability.

■

17. Thinking about autoregressive models, which of the following statements are correct? (multiple answer):

• (a) In the MADE architecture the first output of the output layer is independent of the input variables.

(b) Sampling in the MADE autoregressive model is cheap because we use masking and that allows all variables to be sampled at the same time.

• (c) The Wavenet architecture uses dilated convolutions to increase the receptive field of the network so that more information from the past can be used.

(d) An advantage of the ImageGPT model is that it can generate high-resolution images because the attention mechanism has a linear cost in the image size.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because In MADE the values of the output variables must be obtained in autoregressive order, so the network has to be executed as many times as there are variables using the values of the previous variables.

The answer (d) is not correct because Attention has a quadratic cost in the image size, so it is not possible to generate high resolution images.

■

18. Thinking about autoregressive models, which of the following statements are correct? (multiple answer):

(a) One of the advantages of recurrent networks for representing autoregressive models is that each variable in the distribution uses a different set of parameters in the network.

(b) One of the advantages of recurrent networks is that sampling is fast because all variables can be sampled at the same time.

• (c) The Gated PixelCNN architecture is composed of two parallel branches that combine the result of 2D and 1D convolutions using sigmoidal activations and hyperbolic tangents as gates to decide how they are combined.

• (d) Gated PixelCNN improves upon PixelCNN by using masked 2D convolutions along with one-dimensional causal convolutions to avoid blind spots in the network's receptive field.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because In recurrent networks all variables share the parameters of the network.

The answer (b) is not correct because In recurrent networks (like all autoregressive models) the values of the output variables must be obtained in autoregressive order, so the network has to be run as many times as there are variables using the values of the previous variables.

■

19. Thinking about flow models, which of the following statements are correct? (multiple answer):

(a) A normalizing flow transforms data from any probability distribution to a unit uniform distribution.

(b) Autoregressive flows assume that each variable in a distribution is independent of the others to learn flows between distributions with multiple variables.

• (c) When applying flows to multiple variables at the same time, a scalability factor is that the Jacobian matrix corresponding to the derivative of the flow transformation is cheap to calculate, being for example a triangular matrix.

• (d) The advantage of elementary flows is that they make the variable-to-variable transformation independently and that ensures that the determinant of the Jacobian matrix is diagonal.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because Normalizing flows transform data to a unitary Gaussian distribution.

The answer (b) is not correct because Autoregressive flows allow the calculation of the flow between distributions with multiple variables by factoring the probability distribution using the product rule.

■

20. Thinking about flow models, which of the following statements are correct? (multiple answer):

• (a) The use in NICE/RealNVP of layers that make a permutation of the variables allows all the variables to influence the transformation of the others.

(b) An advantage of learning flows using neural networks is that any activation function we use makes the network an invertible function, unlike in other unsupervised models.

(c) An affine flow is a type of multidimensional flow that is efficient since the matrix that computes the linear transformation always has a triangular Jacobian.

• (d) Flows are models that guarantee a one-to-one mapping from one distribution to another because they are monotonic and invertible functions.

Solution:

The correct answers are (a) and (d).

The answer (b) is not correct because Not all activation functions allow a neural network to be invertible.

The answer (c) is not correct because There is no guarantee that any arbitrary linear transformation has a triangular Jacobian.

■

21. Thinking about latent variables models/VAEs, which of the following statements are correct? (multiple answer):

• (a) The reparametrization trick allows to train variational autoencoders (VAEs) with SGD by replacing the probabilistic operations necessary to train the autoencoder with deterministic nodes and Gaussian noise.

(b) Conditioning a variational autoencoder (VAE) using a categorical variable reduces the expressive capacity of the model, so it is better to use a normalizing flow to generate data samples that, for example, follow the classes of the data set.

• (c) Variational autoencoders (VAEs) for latent disentanglement modify the loss function so that the latent variables are more independent of each other.

(d) A variational autoencoder (VAE) is a neural network that deterministically maps data samples to samples in a latent variable space.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because There is no problem in conditioning a variational autoencoder (VAE) to a categorical variable.

The answer (d) is not correct because A variational autoencoder (VAE) computes probability distributions, so the mapping is nondeterministic.

