Home Exercises for *Statistical Learning*

Dong Liu

使用说明：

- 平时作业我们更关注大家是否按时、独立完成，做得正确与否不是评分的依据。
- 每次布置作业后，要求一周内完成，下次上课时交作业。
- 题目前标有 表示要求用计算机编程完成。
- 作业提交方式很灵活，可以提交纸质版（手写或打印），也可以提交电子版。如果提交手写纸质版，对于编程作业无需抄写源代码，但我们建议你写下程序设计的思路和程序运行的结果。如果提交打印版或电子版，不妨试试 Jupyter Notebook，它能把代码、文档和结果融为一体，又能输出成 PDF、HTML 等格式。

习题：

1. Is there any other method for machine learning beyond statistical learning? Hint: You may refer to the textbook *Machine Learning* written by T. M. Mitchell.

2. Learn to install Anaconda (version >= 3.7) and OpenCV-Python (version >= 3.4) on your PC. Write down the installation steps, which can be very helpful for the following exercises and projects.

3. Prove that the mean, median, and mode of a Gaussian random variable are equal.

4. Prove that the maximum likelihood estimator for σ^2 of a Gaussian distribution, i.e.
$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \tag{1}$$
where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, is biased.

5. Choose a pair of parameters (μ, σ^2) to generate N random numbers that follow the Gaussian distribution, then calculate the mean and variance of the random numbers and compare with the ground-truth (μ, σ^2). Set N to 100, 1000, 10000, 100000, …, to observe the change of results.

6. We already know that if we use $\hat{y} = \frac{\sum_i y_i}{N}$ to estimate a variable, it corresponds to minimizing least squares $\min_y \sum_i (y_i - y)^2$. Now we use $\hat{y} = \sqrt[N]{\prod_i y_i}$ to estimate a variable, what can be the corresponding minimization problem?

7. If the basis function is constant, i.e. $\phi(\boldsymbol{x}) = 1$, calculate the corresponding equivalent kernel function.

8. Solve the weighted least squares problem:

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} r_i (y_i - \boldsymbol{w} \cdot \boldsymbol{x_i})^2 \tag{2}$$

where $r_i > 0$ is the weight of $(\boldsymbol{x_i}, y_i)$.

9. Solve the regularized weighted least squares problem:

$$\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{N} r_i (y_i - \boldsymbol{w} \cdot \boldsymbol{x_i})^2 + \frac{\lambda}{2} ||\boldsymbol{w}||_2^2 \tag{3}$$

where $r_i > 0$ is the weight of $(\boldsymbol{x_i}, y_i)$.

10. Solve the following optimization problem.

$$\min_{\boldsymbol{x}} ||\boldsymbol{x}||_1, \text{subject to } ||\boldsymbol{x}||_2 = c \tag{4}$$

11. Solve the following optimization problem.

$$\min_{\boldsymbol{x}} ||\boldsymbol{x}||_p, \text{subject to } ||\boldsymbol{x}||_q = c \tag{5}$$

where $p > 0$ and $q > 0$. Hint: Consider the cases $p > q$, $p = q$, and $p < q$ separately.

12. Consider the following two optimization problems,

$$\min_{\boldsymbol{x}} J_1(\boldsymbol{x}) = ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}||_2^2 + \lambda_1 ||\boldsymbol{x}||_1 + \lambda_2 ||\boldsymbol{x}||_2^2 \tag{6}$$

$$\min_{\boldsymbol{x}} J_2(\boldsymbol{x}) = ||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{A}}\boldsymbol{x}||_2^2 + c\lambda_1 ||\boldsymbol{x}||_1 \tag{7}$$

Prove that, by choosing appropriate $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{A}}$, we can ensure that $\arg\min J_1(\boldsymbol{x}) = c \times \arg\min J_2(\boldsymbol{x})$. (This conclusion is useful when we want to solve the first problem, which seems more difficult, through solving the second one, which seems easier.)

