FIRST NAME: .....

LAST NAME: .....

ID NUMBER: .....

1. The goal of supervised learning is:
   (a) to learn what features are useful
   (b) to learn a model with low true risk
   (c) to learn a model with low empirical risk
   (d) to learn a deep neural network
   (e) none of the above
2. If you flip 10 times a coin that has probability 0.25 to give tail, the probability that you obtain exactly 1 head is:
   (a) $0.25^9 \times (1 - 0.25)$
   (b) $(1 - 0.25)^9 \times 0.25$
   (c) $10 \times 0.25^9 \times (1 - 0.25)$
   (d) $10 \times (1 - 0.25)^9 \times 0.25$
   (e) none of the above
3. Let $\mathcal{X}, \mathcal{Y}, \mathcal{D}, \ell(\mathbf{x}, y), \mathcal{H}, h, S$ defined as usual during the course. The definition of *training* error $L_S(h)$ is:
   (a) $L_S(h) = \mathbf{E}_{\mathbf{x},y\sim\mathcal{S}}[(h(\mathbf{x}) - y)^2]$
   (b) $L_S(h) = \frac{1}{|S|}\sum_{i=1}^{|S|}(h(\mathbf{x}_i) - y_i)^2$
   (c) $L_S(h) = \mathbf{E}_{\mathbf{x},y\sim\mathcal{S}}[\ell(h, (\mathbf{x}, y))]$
   (d) $L_S(h) = \frac{1}{|S|}\sum_{i=1}^{|S|}\ell(h, (\mathbf{x}_i, y_i))$
   (e) none of the above
4. Let $\mathcal{X}, \mathcal{Y}, \mathcal{D}, \ell(\mathbf{x}, y), \mathcal{H}, h, S$ defined as usual during the course. The definition of *generalization* error $L_{\mathcal{D}}(h)$ is:
   (a) $L_{\mathcal{D}}(h) = \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[(h(\mathbf{x}) - y)^2]$
   (b) $L_{\mathcal{D}}(h) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(h(\mathbf{x}_i) - y_i)^2$
   (c) $L_{\mathcal{D}}(h) = \mathbf{E}_{\mathbf{x},y\sim\mathcal{D}}[\ell(h, (\mathbf{x}, y))]$
   (d) $L_{\mathcal{D}}(h) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\ell(h, (\mathbf{x}_i, y_i))$
   (e) none of the above
5. In the ERM approach we pick the model for making predictions by:
   (a) finding the hypothesis with smallest training error
   (b) finding the hypothesis with smallest generalization error
   (c) finding the hypothesis with smallest complexity
   (d) finding the hypothesis that minimizes the expected regularization
   (e) none of the above
6. Ideally, the choice of the loss function should depend on:
   (a) the real-world problem you are trying to solve
   (b) the hypothesis class you want to use
   (c) the implementation that is available
   (d) the amount of data you have
   (e) none of the above
7. What does "overfitting" refer to?
   (a) Learning a model that has perfect accuracy on all datasets
   (b) Learning a model that performs well on both training data and validation data
   (c) Learning a model that performs well on training data but poorly on new data
   (d) Failing to converge during training
   (e) none of the above

8. A dataset is linearly separable if:
   (a) you can connect its instances with a line
   (b) you can learn a liner model from it
   (c) there is linear model that perfectly classifies it
   (d) you can plot it in 2 dimensions
   (e) none of the above
9. In the context of machine learning algorithms, what does the term "gradient descent" refer to?
   (a) A method for solving linear equations
   (b) An optimization algorithm used to minimize a cost function
   (c) A technique for clustering data points
   (d) A form of unsupervised learning
   (e) none of the above
10. What is the main idea behind the concept of bias-complexity trade-off?
   (a) Balancing the trade-off between model simplicity and interpretability
   (b) Balancing the trade-off between accuracy and training time
   (c) Balancing the trade-off between estimation error and approximation error
   (d) Balancing the trade-off between feature selection and feature engineering
   (e) none of the above
11. What is the purpose of regularization?
   (a) To reduce model complexity so to prevent overfitting
   (b) To increase model complexity so to prevent overfitting
   (c) To improve training speed
   (d) To eliminate bias in the model
   (e) none of the above
12. What is the purpose of using a test set?
   (a) To test the model on multiple datasets
   (b) To assess the model's performance on the training set
   (c) To compare different models on unseen data
   (d) To obtain a good estimate of the generalization error
   (e) none of the above
13. What is a main difference between SVMs and linear models?
   (a) there is no difference
   (b) SVM can be used to learn models that are polynomial in the features
   (c) SVMs can be used only for linearly separable data
   (d) SVMs consider the margin of the model, linear models do not
   (e) none of the above
14. The VC dimension is a measure of:
   (a) the dimension of each point in a dataset
   (b) the complexity of an hypothesis class
   (c) the number of features in a model
   (d) the generalizability of an hypothesis
   (e) none of the above
15. What is the key idea behind hierarchical clustering?
   (a) Assigning each data point to the nearest centroid
   (b) Dividing the dataset into a fixed number of clusters
   (c) Creating a tree-like structure representing various clusterings
   (d) Determining the density of data points in different regions
   (e) none of the above