■

22. Thinking about latent variables models/VAEs, which of the following statements are correct? (multiple answer):

• (a) Introducing additional supervised information into the encoder and decoder of a variational autoencoder (VAE) biases the model so that its output depends on the additional information.

• (b) In the variational autoencoder (VAE), a Gaussian distribution is usually used as the variational distribution and it is usually assumed that its variables are independent (diagonal covariance matrix) to reduce the number of parameters to be estimated.

(c) The advantage of choosing the Gaussian distribution as the variational distribution in variational autoencoders (VAEs) is that the generated samples will follow better the distribution of the real data.

(d) In the variational autoencoder (VAE) with variational Gaussian distribution, the encoder directly calculates samples of the latent variables that are transformed by the decoder to generate samples of the output data.

Solution:

The correct answers are (a) and (b).

The answer (c) is not correct because Choosing the Gaussian distribution as the variational distribution in variational autoencoders (VAEs) does not guarantee that the generated samples follow the distribution of the real data.

The answer (d) is not correct because The encoder in the variational autoencoder (VAE) with variational Gaussian distribution calculates the mean and standard deviation of the distribution of the latent variables from the input data.

■

23. Thinking about latent variables models/VAEs, which of the following statements are correct? (multiple answer):

(a) In variational inference, the evidence lower bound (ELBO) calculates the closeness between the intractable probability distribution function that links the data with the latent variables and the tractable distribution that we choose to approximate it.

(b) Latent variables models have in common with flows and autoregressive models that the log likelihood of samples can be calculated exactly.

• (c) Introducing a normalizing flow after the encoder of a variational autoencoder (VAE) allows the variational distribution to better fit the true distribution of the latent variables.

• (d) One way to improve variational autoencoders (VAEs) is to use a more flexible variational distribution, such as a multimodal normal distribution (mixture of Gaussians), which allows us to better approximate the real distribution of the latent variables.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because ELBO calculates a lower bound for the probability of the marginal distribution of the data.

The answer (b) is not correct because Latent variables models approximate the log likelihood, it cannot be calculated exactly.

■

24. Thinking about implicit models/GANs, which of the following statements are correct? (multiple answer):

• (a) To evaluate the quality of the data generated by a GAN, the value of the loss of the generator and the discriminator is not useful, it is necessary to use other substitute methods to evaluate it.

• (b) The Frechet Inception Distance (FID) allows to evaluate sets of samples from a GAN using a pre-trained classification network, by using one of its layers to generate a feature space where it can be measured the distance between the distributions of the generated samples and the real samples.

(c) The truncation trick consists of truncating the values of the input noise distribution to the generator to a maximum value so that the generated samples are more diverse, at the cost of their realism.

(d) BigGAN is a GAN that uses a different architecture from the usual one for the generator, which consists of a mapping network that transforms the input latent noise into a new space (style vectors), and a generating network that generates the image using those style vectors.

Solution:

The correct answers are (a) and (b).

The answer (c) is not correct because The truncation trick consists of truncating the values of the input noise distribution to the generator to a maximum value so that the generated samples are more realistic, at the cost of their diversity.

The answer (d) is not correct because It is the StyleGAN network that uses a different architecture from the usual one for the generator, which consists of a mapping network that transforms the input latent noise into a new space (style vectors), and a generator network that generates the image using those style vectors.

■

25. Thinking about implicit models/GANs, which of the following statements are correct? (multiple answer):

(a) Generative Adversarial Networks (GANs) are based on minimizing the Evidence Lower Bound (ELBO) of the joint probability distribution learned by two neural networks (generator/discriminator).

• (b) The goal of the discriminator in Generative Adversarial Networks (GANs) is to determine whether a sample is real or generated.

• (c) Conditional GANs are a variation of GANs where the generator additionally to the noise input gets side information that bias what the generator learns so the results are conditioned to the side input.

(d) Mode collapse is one of the problems that can arise when training GANs, and refers to the circumstance that the discriminator becomes so bad that it cannot distinguish real data from generated samples.

Solution:

The correct answers are (b) and (c).

The answer (a) is not correct because GANs are based on training two networks that play a minimax game.

The answer (d) is not correct because Mode collapse refers to the circumstance that the generator is only capable of generating samples of some of the distribution modalities of the real data.

