CS 560 Statistical Machine Learning: Mid-Term Exam

Instructor: Jie Shen

March 15, 2020, 15:00 – 17:30 EST

Instructions:

- Open-book exam, notes allowed;
- No electronic device, no discussion, no sharing;
- 20 points for each problem, totally 5.

Write down your name. (10 pts)

1. In which aspect(s) does the course reshape your understanding of machine learning?

2. Let D be a distribution over $\mathbb{R}$ where the mean is 5 and variance is 9. Suppose $x_1, \dots, x_{10}$ are independent draws from D. Plot the possible positions of these random variables on the real line.

3. Let $w \in \mathbb{R}^d$ be the variable, and let $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ be given. Calculate the gradient of the following functions with respect to w:

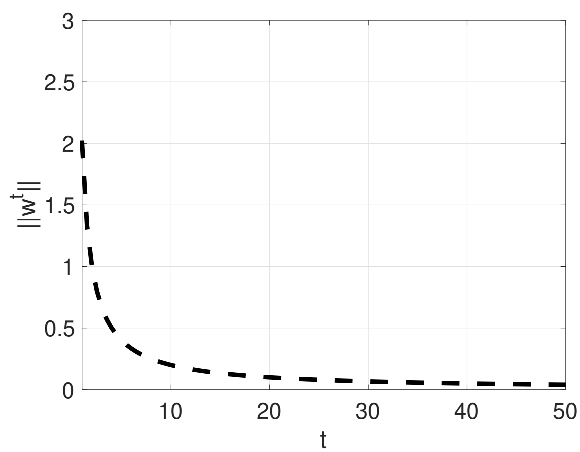
- $F(w) = (y - w \cdot x)^{100}$;
- $F(w) = \frac{1}{y + w \cdot x}$;
- $F(w) = \log(1 + yw \cdot x)$;
- $F(w) = e^{(w \cdot x)^2}$.

4. Many machine learning problems boil down to solving the following optimization program:

$$\min_{\boldsymbol{w}} \ F(\boldsymbol{w}), \quad \mathrm{s.t.}\ \boldsymbol{w} \in \mathbb{R}^d. \tag{1}$$

Suppose that $d = 2$ and $F(\boldsymbol{w}) = \frac{1}{2}\left(w_1^2 + (w_1 + w_2)^2\right)$ where w_1 and w_2 are the first and second coordinates of $\boldsymbol{w}$ respectively.

- Calculate the gradient and the Hessian matrix of $F(\boldsymbol{w})$;
- Show that $F(\boldsymbol{w})$ is a strongly convex and smooth function, and calculate the strong convexity parameter α and smoothness parameter L;
- Consider that we run gradient descent (GD) to find the global optimum of $F(\boldsymbol{w})$, starting from the initial iterate $\boldsymbol{w}^0 = (1, 1)$ and proceed with learning rate $\eta = 1/2$. Calculate the iterates $\boldsymbol{w}^1, \boldsymbol{w}^2, \boldsymbol{w}^3$.
- Suppose we are able to calculate more iterates $\boldsymbol{w}^4, \boldsymbol{w}^5, \ldots, \boldsymbol{w}^t, \ldots$ with $\eta = 1/2$, and we plot the curve "$\left\|\boldsymbol{w}^t\right\|_2$ v.s. t" as below. If we run GD with $\eta = 2/3$, what will the curve likely be? What about $\eta = 2$? Please plot them in the same figure and explain how you obtain these curves.



- Now consider minimizing the same function with stochastic GD, where the learning rate $\eta_t = 1/t$ at the t-th iteration. Plot "$\left\|\boldsymbol{w}^t\right\|_2$ v.s. t" in the figure above.

5. Let D be the distribution of training data and D' be that of test data. A key condition under which classical PAC learning results hold is that $D' = D$. Give an example to show that when $D' \neq D$, any learner with access to finite training data, even with unlimited computational power, may incur a high testing error.