CS 189 Introduction to Machine Learning
Fall 2018 HW1

1 Getting Started

Read through this page carefully. You may typeset your homework in latex or submit neatly handwritten/scanned solutions. Please start each question on a new page. Deliverables:

1. Submit a PDF of your writeup to assignment on Gradescope, "HW1 Write-Up". If there are graphs, include those graphs in the correct sections. Do not simply reference your appendix.

(a) Who else did you work with on this homework? In case of course events, just describe the group. How did you work on this homework? Any comments about the homework?

(b) Please copy the following statement and sign next to it. We just want to make it *extra* clear so that no one inadverdently cheats.

I certify that all solutions are entirely in my words and that I have not looked at another student's solutions. I have credited all external sources in this write up.

This homework is due **Monday, September 3rd at 10 p.m.**

2 The accuracy of learning decision boundaries

This problem exercises your basic probability (e.g. from 70) in the context of understanding why lots of training data helps to improve the accuracy of learning things.

For each $\theta \in (1/3, 2/3)$, define $f_\theta : [0,1] \to \{0,1\}$, such that

$$f_\theta(x) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ otherwise.} \end{cases}$$

The function is plotted in Figure 1.

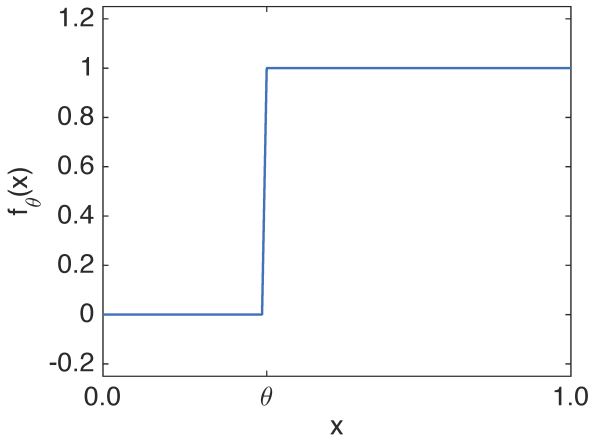

Figure 1: Plot of function $f_\theta(x)$ against x.

We draw samples $X_1, X_2, \ldots, X_n$ uniformly at random and i.i.d. from the interval $[0,1]$. Our goal is to learn an estimate for θ from n random samples $(X_1, f_\theta(X_1)), (X_2, f_\theta(X_2)), \ldots, (X_n, f_\theta(X_n))$.

Let $T_{min} = \max(\{\frac{1}{3}\} \cup \{X_i | f_\theta(X_i) = 0\})$. We know that the true θ must be larger than T_{min}.

Let $T_{max} = \min(\{\frac{2}{3}\} \cup \{X_i | f_\theta(X_i) = 1\})$. We know that the true θ must be smaller than T_{max}.

The gap between T_{min} and T_{max} represents the uncertainty we will have about the true θ given the training data that we have received.

(a) **What is the probability that $T_{max} - \theta > \epsilon$ as a function of ϵ? And what is the probability that $\theta - T_{min} > \epsilon$ as a function of ϵ?**

(b) Suppose that you would like the estimator $\hat{\theta} = (T_{max} + T_{min})/2$ for θ that is ϵ-close (defined as $|\hat{\theta} - \theta| < \epsilon$, where $\hat{\theta}$ is the estimation and θ is the true value) with probability at least $1 - \delta$. Both ϵ and δ are some small positive numbers. **Please bound or estimate how big of an n do you need?** You do not need to find the optimal lowest sample complexity n, an approximation using results of question (a) is fine.

(c) Let us say that instead of getting random samples $(X_i, f(X_i))$, we were allowed to choose where to sample the function, but you had to choose all the places you were going to sample in advance. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?** (**Hint:** You need not use a randomized strategy.)

