CSCI-UA 9473
Introduction to Machine Learning
Practice (theory) questions

May 2019

Part I Statistics

1. Explain the difference between the Bayesian and the Frequentist approaches in parameter estimation. Give an estimator from each framework and illustrate the connection between the two frameworks through Bayes's theorem.

Part II Linear regression

2. What minimization problem do you solve when you want to fit a line $y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ to a set of pairs $\{\boldsymbol{x}_\mu, t_\mu\}_{\mu=1}^N$such that $y(\boldsymbol{x}_\mu) \approx t_\mu$. Solve the problem and give the final expression for $(\boldsymbol{w}, b)$ as a function of the matrix $\boldsymbol{X}$ encoding the prototypes $\boldsymbol{x}_\mu$ (or the matrix $\boldsymbol{\Phi}$ encoding the feature vectors $\phi_\mu(\boldsymbol{x})$).

3. What is meant by "bias/variance trade-off"?, Explain the connection of this notion to the prediction error and to the complexity of a model.

4. In linear regression, we have seen that correlated features lead to a larger variance contribution in the bias variance decomposition. Give three approaches that can be used to reduce the variance and hence the prediction error in this case. How does each approach extend the simple linear regression model (give the mathematical formulation of each extension)?

5. Give an example of a dataset that is not linearly separable in $\mathbb{R}^m$ but is linearly separable in $\mathbb{R}^{m'}$ with $m' > m$.

Part III Linear classification

6. What is the simplest way to define a binary (two-classes) classifier? Give the mathematical expression of such classifier and the problem that one has to solve to get the coefficients that appear in it.

7. In Logistic Regression, how are the posterior class probabilities modeled? How can we learn this model from data $\{\boldsymbol{x}_\mu, t_\mu\}$?

8. Explain how one can extend a binary (or two-class) classifier to a multiclass problem (Give two extensions).

Part IV Kernels and SVM

9. When are kernels particularly useful? Illustrate with an example.

10. Give an example of a kernel defined from feature vectors $\{\phi_\mu(x)\}$.

11. Explain how the kernel trick can be used to "transform" the linear regression formulation into a formulation defined only in terms of the Kernel matrix.

12. Give the expression of the distance of a point to a plane and illustrate your derivation with a drawing.

13. What optimization problem do you solve when you want to compute a maximal margin classifier for a two-class dataset $\mathcal{D} = \{\boldsymbol{x}_\mu, t_\mu\}_{\mu=1}^N$ where $t_\mu \in \{0, 1\}$? Give the mathematical formulation of the problem. Also explain the terms that appear in the formulation by relying on the notion of distance of a point to a hyperplane.

14. Give the final expression (mathematical expression of the classifier $y(\boldsymbol{x})$ as a function of the training set $\{\boldsymbol{x}_\mu, t_\mu\}$) for the maximal margin classifier.

15. What is a support vector? What are the two main characteristics of Sparse Vector Machines? Why are those classifiers particularly interesting with respect to linear regression for example? Illustrate with a drawing.

Part V Neural networks

16. Give the mathematical expression of the perceptron.

17. Explain the perceptron learning rule and state the associated perceptron convergence theorem.

18. Give the mathematical expression of a neural network (at least one hidden layer with two neurons and a general activation function) and illustrate this expression with a diagram.

19. During the programming session, we sometimes trained neural network with a stochastic gradient algorithm. What does stochastic mean here?

20. List and briefly explain the various regularization approaches that can be used in the training of neural networks.

21. Explain the difference between SGD, batch gradient descent and minibatch gradient descent.

22. Explain how gradient descent can be used to find the minimum of a function.

23. Give the expression of the cross entropy or log loss.

24. Let $\sigma(x)$ denote a sigmoid activation function. Show that $\sigma'(x) = \sigma(x)(1 - \sigma(x))$

25. Given a one or two hidden layers neural network, derive the gradient through backpropagation. (You can assume a cross entropy or log loss with sigmoid activations in each neuron).

Part VI Clustering

26. Give the pseudo-code for K-means, K-medoid. What is the strength/weakness of each approach ?

27. Give four possible initializations for K-means and describe each of them briefly.

28. What are the two main classes of hierarchical clustering algorithms?

29. In Agglomerative clustering, there are three main criteria used to select the two clusters to be merged. List those criteria and characterize each of them in terms of the dissimilarity used.

30. Give one particular example of a divisive clustering algorithm. How are the clusters split in this particular algorithm?

Part VII Linear latent variable models

31. What is Principal Component Analysis? This problem can read as a minimization problem, give the expression of this problem.

32. Explain how the principal components of a dataset can be obtained from the eigenvalue decomposition of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ (or the singular value decomposition of $\boldsymbol{X}$.) where the rows of $\boldsymbol{X}$ encode the prototypes $\{\boldsymbol{x}_\mu\}_{\mu=1}^N$.

33. Represent on a simple 2D dataset, the first and second principal directions.

34. What is Independent Component Analysis? Give 2 applications of this model.

Part VIII: Manifold learning

35. What is multidimensional scaling (MDS) ? How can it be extended to work with graph distances? Give the pseudo code for both MDS and its extension to graph distances (ISOMAP).

36. In ISOMAP, how do we define the graph representation of the data? Give two possible approaches.

37. Explain the intuition behind Locally Linear Embedding (LLE). What minimization problem(s) does the method solve to find a low dimensional representation of the data?

38. Explain how Self Organizing Maps (SOM) can be used to compute a low dimensional representation of the data.

Part IX: Reinforcement learning

39. What are the five constitutive elements of a reinforcement learning model? give a brief description of each element.

40. Give the pseudo-code for the Bandit algorithm.

41. Explain the difference between greedy and ε-greedy methods in RL? Why/when is it good to follow an ε-greedy strategy? When might it be a good idea to follow a greedy strategy? Following your answers to the previous two questions, how could you imagine adapting ε through time?

42. What is Q-learning ? How does the agent decide which action to take in this model (during training first and then during test)? How is the Q-table updated during training? Explain each of the terms that appear in the update. Give the pseudo code of Q-learning algorithm.