■

26. Thinking about implicit models/GANs, which of the following statements are correct? (multiple answer):

(a) The Wasserstein GAN with gradient penalty (WGAN-GP) restricts the function learned by the generating network to be 1-Lipschitz by limiting gradients that are greater than one to a constant, so they are penalized.

• (b) Generative Adversarial Networks (GANs) are implicit models, that is, there is no explicit representation of the parameters of the probability distribution that is learned, but they allow to generate samples from the data distribution.

• (c) The Wasserstein GAN (WGAN) reformulates the loss of the original GAN to a problem where the goal is to minimize the distance between the probability distribution of the real data and the data generated using the function that calculates the discriminator.

(d) The original loss function of GANs has a vanishing gradient problem when the discriminator is very confident in classifying real samples from generated ones, this can be solved by inverting the game that is solved, making it a maximin problem instead of a minimax problem.

Solution:

The correct answers are (b) and (c).

The answer (a) is not correct because WGAN-GP penalizes the loss function to make the gradient of the discriminator function close to 1.

The answer (d) is not correct because To resolve vanishing gradient, the generator optimizes the loss function inverse to the one optimized by the discriminator for the generated samples.

■

27. Thinking about diffusion models, which of the following statements are correct? (multiple answer):

(a) The advantage of using an ODE on score based modeling with SDE is that the denoising process is deterministic, so there is a one to one mapping between the generated and the real data.

(b) The denoising network commonly used by DDPM is an autoencoder that receives a sample and a time step that is encoded with an embedding that uses positional encoding.

• (c) Latent diffusion makes it possible to accelerate the diffusion process by transforming the examples to a latent space of lower dimensionality (for example with a VAE) and carrying out the diffusion process in that space.

• (d) A problem with DDPM is that because the defined diffusion process is only Gaussian when two consecutive diffusion steps are very close, many steps must be performed to obtain a good result.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because Using an ODE defines a one to one mapping between the data and the noise distribution.

The answer (b) is not correct because DDPM uses a U-Net as a denoising network, not an autoencoder.

■

28. Thinking about diffusion models, which of the following statements are correct? (multiple answer):

(a) DDPM is an alternative formulation to DDIM that defines a non-Markov diffusion process that allows accelerating the diffusion process by skipping time steps.

(b) Diffusion models like flows and autoregressive models can generate high quality samples.

● (c) The DDPM training process optimizes each term of the loss function independently by selecting a random time step of the diffusion process and using a sample diffused at that time step.

● (d) One advantage of defining Score based modeling with SDE as an ODE is that ODEs can be integrated using advanced solvers of systems of differential equations that allow taking larger steps in the reverse process, so sampling is faster.

Solution:

The correct answers are (c) and (d).

The answer (a) is not correct because DDIM is what defines the diffusion process as non-Markovian.

The answer (b) is not correct because Diffusion models do generate high quality samples, but not flows or autoregressive models.

■

29. Thinking about diffusion models, which of the following statements are correct? (multiple answer):

● (a) One of the advantages of defining Score based modeling with SDE using an ODE is that we can perform semantic interpolation between latents as in GANs.

(b) DDPM defines a continuous denoising process based on a Gaussian Markov chain.

● (c) The forward process of denoising diffusion models is cheap to compute, because the noise to be added to an example at step t can be computed analytically.

(d) We can speed up the DDPM sampling process by jumping from k to k time steps uniformly obtaining good quality samples.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because The denoising process defined by DDPM is discrete.

The answer (d) is not correct because Jumping uniformly from k to k time steps does not give good results, since sample information is lost.

■

30. Thinking about self-supervised deep learning methods (multiple answer):

● (a) Self-supervised contrastive learning methods train two encoding networks that have the same architecture, but do not share weights.

(b) Word embeddings based on skip-grams are vector representations of words learned using the task of predicting the probability of a word from its neighboring words in a text.

● (c) Self-supervised methods using contrastive learning train two encoding networks, each of which generates codes for different augmentations of an example.

(d) In contrastive learning, the patch masking technique is used to obtain the training examples.

Solution:

The correct answers are (a) and (c).

The answer (b) is not correct because Skip-grams are vector representations of words learned using the task of predicting from a word the probability of its neighboring words in a text.

The answer (d) is not correct because In contrastive learning, data augmentation techniques are used to obtain training examples.

■

End Quiz

☼