Exam Protocol
Machine Learning 2

1. Multiple Choice

Which of the following is True: Let $k(x,x') = \sum_{i=1}^{d} \#w_i(x) \cdot \#w_i(x')$ be a string kernel, where $\#w_i(x)$ indicates the number of occurrences of word w_i in sequence x. A feature map associated to this kernel is given by:

□ $\phi : x \mapsto \sum_{i=1}^{d} \#w_i(x)$

□ $\phi : x \mapsto (\#w_1(x), ..., \#w_d(x))$

□ $\phi : x \mapsto (\#w_1(x), ..., \#w_{\sqrt{d}}(x))$

□ $\phi : x \mapsto \sum_{i=1}^{d} \sqrt{\#w_i(x)}$

Which of the following is True: The role of spatial pooling in a convolutional network (CNN) is:

□ To use the same feature detectors for each part of the input image.

□ To have some translation invariance in the pixel space.

□ To reuse low-level features for different classes.

□ To make the network recurrent in space.

Which of the following is True: Sparse auto-encoders:

□ Learn a feature space representation of sparse data.

□ Learn a model whose parameters are regularized for sparsity.

□ Learn a model whose reconstructions are regularized for sparsity.

□ Learn a sparse feature space representation of the data.

Which of the following is True: A sparse code is:

□ An encoding of the input data in terms of 0's and 1's that allows efficient storage on a computer.

□ A representation of the data that is suitable for transmissions over a noisy channel.

□ A representation of the data consisting of a small number of "dictionary" 'elements.

□ A low-dimensional projection of the data that preserves most of the information in the data.

2. Practical ML

We consider the practical scenario of a furniture store, in which we would like products to be arranged in an optimized manner. Specifically, we would like similar products to be located nearby in the store. Similarity might be e.g. sharing the same brand, being of same type (e.g.chair) or being made of the same construction material (e.g. oak). Because of the large number of products, we adopt a data driven approach, where we have first produced for each product a tabular description containing the brand, the construction material, the type of object, the color, etc.

2.1 *Select* a combination of methods that can be used to address the problem above.

□ neural networks + Hidden Markov models

□ t-SNE + structured kernels

□ neural networks + explainable AI

□ One-class SVM + structured kernels

2.2 *Describe* for this particular application how the data should be processed (e.g. which features are extracted) and if applicable, how to build a representation of the data or some similarity model.

2.3 *Explain* how the data and its representations/similarity model is fed to the machine learning algorithm, how to set the parameters of the algorithm for the purpose of the application and what the machine learning algorithm optimizes.

2.4 Explain how the output of the learning algorithm or analysis is used concretely to solve the application problem above.

3. String Kernels

A kernel is positive semi-definite, if it satisfies the property

$$\sum_{i=1}^{N}\sum_{j=1}^{N} c_i c_j k(x_i, x_j) \geq 0$$

for all inputs $x_1, ..., n_N$ and choice of real numbers $c_1, ..., c_N$.

Let T and T' be two documents represented by the set of words composing them. Let W denote the set of all words in the dictionary. We consider the structured kernel

$$k(T, T') = \left(\sum_{a \in T}\sum_{b \in T'} I(a = b)\right) - \theta$$

where I is an indicator function and where we assume $\theta \geq 0$.

3.1 Considering that the input domain is restricted to documents which always include the footer “for confidential use only”, give the maximum value of the parameter θ which ensures that the kernel remains positive semi-definite.

3.2 Give for the special case $\theta = 0$ a feature map $\phi(T)$ associated to this kernel.

3.3 Give for the general case θ, and incorporating the previously stated constraints on the domain, a feature map $\phi(T)$ associated to this kernel. (Note: the feature map must be a vector of real-valued numbers, in particular, it cannot incorporate complex numbers).

3.4 Compute $||\phi(T)||$ where T is a document taken from the domain stated above and in addition to words that are included in every document of the domain, also includes the additional sentence "for more detailed information we refer to the appendix".

4. Conditional Restricted Boltzmann Machine

We consider a conditional RBM (CRBM) model composed of one input unit $x \in \{0,1\}$, one hidden unit $h \in \{0,1\}$ and one output unit $y \in \{0,1\}$. The CRBM to each joint configuration of units the probability:

$$P(x,h,y) = \frac{1}{Z}\exp(-E(x,h,y))$$

Where E is an energy function to be specified and where Z is a normalization constant. In the following, we will be interested in the marginalized distribution $P(x,y)$ for the different states $(x,y) \in \{0,1\}^2$

4.1 We consider the energy function:

$$E(x,h,y) = -xwh + hwy$$

where $w \in \mathbb{R}$ is a parameter of the model. Using the equations above, compute the probability function $P(x,y)$ for each state $(x,y) \in \{0,1\}^2$.

4.2 We now consider the "structured output" scenario where we would like to learn the parameter w such that we can predict y from x. Compute the conditional probability $p(y|x)$ for all cases $(x,y) \in \{0,1\}^2$.

4.3 Explain how the parameter w controls the correlation between the input x and the output of the model y.

4.4 Write the free energy function associated to the CRBM defined above.

5. Programming

We consider a Hidden Markov Model defined by the transition matrix A of size $h \times h$, the emission matrix B of size $h \times d$ and the initial state vector π of size h. All these matrices are stored in numpy arrays. The number of states h is a hyperparameter to be selected and the number of dimensions d is problem dependent.
We would like to build a function that checks, whether the HMM given as input is a valid HMM (i.e. matrices have correct dimensions and entries of the matrices correspond to valid probability distributions).

5.1 Implement such function for the case where the HMM models sequences of musical notes. (In such case, we choose the set of observed symbols to cover 5 octaves, each of which containing 12 musical notes. Therefore, $d = 60$.

5.2 Create a method that simulates the HMM described above for a given number of iterations T, and as a result returns a sequence of observed symbols.