Machine Learning — Statistical Methods for Machine Learning

Quiz list for the written test

Instructor: *Nicolò Cesa-Bianchi* version of September 14, 2025

A variable subset of this quiz list will be used to create the written test in each exam session. Bonus quizzes (not in this list) will be added to each test for extra points.

1. Write the formulas for the square loss, the zero-one loss, the hinge loss, and the logarithmic loss.
2. Describe what a learning algorithm receives in input and what it produces in output.
3. Write the mathematical expression defining the training error $\ell_S(h)$ of a predictor h.
4. Write the mathematical expression defining the ERM algorithm over a class $\mathcal{H}$ of predictors. Define the main quantities occurring in the formula.
5. Explain in words how overfitting and underfitting are defined in terms of the behavior of a learning algorithm on the training and test sets.
6. Name and describe three reasons why labels may be noisy.
7. Describe the classifier obtained when running k-NN on a training set of size k.
8. Suppose you run 1-NN on a training set of size m for binary classification (containing no duplicate datapoints). What can you say about the training error of the resulting classifier?
9. Write the classification rule used by k-NN for binary classification.
10. Write the classification rule used by k-NN for regression.
11. Is k-NN more likely to overfit when k is large or small? Justify your answer.
12. Use Jensen's inequality to prove that the training error of a tree classifier never increases after a split based on a concave splitting criterion.
13. Write the formulas for at least two splitting criteria ψ used in practice to build tree classifiers.
14. Explain the relationship between tree predictors and DNF formulas.
15. Write the formula defining the statistical risk $\ell_{\mathcal{D}}(h)$ of a predictor h with respect to a generic loss function and data distribution.
16. Write the formula defining the Bayes optimal predictor f^* for a generic loss function and data distribution.
17. Write the formula defining the Bayes optimal predictor and the Bayes risk for the zero-one loss. Define the main quantities occurring in the formula.

18. Can the Bayes risk for the zero-one loss be zero? Justify your answer.

19. Write the formula defining the Bayes optimal predictor and the Bayes risk for the square loss.

20. Explain in mathematical terms the relationship between test error and statistical risk.

21. State the Chernoff-Hoeffding bounds.

22. Write the bias-variance decomposition for a generic learning algorithm A and associate the resulting components to overfitting and underfitting.

23. Write the upper bound on the estimation error of ERM run on a finite class $\mathcal{H}$ of predictors. Define the main quantities occurring in the formula.

24. What is the number of nodes that a binary tree classifier must have so that by assigning labels to its leaves we can compute any function of the form $h : \{0,1\}^d \to \{-1,1\}$?

25. Write an upper bound on the cardinality of the set $\mathcal{H}_N$ of all classifiers computed by complete binary tree predictors with exactly N nodes on d binary features.

26. Write the upper bound on the estimation error of ERM run on a the class of complete binary tree predictors with at most N nodes on d binary features.

27. How many bits are sufficient to encode an arbitrary tree predictor with N nodes on d binary features so that no tree has an encoding that is a prefix of the encoding of a different tree?

28. Write the bound on the difference between risk and training error for an arbitrary complete binary tree classifier h on d binary features in terms of its number N_h of nodes. Bonus points if you provide a short explanation on how this bound is obtained.

29. Write the formula for the K-fold cross validation estimate. Explain the main quantities occurring in the formula.

30. Define the leave-one-out estimate and explain its mathematical relation with the expected risk of a generic learning algorithm A.

31. Given a training set S, describe a procedure to find the predictor in $\{A_\theta(S) : \theta \in \Theta\}$ with smallest risk.

32. Write the pseudo-code for computing the nested cross validation estimate.

33. Define the quantity estimated by the nested cross valdation estimate.

34. Write the mathematical definition of consistency for an algorithm A. Define the main quantities occurring in the formula.

35. Write the mathematical definition of nonparametric learning algorithm. Define the main quantities occurring in the formula.

36. Write an upper bound on the expected risk of the binary classifier returned by k-NN (with odd k) as the training set size m goes to infinity.

