1 Question-answering (76)

1. (2) If an AI program passes the Turing test in real life, does it mean it can think like human beings? Why?
2. (1) Why do people say Dartmouth conference is the birth of AI?
3. (2) Define empirical distribution.
4. (2) What is the difference between test loss and population loss?
5. (1) What is the difference between regression and classification?
6. (2) What is the most natural loss function for classification problem (that we taught in class)? Why doesn't it work?
7. (2) What is memorization? Please describe the simplest example.
8. (1) What is mechanical turk?
9. (2) In terms of modern view & classical view of overfitting, compare explicit regularization and implicit regularization.
10. (2) How do you use a generative model to describe a bizarre distribution?
11. (2) PCA is one of the most widely used techniques for dimension reduction. For what kind of data, PCA is not very useful?
12. (2) What is zero-th order method?
13. (3) What is the smoothness assumption?
14. (2) Why do we say GD is essentially a greedy algorithm?
15. (3) Describe the connection between smoothness/convexity and the eigenvalues of Hessian matrix.
16. (2) What is SGD algorithm? State the updating rule.
17. (3) What is the converging rate of SGD/GD for convex and smooth function?
18. (3) State the SVRG algorithm.
19. (3) State the perceptron algorithm. What's the loss function that perceptron algorithm is designed for?
20. (2) We can use both ℓ_1 loss and cross entropy loss to measure the distance between the two probability distributions. Why do people usually use cross entropy empirically?
21. (2) In Lasso, why do we use $\|w\|_1$ as the regularizer, instead of $\|w\|_0$?
22. (2) Please describe the compressed sensing setting. What is the measurement matrix A?
23. (2) What is the RIP condition?

24. (3) In class, we mentioned the non-linear version of compressed sensing. Please state inequality appeared in the theorem, and describe the three terms on the right hand side. What do they mean?

25. (3) In class, we learned one theorem saying that if U is an orthonormal matrix, W is a random Guassian matrix, then with decent probability, WU is RIP. Please describe why we need U here in practice.

26. (3) Write down the program for solving hard SVM.

27. (3) In the dual program of SVM, how many variables are we optimizing?

28. (3) What is kernel trick for SVM?

29. (3) What is Mercer's theorem?

30. (3) No free lunch theorem says there is no universal learner. However, in practice, models like ResNet/DenseNet work really well for almost all the CV applications. Why?

31. (3) What is VC dimension?

32. (4) What is Massart lemma? What does it mean?

2 Convergence analysis of GD (6)

Consider a L-smooth and convex function f, and we run gradient descent algorithm on it. We already know:

- By smoothness, we have $f(w_{i+1}) \le f(w_i) - \frac{\eta}{2}\|\nabla f(w_i)\|^2$
- By convexity, we have $f(w_i) \le f(w^*) + \langle \nabla f(w_i), w_i - w^* \rangle$

Prove:

$$f(w_{i+1}) \le f(w^*) + \frac{1}{2\eta}\|w_i - w^*\|^2 - \frac{1}{2\eta}\|w_{i+1} - w^*\|^2$$

3 Almost orthogonality (6)

Recall when proving the compressed sensing theorem, we used the almost orthogonality property of RIP matrix. We have the following theorem: If W is $(\epsilon, 2s)$-RIP, $\forall I, J$ disjoint sets of size $\le s$, for any vector μ we have $\langle W\mu_I, W\mu_J \rangle \le \epsilon \|\mu_I\| \|\mu_J\|$. Recall W is $(\epsilon, 2s)$-RIP, $y = Wx$, $x^* \in \arg\min_{v: Wv = y} \|v\|_1$, $h = x^* - x$, and T is defined as:

- T_0 contains the s largest elements in absolute value in x, $T_0^c = [d] \backslash T_0$
- T_1 contains the s largest elements in $h_{T_0^c}$, $T_{0,1} = T_0 \cup T_1$
- T_2 contains the s largest elements in $h_{T_{0,1}^c}$. T_3, T_4 are constructed in the same way.

Prove:

$$\|Wh_{T_{0,1}}\|_2^2 \le \sqrt{2}\epsilon \|h_{T_{0,1}}\|_2 \sum_{j \ge 2} \|h_{T_j}\|_2$$

4 Perceptron (6)

In the proof of convergence of Perceptron algorithm, how do we show $\|w_{t+1}\| \geq t\gamma$? Notice that t is the number of mistakes that we made, and we assume there exists w^* s.t. $\|w^*\| = 1$, and $\exists\gamma > 0$ s.t. $\forall i, y_i\langle w^*, x_i\rangle \geq \gamma$, and we start from $w_0 = \mathbf{0}$ (all zero vector).

5 Rademacher complexity (6)

In the proof for showing the relationship between representativeness and Rademacher complexity, we have the following equation:

$$E_{S,S'}\left[\sup_{f\in F}\sum_{i=1}^{m}(f(z_i') - f(z_i))\right] = E_{S,S',\sigma}\left[\sup_{f\in F}\sum_{i=1}^{m}\sigma_i(f(z_i') - f(z_i))\right]$$

Where S, S' are two training set samples of size m, and σ_i is the Rademacher random variable. Please prove this equation.