(d) Suppose that you could pick where to sample the function adaptively — choosing where to sample the function in response to what the answers were previously. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?**

(e) In the three sampling approaches above: random, deterministic, and adaptive, **compare the scaling of n with ϵ (and δ as well for the random case).**

(f) **Why do you think we asked this series of questions? What are the implications of those results in a machine learning application?**

3 Gambling is for deviations

Suppose you go to a casino which has $n \geq 2$ slot machines, where the payouts from the i-th slot machine are i.i.d. random variables with distribution $\mathcal{N}(\theta_i, 1)$, where θ_i are real numbers. Assume that one of the slot machines has mean payout $\theta_{\max}$, and all other machines have mean payout $\theta_{\min} < \theta_{\max}$. For a fixed level of confidence $\delta \in (0, 1)$, your goal is to identify the slot machine with the highest mean payout with probability of error δ. You are allowed to do this by pulling each slot machine T times, guessing the best slot machine based based on the $T \times n$ payouts observed. At the end, you want to ensure that your guess is correct with probability at least $1 - \delta$. In these steps, you will determine an upper bound on the number of times T you neeed to identify the best slot machine.

(a) **Show that if X is a real valued random variable, you can bound $\mathbf{P}[X \geq t] \leq \mathbb{E}[X\mathbf{I}(X \geq t)]/t$ for all $t > 0$, where $\mathbf{I}(X \geq t) = 1$ if $X \geq t$, and 0 otherwise**. If it's easier, you may assume X has a continuous density $p(x)$.

(b) Let Z be distributed as $\mathcal{N}(0, 1)$. **Show that**

$$\forall t > 0, \quad \mathbf{P}[Z \geq t] \leq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}$$

Use this to bound $\mathbf{P}[|Z| \geq t]$.

(c) Let $Z_1, \ldots, Z_n$ be distributed $\mathcal{N}(0, 1)$ (not necessarily independent!), and let $n \geq 2$. **Show that for any $t \geq 1$**

$$\mathbf{P}[\max_i |Z_i| \geq t] \leq n \cdot \sqrt{\frac{2}{\pi}} e^{-t^2/2}.$$

(d) Suppose $n = 2$. **Show that that, in order to identify the slot machine with the highest payout with probability $1 - \delta$, it suffices to take**

$$T \geq \max\left\{1, \frac{4\log(2/\delta)}{(\theta_{\max} - \theta_{\min})^2}\right\} \text{ samples.}$$

You should use the inequalities developed earlier in the problem.

(e) **Generalize the above result to $n \geq 2$ slot machines.** When δ is a constant (say $\delta = 1/2$), there should be a $\log n$ somewhere in your answer.

4 Much ado about norms

Recall that a norm $\|\cdot\|$ is a function from $\mathbb{R}^d \to \mathbb{R}$ which satisfies the following properties:

(a) For all $x \in \mathbb{R}^d$, $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$

(b) For any real number α, $\|\alpha x\| = |\alpha|\|x\|$

(c) For any $x, y \in \mathbb{R}^d$, $\|x + y\| \leq \|x\| + \|y\|$

(a) For $p \in (0, \infty)$, let $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, and let $\|x\|_\infty = \max_i |x_i|$. **Please prove the following inequalities:**

(a) Show that $\|x\|_2$, $\|x\|_1$, and $\|x\|_\infty$ are norms.
(b) Show that $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$.
(c) Show that $\|x\|_2^2 \leq \|x\|_1 \|x\|_\infty$.
(d) Show that $\|x\|_1 \leq \sqrt{d}\|x\|_2$ and $\|x\|_2 \leq \sqrt{d}\|x\|_\infty$.