37. For what values of k is k-NN for binary classification consistent when the Bayes risk is zero?
38. Write the statement of the no-free-lunch theorem.
39. Name one nonparametric learning algorithm and one parametric learning algorithm.
40. Provide an example of a nonparametric learning algorithm that is not consistent.
41. Provide an example of a choice of k (as a function of the training set size m) that makes k-NN consistent.
42. Write the formula for the Lipschitz condition in a binary classification problem. Define the main quantities occurring in the formula.
43. Write the typical rate at which the risk of a consistent learning algorithm for binary classification vanishes as a function of the training set size m and the dimension d under Lipschitz assumptions.
44. Explain the curse of dimensionality.
45. Write the bound on the risk of the 1-NN binary classifier under Lipschitz assumptions.
46. When the Bayes risk is zero, 1-NN converges to the Bayes risk at rate of order $m^{-1/(d+1)}$. Is this statement true or false? Justify your answer.
47. Can ERM over linear classifiers be computed efficiently for the zero-one loss? Can it be approximated efficiently? Motivate your answers.
48. Write the system of linear inequalities defining the condition of linear separability for a training set in binary classfication.
49. Write the pseudo-code for the Perceptron algorithm.
50. Write the statement of the Perceptron convergence theorem. Define the main quantities occurring in the formula.
51. Write the closed-form formula (i.e., not the argmin definition) defining the Ridge Regression predictor. Define the main quantities occurring in the formula.
52. Write the pseudo-code for the projected online gradient descent algorithm.
53. Write the upper bound on the regret of projected online gradient descent on convex functions. Define the main quantities occurring in the bound.
54. Write the upper bound on the regret of online gradient descent on σ-strongly convex functions. Define the main quantities occurring in the bound.
55. Write the mistake bound for the Perceptron run on an arbitrary data stream for binary classification. Define the main quantities occurring in the bound.
56. Write the formula for the polynomial kernel of degree n.
57. Write the formula for the Gaussian kernel with parameter γ.

58. Write the pseudo-code of the kernel Perceptron algorithm.

59. Write the mathematical definition of the linear space $\mathcal{H}_K$ of functions induced by a kernel K.

60. Let f be an arbitrary element of the linear space $\mathcal{H}_K$ induced by a kernel K. Write $f(\boldsymbol{x})$ in terms of K.

61. Write the mistake bound of the Perceptron convergence theorem when the Perceptron is run with a kernel K. Define the main quantities occurring in the bound.

62. Write the closed-form formula (i.e., not the argmin definition) of the kernel version of the Ridge Regression predictor. Define the main quantities occurring in the formula.

63. Write the convex minimization problem with linear constraints that defines the SVM hyperplane in the linearly separable case.

64. Write the convex minimization problem with linear constraints that defines the SVM hyperplane when the training set is not necessarily linearly separable.

65. Write the unconstrained optimization problem whose solution defines the SVM hyperplane when the training set is not necessarily linearly separable.

66. Write the Fritz John optimality condition. Define the main quantities occurring in the condition.

67. Write the pseudo-code of the Pegasos algorithm.

68. Write the bound on the expected value of the SVM objective function achieved by Pegasos. Provide also a bound on the expected squared norm of the loss gradient in terms of a bound X on the Euclidean norm of the training points.

69. Write the definition of ε-stability for a learning algorithm. Define the main quantities occurring in the formula.

70. Write the value of ε for which SVM is known to be stable. The value depends on the radius X of the ball where the training data points live, the training set size m, and the regularization coefficient λ.

71. Provide an example of a choice of the regularization coefficient λ (as a function of the training set size m) ensuring consistency for the SVM algorithm wih Gaussian kernel.

72. Consider the class $\mathcal{F}_d$ of all functions of the form $f : \{-1,1\}^d \to \{-1,1\}$. Let $\mathcal{F}_{G,\mathrm{sgn}}$ be the class of functions computed by a feedforward neural networks with the sgn activation function and graph $G = (V,E)$. Provide asymptotic upper and lower bounds on $|V|$ such that $\mathcal{F}_d \subseteq \mathcal{F}_{G,\mathrm{sgn}}$.

73. Describe a class of neural networks for which the ERM problem with the square loss is NP-hard.

74. Write the update line of the stochastic gradient descent algorithm. Explain the main quantities occurring in the formula and the reason why the algorithm is stochastic.

75. Write the recursive definition of the partial derivative $\frac{\partial \ell_t(W)}{\partial w_{i,j}}$ used in the backpropagation algorithm. Explain the main quantities occurring in the definition.

76. Write the definition of logistic loss for logistic regression with linear models.

77. Write the gradient descent update for regularized logistic regression with a linear model.