Machine Learning, 2021 Fall
Assignment 1

Notice

Due 23:59 (CST), Oct. 23, 2021
Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

1 Gradient Descent

In order to minimize $f(\boldsymbol{x})$ where $x \in R^n$, we takes iteration:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{p}^k$$

where $\boldsymbol{p}^k = \boldsymbol{H}^k \nabla f\left(\boldsymbol{x}^k\right)$ and $\alpha^k \to 0^+$. What kind of $\boldsymbol{H}^k$ can guarantee that $\boldsymbol{p}^k$ is a descent direction ? Give a detailed proof.[1pts]

2 Convex

(1) Prove that $f: R^n \to R$ is a convex function if and only if $epi f = \{(x,t) \in R^{n+1} | x \in dom(f),\ f(x) \leq t\}$ is a convex set. [0.5pts]
(2) Let $f_1, f_2, ..., f_k$ be convex functions on R^n, prove that $f(x) = max\{f_1(x), f_2(x), ..., f_k(x)\}$ is also a convex function. [0.5pts]
(3) Prove that $f: R^n \to R$ is a convex function if and only if $g: R \to R$

$$g(t) = f(x+tv) \quad dom(g) = \{t | x + tv \in dom(f)\}$$

is convex for any $x \in dom(f)$ and $v \in R^n$. [0.5pts]
(4) Prove that $f(x) = log(det(x)) \quad dom(f) = S^n_{++}$ is a concave function. [0.5pts]

3 Learning

Assume that $\mathcal{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N, \mathrm{x}_{N+1}, \ldots, \mathrm{x}_{N+M}\}$, $N, M \in \mathbb{N}^+$ and $\mathcal{Y} = \{-1, +1\}$ with an unknown target function $f: \mathcal{X} \to \mathcal{Y}$. The training data set $\mathcal{D}$ is $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$. Define the off-training-set error of a hypothesis h with respect to f by

$$E_{\text{off}}(h, f) = \frac{1}{M} \sum_{m=1}^{M} \left[h\left(\mathbf{x}_{N+m}\right) \neq f\left(\mathbf{x}_{N+m}\right)\right]$$

(a) Say $f(\mathrm{x}) = +1$ for all x and

$$h(\mathbf{x}) = \begin{cases} +1, & \text{for } \mathbf{x} = \mathbf{x}_k \text{ and } k \text{ is odd and } 1 \leq k \leq M+N \\ -1, & \text{otherwise} \end{cases}$$

What is $E_{\text{off}}(h, f)$? [0.5pts]

(b) We say that a target function f can 'generate' $\mathcal{D}$ in a noiseless setting if $y_n = f(\mathbf{x}_n)$ for all $(\mathbf{x}_n, y_n) \in \mathcal{D}$. For a fixed $\mathcal{D}$ of size N , how many possible $f : \mathcal{X} \to \mathcal{Y}$ can generate $\mathcal{D}$ in a noiseless setting? [0.25pts]

(c) For a given hypothesis h and an integer k between 0 and M , how many of those f in (b) satisfy $E_{\text{off}}(h, f) = \frac{k}{M}$? [0.25pts]

(d) For a given hypothesis h , if all those f that generate $\mathcal{D}$ in a noiseless setting are equally likely in probability, what is the expected off trainingset error $\mathbb{E}_f\left[E_{\text{off}}(h, f)\right]$? [0.5pts]

(e) A deterministic algorithm A is defined as a procedure that takes $\mathcal{D}$ as an input, and outputs a hypothesis $h = A(\mathcal{D})$. Argue that for any two deterministic algorithms A_1 and A_2. [0.5pts]

$$\mathbb{E}_f\left[E_{\text{off}}\left(A_1(\mathcal{D}), f\right)\right] = \mathbb{E}_f\left[E_{\text{off}}\left(A_2(\mathcal{D}), f\right)\right]$$

4 MAE

The Empirical risk minimization(ERM) principle is meant to choose a hypothesis $\hat{h}$ which minimizes the empirical risk $\hat{R}_{\mathcal{D}}[h]$.

(a) Consider the following hypothesis and loss function

$$\mathcal{H} = \{h_\theta(x) = \theta_1 x : \theta_1 \in \mathbb{R}\},$$
$$\mathcal{L}(\theta_1) = \frac{1}{N} \sum_{i=1}^{N} \left| h_\theta\left(x^{(i)}\right) - y^{(i)} \right|$$

Assume we already have a dataset $\mathcal{D} = \{(1, 3), (-1, -2), (2, 4)\}$. Derive the value of θ_1 which minimizes the empirical risk. [0.5pts]

(b) Assume we have a dataset $\mathcal{D} = \{x_1, x_2, \cdots, x_n\}$ where $x_i \in \mathcal{R}$. Consider the hypothesis $\mathcal{H}$ to be

$$\mathcal{H} = \{h_\theta = \theta_0 : \theta_0 \in \mathbb{R}\}$$

Derive the hypothesis h^* which minimizes the empirical risk. [0.5pts]

$$h^* = \arg\min_h \frac{1}{N} \sum_{i=1}^{N} \left| h - x^{(i)} \right| \quad \text{s.t.} \quad h \in \mathcal{H}$$