13. If the prior is $p(w) = \mathcal{N}(w|\mu, \sigma_w^2)$, and the likelihood function is $p(y|x, w) = \mathcal{N}(y|wx, \sigma_e^2)$, calculate the posterior $p(w|y, x)$.

14. (a) Use the Lagrange multiplier method to solve the following optimization problem; (b) Invoke the constrained nonlinear optimization function in SciPy to solve the following optimization problem.

$$\min_{\boldsymbol{x}} 10 - x_1^2 - x_2^2, \text{subject to } x_2 \geq x_1^2, x_1 + x_2 = 0 \tag{8}$$

15. Randomly generate 100 datasets, each of which consists of 25 points that are samples of $\mathbf{y} = \sin(2\pi x) + \mathbf{e}$, where $x \in \{0.041 \times i, i = 0, 1, \dots, 24\}$, and $\mathbf{e}$ is additive white Gaussian noise with $\mathcal{N}(0, 0.3^2)$. Perform ridge regression on each dataset with 7th-order polynomial (with 8 free parameters) with different values of λ. Observe the results with respect to λ.

16. Calculate the differential entropy of Gaussian distribution $\mathcal{N}(x|\mu, \sigma^2)$.

17. Calculate the Kullback-Leibler divergence between two Gaussian distributions $\mathcal{N}(x|\mu, \sigma^2)$ and $\mathcal{N}(x|\nu, \chi^2)$.

18. Does log-normal distribution $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}x}\exp(-\frac{(\ln x - \mu)^2}{2\sigma^2})$ belong to the exponential family?

19. As we know Bernoulli distribution belongs to the exponential family, so how can we derive the variance of a Bernoulli variable using its cumulant function?

20. If a discrete random variable is defined on a finite set of integers, i.e. $x \in \{1, 2, \dots, N\}$, then under which distribution $P(x)$, the variable has the maximum entropy?

21. The XOR function can be represented by the dataset in Table 1, prove that the dataset is not linearly separable.

22. Design an algorithm to judge whether a dataset for binary classification is linearly separable or not.

23. Implement the perceptron learning algorithm by yourself. Use the dataset in Table 2 to test your algorithm.

24. Use the dataset in Table 2 to calculate a linear SVM, and indicate which samples are support vectors.

25. Prove that $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\frac{||\boldsymbol{x}-\boldsymbol{y}||_2^2}{2\sigma^2})$ is a kernel function. Hint: Use Taylor's expansion.

26. Invoke the SVM function in scikit-learn to calculate a linear SVM for the dataset given in Table 2.

27. We are training a classifier with 0-1 loss function. The hypothesis space consist of M functions. The training set consist of N samples. Prove that $\sup_\alpha(R(\alpha) - R_{\text{emp}}(\alpha)) \leq \sqrt{\frac{1}{2N}(\ln M - \ln \eta)}$ is satisfied with probability at least $1 - \eta$. Hint: If $M \geq 1$ and $0 < x < 1$, then $(1-x)^M \geq 1 - Mx$.

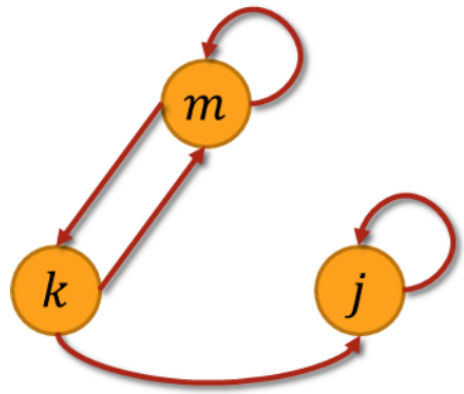

Figure 1: A graph for PageRank.

28. Suppose we are training a binary classifier. If our objective is to maximize the precision, which loss function is suitable? What if our objective is to maximize the recall?

29. What is k-d tree? Why is k-d tree useful for k-NN?

30. Use the gradient descent strategy to solve the sparse coding problem:

$$\min_{\boldsymbol{\alpha}} ||\boldsymbol{\alpha}||_1, \text{subject to } \boldsymbol{x} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x_i} \tag{9}$$

    Write down an algorithm in pseudo-code.

