FUNDAMENTALS OF MACHINE LEARNING
AA 2023-2024
Prova Finale (FACSIMILE)

18 Dicembre, 2023

Istruzioni: Niente libri, niente appunti, niente dispositivi elettronici, e niente carta per appunti. Usare matita o penna di qualsiasi colore. Usare lo spazio fornito per le risposte.
Instructions: No books, no notes, no electronic devices, and no scratch paper. Use pen or pencil. Use the space provided for your answers.
This exam has 5 questions, for a total of 100 points and 10 bonus points.

Nome: _____

Matricola: _____

1. **Multiple Choice**: Select the correct answer from the list of choices.
   (a) [5 points] True or False: A K-nearest neighbor classifier is only able to learn linear discriminant functions. ◯ True ◯ False
   (b) [5 points] True or False: Projecting a dataset onto its first principal component maximizes the variance of the projected data. ◯ True ◯ False
   (c) [5 points] True or False: The K-means algorithm is guaranteed to find the best cluster centers for any dataset. ◯ True ◯ False
   (d) [5 points] True or False: A Parzen kernel density estimator uses only the nearest sample in the dataset to estimate the probability of an input sample $\mathbf{x}$. ◯ True ◯ False
   (e) [5 points] How many parameters will a Multilayer Perceptron (MLP) for binary classification with a single hidden layer of width 10 and an input dimensionality of 8 have?
   ◯ 80 ◯ 99 ◯ 88 ◯ None of the above
   (f) [5 points] What will the entries of the Gram matrix be for a linear kernel?
   ◯ $K[i,j] = (\mathbf{x}_i^T \mathbf{x}_j)^\gamma$
   ◯ $K[i,j] = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||_2^2)$
   ◯ $K[i,j] = \mathbf{x}_i^T \mathbf{x}_j$
   ◯ None of the above
   (g) [5 points] Which of the following loss functions is called the negative log likelihood?
   ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} (\ln y_c - \ln \hat{y}_c)^2$
   ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} (y_c - \ln \hat{y}_c)^2$
   ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} y_c \ln \hat{y}_c$
   ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \ln \hat{y}_c$
   (h) [5 points] How many iterations of gradient descent must we perform for an epoch of minibatch Stochastic Gradient Descent with a dataset of 1024 samples and a batch size of 16?
   ◯ 1024 ◯ 1 ◯ 32 ◯ 64

Total Question 1: 40

2. **Multiple Answer**: Select **ALL** correct choices: there may be more than one correct choice, but there is always at least one correct choice.

(a) [5 points] What are the advantages of projecting data onto $K < D$ principal components?

- ◯ We eliminate noise in the original representation.
- ◯ Classes are guaranteed to be linearly separable.
- ◯ It is a nonlinear embedding that makes learning easy with simpler models.
- ◯ Models trained on the reduced data are simpler.

(b) [5 points] Which of the following are advantages of Ensemble Models (e.g. Committees)?

- ◯ They reduce the variance of the resulting model.
- ◯ They are much more efficient than the base model.
- ◯ They can reduce the expected error of the final model.
- ◯ The resulting model is nonlinear even if the base model is linear.

(c) [5 points] Which of the following are causes of the vanishing gradients when training neural networks?

- ◯ Saturated inputs to activation functions with near-zero derivatives when saturated.
- ◯ Badly scaled input values.
- ◯ Very deep models.
- ◯ Bad random initialization of the network parameters.

(d) [5 points] If we want to penalize classification errors less when training an SVM we should

- ◯ Increase the hyperparameter C.
- ◯ Use a radial basis kernel.
- ◯ Decrease the hyperparameter C.
- ◯ None of the above.

(e) [5 points] Which of the following are requirements for applying backpropagation to compute gradients in a deep network?

- ◯ The network must not be too deep.
- ◯ The network must be a directed acyclic graph.
- ◯ All activation functions must be differentiable.
- ◯ All activation functions must be continuous.

(f) [5 points] Which of the following are true of the Nadaraya-Watson estimator?

- ◯ It only requires some of the training data at test time.
- ◯ It is a nonparametric method.
- ◯ It estimates a nonlinear function of the input.
- ◯ It estimates a linear function of the input.

(g) [5 points] Which of the following models are nonparametric?

- ◯ The Multilayer Perceptron (MLP).
- ◯ Logistic regression.
- ◯ The K-Nearest Neighbor Classifier
- ◯ Decision Trees.

Total Question 2: 35

3. [10 points] Show that a Committee Ensemble model using N bootstrapped linear regression models is a linear regression (i.e. that can be expressed as $\mathbf{w}^T\mathbf{x} + b$ for some $\mathbf{w}$ and b).

4. [15 points] Show that a Multilayer Perceptron with two hidden layers with activation function $\sigma(x) = x$ is only capable of learning linear functions.

5. [10 points (bonus)] Design a Deep Convolutional Neural Network (with at least three convolutional layers and one or more pooling layers) to classify MNIST images (input size 28×28). Draw the network (or write pseudocode for its definition) and indicate how many parameters each layer has and the sizes of the intermediate feature maps.