(b) **For each the following functions $f(x) : \mathbb{R}^d \to \mathbb{R}$, state if f is always, sometimes, or never a norm. Give a proof. If the answer is 'sometimes', give a necessary and sufficient condition for f to be a norm**

(a) $f(x) = \log\cosh \|x\|_2$.
(b) $f(x) = \sum_{i=1}^n x_i^2$.
(c) $f(x) = \|Ax\|$, where $\|x\|$ is a norm on $\mathbb{R}^d$, and $A \in \mathbb{R}^{d \times d}$.
(d) $f(x) = \sqrt{x^\top \Sigma x}$ where Σ is symmetric and has positive eigenvalues.
(e) $f(x) = \sqrt{x^\top \Sigma x}$, where $\Sigma = \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}$, and $A \in \mathbb{R}^{n \times m}$ where $m + n = d$.
(f) $f(x) = \sum_i \alpha_i |x_i|$, $\alpha_i \in \mathbb{R}$.
(g) $f(x) = \sup_{w \in \mathcal{C}} \langle w, x \rangle$, where $\mathcal{C} \subset \mathbb{R}^d$ has the following properties:
- $x \in \mathcal{C}$ if and only if $-x \in \mathcal{C}$.
- $\mathcal{C}$ is bounded; that is $\sup_{x \in \mathcal{C}} \|x\| < \infty$.
- There exists an orthonormal basis $\{e_1, e_2, \ldots, e_d\} \subset \mathcal{C}$.

5 Projecting your problems

Given $1 \leq d \leq n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(d) = P$, $P = P^\top$ and $P^2 = P$.

(a) **Prove that P is projection matrix if and only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^\top$ and $U^\top U = I$**

(b) **Prove that if P is a rank d projection matrix, then $\text{tr}(P) = d$.**

(c) **Prove that, for all $v \in \mathbb{R}^n$,**

$$Pv = \arg \min_{w \in \text{range}(P)} \|v - w\|_2^2.$$

(d) **Prove that if $X \in \mathbb{R}^{d \times d}$ and $\text{rank}(X) = d$, then $X(X^\top X)^{-1}X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U?**

(e) Let $y = X\theta_* + z$, where $X \in \mathbb{R}^{n \times d}$, $\theta_* \in \mathbb{R}^d$, $y \in \mathbb{R}^n$, and $z = \mathcal{N}(0, I) \in \mathbb{R}^n$, and suppose $\text{rank}(X) = d$. **Prove that if $\widehat{\theta} = \arg\min_\theta \|X\theta - y\|_2^2$, then**

$$\mathbb{E}[\|\theta_* - \widehat{\theta}\|_2^2] = \text{tr}\Big((X^\top X)^{-1}\Big)$$

(f) **In the setting of the Part (e), show that**

$$\frac{1}{n}\mathbb{E}[\|X(\theta_* - \widehat{\theta})\|_2^2] = \frac{d}{n}.$$

How does the answer change if $\text{rank}(X) < d$?

6 Installing Python

Set up your Python environment for the course by downloading Anaconda 5.2 here: `https://www.anaconda.com/download/#macos` (Mac), `https://www.anaconda.com/download/#windows`(Windows) or `https://www.anaconda.com/download/#linux`(Linux). Make sure to install the *Python 3.6* version.

7 Your Own Question

Write your own question, and provide a thorough solution.

Writing your own problems is a very important way to really learn the material. The famous "Bloom's Taxonomy" that lists the levels of learning is: Remember, Understand, Apply, Analyze, Evaluate, and Create. Using what you know to create is the top-level. We rarely ask you any HW questions about the lowest level of straight-up remembering, expecting you to be able to do that

yourself. (e.g. make yourself flashcards) But we don't want the same to be true about the highest level.

As a practical matter, having some practice at trying to create problems helps you study for exams much better than simply counting on solving existing practice problems. This is because thinking about how to create an interesting problem forces you to really look at the material from the perspective of those who are going to create the exams.

Besides, this is fun. If you want to make a boring problem, go ahead. That is your prerogative. But it is more fun to really engage with the material, discover something interesting, and then come up with a problem that walks others down a journey that lets them share your discovery. You don't have to achieve this every week. But unless you try every week, it probably won't happen ever.