31. Consider the following clustering problem: we need to find a function $q : \mathbb{R} \to \{1, \ldots, k\}$, and the "center" of the j-th cluster is $c_j, j = 1, \ldots, k$. Our optimization target is $\min \sum_{i=1}^{N} |x_i - c_{q(x_i)}|$. How to solve this problem? Write down an algorithm in pseudo-code.

32. Given a set of data points: -67, -48, 6, 8, 14, 16, 23, 24, 28, 29, 41, 49, 56, 60, 75. Assume these data follow a two-Gaussian-mixture distribution, calculate the parameters of the distribution.

33. Calculate the volume of (a) a hypersphere in the D-dimensional space, whose radius is r; (b) a hypercube in the D-dimensional space, whose edge length is $2r$. Then calculate the ratio between the two volumes, and consider the ratio when $D \to \infty$.

34. Fig. 1 shows a graph with 3 nodes and 5 edges, all edge weights equal 1.0. Calculate the ranks of these nodes, using the PageRank algorithm, and setting the damping factor to 0.85 or 1.

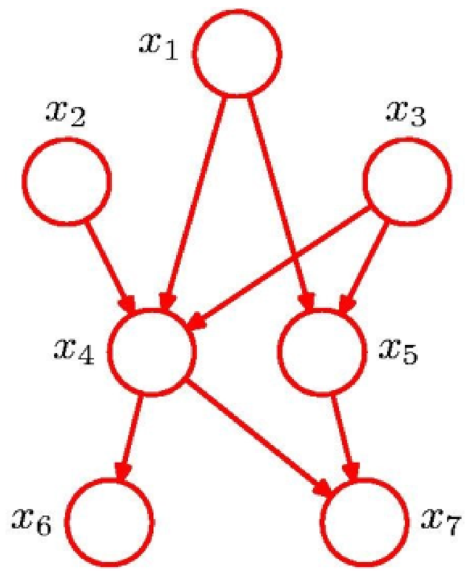

Figure 2: A Bayes network.

35. Implement the AdaBoost algorithm, with decision stamp as base learner, by yourself. Use the dataset in Table 3 for training, and $\boldsymbol{x} = (1, M)$ for test.

36. Implement an algorithm for building regression tree, where the leaf nodes are 3th-order polynomials rather than constants, use an adjustable threshold T to early terminate tree building. Test your algorithm with one dataset that you had generated in the Exercise 15. Observe the results with respect to T.

37. If $p(x_1, x_2, x_3, x_4, y) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_1, x_3)p(y|x_3, x_4)$, draw the corresponding Bayes network and factor graph, and convert it to Markov random field.

38. A Bayes network is shown in Fig. 2, (1) are x_5, x_6 conditionally independent given x_1, x_3? (2) what about x_5, x_6 given x_2, x_3? (3) what about x_2, x_5 given x_6, x_7?

39. Implement the naive Bayes algorithm (with Laplace smoothing and adjustable α) by yourself. Use the dataset in Table 3 for training, and $\boldsymbol{x} = (1, M)$ for test. Give the results when $\alpha = 0$ and $\alpha = 1$.

40. Implement the back propagation algorithm for multi-layer perceptron (MLP) network, where the activation function is sigmoid: $\frac{1}{1+\exp(-wx)}$. Test your algorithm, with 2-2-1 MLP, with the dataset in Table 1. Change the activation function to ReLU: $\max(x, 0)$, and test again.

Table 1: Dataset for XOR

x_1	0	0	1	1
x_2	0	1	0	1
y	-1	1	1	-1

Table 2: Dataset for binary classification on 2-D plane

$\boldsymbol{x_i}$	(1, 2)	(2, 3)	(3, 3)	(2, 1)	(3, 2)
y_i	1	1	1	-1	-1

Table 3: Dataset for binary classification with discrete input

x_1	1	1	1	1	1	2	2	2	2	2	3
x_2	S	M	M	S	S	S	M	M	L	L	L
y	-1	-1	1	1	-1	-1	-1	1	1	1	1