FUNDAMENTALS OF MACHINE LEARNING
AA 2025-2026
Prova Finale (FACSIMILE)

16 December, 2025

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.

First and Last Name: _____

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: 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

   (c) [5 points] How many parameters will a Multilayer Perceptron (MLP) for a three-class classification problem with a single hidden layer 10 units and an input dimensionality of 8 have?
   ◯ 120 ◯ 90 ◯ 123 ◯ None of the above

   (d) [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

   (e) [5 points] Which of the following loss functions is called the negative log likelihood?
   - ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} (\ln y_c - \ln \hat{y}_c)^2$
   - ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} (y_c - \ln \hat{y}_c)^2$
   - ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} y_c \ln \hat{y}_c$
   - ◯ $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{c=1}^{C} \ln \hat{y}_c$

   (f) [5 points] Which of the following activation functions is called the Rectified Linear Unit (ReLU)?
   - ◯ $\sigma(z) = \min(0, z)$
   - ◯ $\sigma(z) = \frac{1}{1+e^{-z}}$
   - ◯ $\sigma(z) = \max(0, z)$
   - ◯ $\sigma(z) = \frac{1}{e^{-z}}$

   (g) [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: 35

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] 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.

(b) [5 points] Which of the following are true of the vanishing gradient problem for networks using sigmoid activation functions?

- ◯ Using ReLU units instead of sigmoid units can mitigate the risk of vanishing gradients.
- ◯ If a unit has the vanishing gradient problem for one training point, it has the problem for every training point.
- ◯ Networks with sigmoid units don't have this problem if they're trained with the cross-entropy loss function.
- ◯ Deeper neural networks tend to be more susceptible to vanishing gradients.

(c) [5 points] What do residual connections in a Deep Neural Network do?

- ◯ They help mitigate the problem of vanishing gradients.
- ◯ They make training deeper models possibile.
- ◯ They introduce additional nonlinear activations in the network.
- ◯ None of the above

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

- ◯ The model must be a Convolutional Neural Network.
- ◯ The network structure must be a directed acyclic graph (DAG).
- ◯ All activation functions used must be differentiable almost everywhere.
- ◯ The network must only contain linear activation functions.

(e) [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 estimates a linear function of the input.
- ◯ It is a nonparametric method.
- ◯ It estimates a nonlinear function of the input.

(f) [5 points] What does the learning rate control in Stochastic Gradient Descent?

- ◯ The size of gradient steps made in each iteration.
- ◯ The degree of nonlinearity in the model.
- ◯ The number of iterations per epoch.
- ◯ The speed at which the model learns.

(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. [15 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).

   Note: Be sure to state all assumptions you make in your answer.

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.

   Note: Be sure to state all assumptions you make in your answer.

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.

   Note: Be sure to state all assumptions you make in your answer.